Complex Genetic Approaches to Neurodegenerative Diseases

A thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy to the University of London

by

Paresh Rameshchandra Shah

MRC Prion Unit, Institute of Neurology,
University College London
Abstract

Neurodegenerative diseases are fatal disorders in which disease pathogenesis results in the progressive degeneration of the central and/or the peripheral nervous systems. These diseases currently affect ~2% of the population but are expected to increase in prevalence as average life expectancy increases. The majority of these diseases have a complex genetic basis. The work presented in this thesis aimed to investigate the genetic basis of two neurodegenerative diseases, amyotrophic lateral sclerosis (ALS) and the human prion diseases kuru and sporadic Creutzfeldt-Jakob disease (sCJD), using novel complex genetic approaches.

ALS is a fatal neurodegenerative disease in which motor neurons are seen to degenerate. It is a complex disease with ~10% of individuals having a family history and the remaining 90% of non-familial cases having some genetic component. The gene DYNCIHI is involved in retrograde axonal transport and is a good candidate for ALS. In this thesis the genetic architecture of DYNCIHI was elucidated and a mutation screen of exons 8, 13 and 14 was undertaken in familial forms of ALS and other motor neuron diseases. No mutations were found. A linkage disequilibrium (LD) based association study was conducted using two tagging single nucleotide polymorphisms (tSNPs) which were identified as sufficient to represent genetic variation across DYNCIHI. These tSNPs were tested for an association with sporadic ALS (SALS) in 261 cases and 225 matched controls but no association was identified.

Kuru is a devastating epidemic prion disease which affected a highly geographically restricted area of the Papua New Guinea highlands, predominantly affected adult women and children. Its incidence has steadily declined since the cessation of its route of transmission, endocannibalism, in the late 1950’s. Kuru imposed strong balancing selection on codon 129 of the prion gene (PRNP). Analysis of kuru-exposed and unexposed populations showed significant deviations from Hardy-Weinberg equilibrium (HWE) consistent with the known protective effect of codon 129 heterozygosity. Signatures of selection were investigated in the surviving populations, such as deviations from HWE and an increasing cline in codon 129 valine allele frequency, which covaried with disease exposure. A novel PRNP G127V polymorphism was detected which, while common in the area of highest kuru incidence, was absent from kuru patients and unexposed population groups. Genealogical analysis revealed that the heterozygous PRNP G127V genotype confers strong prion disease resistance, which has been selected by the kuru epidemic.

Finally, PRNP copy number was investigated as a possible genetic mechanism for susceptibility to kuru and sCJD. No conclusive copy number changes were identified.
Acknowledgements

Firstly, a massive thank you to my supervisors Professor Elizabeth Fisher and Professor John Collinge. Lizzy, your patience, support and guidance throughout the last three and a half years has gone above-and-beyond the duty of a supervisor and is something that I will always remember. John, thank you for the opportunity to contribute to the outstanding work produced at the MRC Prion Unit. In addition, thanks to Dr Simon Mead – my third supervisor. Simon, your passion for science and your willingness to explain even the most complicated population genetics concepts has been truly influential.

Much of the work in this thesis would not have been possible without the help of several people. Thanks to both Dr Azlina Ahmad-Annuar and Dr Majid Hafezparast for helping me find my feet at the bench at the beginning of my doctorate and your continuing support throughout my studies. I owe a huge debt of thanks to the molecular genetics group at the MRC Prion Unit. James Uphill helped with much of the microsatellite work, Gary Adamson for his help with sequencing and genotyping on the ABI 377, Huda Al-Dujaily and Tracy Campbell for help with extracting DNA from PNG blood samples, John Beck for answering my many questions and Mark Poulter for being the font of all allele discrimination and quantitative PCR knowledge, In addition, thanks to Jerome Whitfield, Mike Alpers and the rest of the PNGIMR field team, for the excellent epidemiology data and samples. Thanks also to Professor Robert Brown and the Day Lab in Boston for your assistance and hospitality during my visit.

The investigations discussed in this thesis would not have been possible without (i) the many thousands of samples collected and shared by various research groups around the world, to whom I am extremely grateful, (ii) all of the families, patients and healthy individuals who donated samples or consented to the collection and use of these samples and (iii) the UK Medical Research Council who funded this project.

I owe much of my perseverance over the last three and a half years to my fellow PhD compatriots, friends and lab mates past and present, without whom, this PhD would have been much more difficult and much less fun. The roll-call is too long to list here but you know who you are and I'll never forget you.

And finally, this thesis is dedicated to my parents who I love more than I have ever dared to say and to Kate, mi vida, without who’s love and support none of this would have been possible.
# Table of contents

Abstract .......................................................................................................................................... 2  
Acknowledgements ....................................................................................................................... 3  
Table of contents ........................................................................................................................... 4  
List of figures ............................................................................................................................... 12  
List of tables ................................................................................................................................. 14  
Abbreviations ............................................................................................................................... 15  

1 Introduction .......................................................................................................................... 19  
1.1 Neurodegenerative diseases ........................................................................................ 19  
  1.1.1 Common features of neurodegenerative diseases ....................................................... 19  
  1.1.2 Related genetic mechanisms ......................................................................................... 20  
1.2 Gene identification in Mendelian forms of neurodegenerative diseases ............... 20  
1.3 The genetic component of non-Mendelian diseases ................................................ 21  
  1.3.1 Familial genetics inform susceptibility loci in sporadic disease .................................. 21  
  1.3.2 Familial clustering of disease ....................................................................................... 22  
1.4 Complex diseases ......................................................................................................... 23  
1.5 Gene identification strategies in complex diseases ................................................... 25  
  1.5.1 Linkage analysis based approaches ............................................................................. 25  
  1.5.2 Association studies ....................................................................................................... 25  
    1.5.2.1 Linkage disequilibrium .......................................................................................... 26  
    1.5.2.2 Linkage disequilibrium patterns in the human genome ....................................... 27  
    1.5.2.3 Tagging SNPs ......................................................................................................... 28  
  1.5.3 Association study design .............................................................................................. 29  
    1.5.3.1 Study power and sample size ................................................................................. 29  
    1.5.3.2 Sample selection ...................................................................................................... 30  
    1.5.3.3 Candidate gene and whole genome approaches .................................................. 31  
    1.5.3.4 Allelic architecture of complex disease ................................................................. 32  
    1.5.3.5 Evolving resources for association studies in complex diseases .......................... 33  
1.6 Evolutionary analyses ................................................................................................. 34  
  1.6.1 Genetic variation is shaped by several forces ............................................................ 35  
  1.6.2 Measuring genetic variation and disease mapping by natural selection ............... 35  
    1.6.2.1 Natural selection ...................................................................................................... 36  
    1.6.2.2 Testing for signatures of selection ........................................................................ 38  
    1.6.2.3 Confounding factors – demography and ascertainment ....................................... 39  
  1.6.3 Hardy-Weinberg equilibrium ..................................................................................... 40  
1.7 Amyotrophic lateral sclerosis ..................................................................................... 40  
  1.7.1 Clinical and pathological features of ALS ................................................................. 40
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.7.2</td>
<td>Epidemiology</td>
<td>41</td>
</tr>
<tr>
<td>1.7.3</td>
<td>The Mendelian genetic basis of ALS</td>
<td>41</td>
</tr>
<tr>
<td>1.7.4</td>
<td>The complex genetic basis of ALS</td>
<td>44</td>
</tr>
<tr>
<td>1.7.4.1</td>
<td>A genetic component to sporadic ALS</td>
<td>45</td>
</tr>
<tr>
<td>1.7.5</td>
<td>Susceptibility genes in sporadic ALS</td>
<td>46</td>
</tr>
<tr>
<td>1.7.6</td>
<td>DYNCIHI as a candidate susceptibility locus</td>
<td>49</td>
</tr>
<tr>
<td>1.7.6.1</td>
<td>The cytoplasmic dyneins and dynactin</td>
<td>49</td>
</tr>
<tr>
<td>1.7.6.2</td>
<td>Cytoplasmic dynein 1 heavy chain subunit and function in neurons</td>
<td>50</td>
</tr>
<tr>
<td>1.7.6.3</td>
<td>Cytoplasmic dynein and human motor neuron degeneration</td>
<td>50</td>
</tr>
<tr>
<td>1.8</td>
<td>Human prion diseases</td>
<td>52</td>
</tr>
<tr>
<td>1.8.1</td>
<td>Aetiology of prion diseases</td>
<td>52</td>
</tr>
<tr>
<td>1.8.1.1</td>
<td>PrP and PRNP</td>
<td>52</td>
</tr>
<tr>
<td>1.8.2</td>
<td>Inherited prion diseases</td>
<td>53</td>
</tr>
<tr>
<td>1.8.3</td>
<td>Sporadic prion disease</td>
<td>54</td>
</tr>
<tr>
<td>1.8.4</td>
<td>Acquired CJD</td>
<td>54</td>
</tr>
<tr>
<td>1.8.5</td>
<td>Kuru</td>
<td>56</td>
</tr>
<tr>
<td>1.8.5.1</td>
<td>Clinical and pathological characteristics</td>
<td>57</td>
</tr>
<tr>
<td>1.8.5.2</td>
<td>Genetics of kuru</td>
<td>57</td>
</tr>
<tr>
<td>1.8.6</td>
<td>Evidence for human genetic susceptibility to prion disease</td>
<td>57</td>
</tr>
<tr>
<td>1.8.6.1</td>
<td>Information from animal studies - inbred mouse lines</td>
<td>57</td>
</tr>
<tr>
<td>1.8.6.2</td>
<td>HLA</td>
<td>60</td>
</tr>
<tr>
<td>1.8.6.3</td>
<td>PRNP codon 129 polymorphism</td>
<td>60</td>
</tr>
<tr>
<td>1.8.6.4</td>
<td>PRNP regulatory elements, PRNP expression and prion disease</td>
<td>61</td>
</tr>
<tr>
<td>1.9</td>
<td>Aims of this thesis</td>
<td>62</td>
</tr>
<tr>
<td>2</td>
<td>Materials and methods</td>
<td>63</td>
</tr>
<tr>
<td>2.1</td>
<td>Materials</td>
<td>63</td>
</tr>
<tr>
<td>2.1.1</td>
<td>General chemicals and reagents</td>
<td>63</td>
</tr>
<tr>
<td>2.1.2</td>
<td>Prepared solutions</td>
<td>64</td>
</tr>
<tr>
<td>2.1.3</td>
<td>Commercial kits</td>
<td>64</td>
</tr>
<tr>
<td>2.1.4</td>
<td>Restriction enzymes and ligases</td>
<td>65</td>
</tr>
<tr>
<td>2.1.5</td>
<td>Equipment</td>
<td>65</td>
</tr>
<tr>
<td>2.1.6</td>
<td>Photography</td>
<td>66</td>
</tr>
<tr>
<td>2.1.7</td>
<td>Software and websites</td>
<td>66</td>
</tr>
<tr>
<td>2.2</td>
<td>Samples</td>
<td>67</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Healthy population DNA samples</td>
<td>67</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Patient/case DNA samples</td>
<td>67</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Papua New Guinea Samples</td>
<td>68</td>
</tr>
</tbody>
</table>
2.3 Methods

2.3.1 DNA isolation from chimpanzee blood

2.3.2 DNA isolation from human blood

2.3.3 Determination of DNA concentration and purity

2.3.4 Primer design

2.3.4.1 Real-time PCR primers and probe design

2.3.5 Polymerase chain reaction

2.3.6 Purification of PCR amplicons for sequencing

2.3.7 Sequencing of purified templates

2.3.8 Post-reaction clean-up with ethanol/salt precipitation

2.3.9 DNA sequencing on the MegaBACE 1000 DNA Analysis System

2.3.10 SNP genotyping by restriction digestion

2.3.11 Microsatellite genotyping

2.3.12 Affymetrix GeneChip 250k NspI Assay

2.3.12.1 NspI digestion

2.3.12.2 NspI adaptor ligation

2.3.12.3 Adaptor-ligated fragment PCR and clean-up

2.3.12.4 Fragmentation

2.3.12.5 Labelling and target hybridisation

2.3.12.6 Washing, staining and scanning

2.3.12.7 Array data analysis

2.3.13 Quantitative real-time PCR

2.3.14 Linkage disequilibrium analysis and “tagging” SNP selection

2.3.15 Haplotype inference

2.3.16 Human and mouse cytoplasmic dyneins nomenclature, map positions and sequences

2.3.17 Human/mouse homology searches

2.3.18 Phylogenetic analyses of the cytoplasmic dynein genes

2.3.19 PRNP codon 127 statistical analysis

2.3.20 Miscellaneous software

3 Mutation screening of DYNC1H1

3.1 Introduction

3.2 Determining the genomic structure of the DYNC1H1 locus in silico

3.2.1 Information available on the genomic organisation of DYNC1H1 in 2003

3.2.2 Fine-scale localisation of DYNC1H1 to the human genome, chromosome 14 assembly

3.2.3 Identifying DYNC1H1 exons and introns

3.2.4 Identifying the 5' and 3' UTRs of exons 1 and 78
3.3 Mutation screening of \textit{DYNC1H1} exons 8, 13 and 14, and intron 13 .......... 95
3.3.1 Prioritising regions of \textit{DYNC1H1} to screen ............................................ 95
3.3.2 Primer design and sequencing of exons 8 and 13 ........................................ 95
3.3.3 Patient samples screened for \textit{DYNC1H1} mutations .................................... 95
3.3.4 Mutation screening of exon 8 ..................................................................... 96
3.3.5 Mutation screening of exons 13, 14 and intron 13 ..................................... 97
3.4 Discussion .......................................................................................................... 98
3.4.1 Updated SNP information ............................................................................ 99
3.4.2 Implications and explanations of a negative screen ...................................... 100

4 Cytoplasmic dynein 1 heavy chain 1 association study ................................... 101
4.1 Introduction ....................................................................................................... 101
4.1.1 Mutation screening of large genes can be problematic ................................. 101
4.2 \textit{DYNC1H1} association study design ............................................................. 101
4.2.1 Evolving resources and methodologies in LD-based association studies influence study design ............................................................. 101
4.2.2 Three phased study design ........................................................................... 102
4.2.3 Estimating LD to assess tagging approach feasibility .................................. 102
4.2.4 Study power and estimating required sample size ....................................... 103
4.3 Phase I - SNP discovery .................................................................................... 104
4.3.1 \textit{In silico} SNP ascertainment ..................................................................... 104
4.3.2 SNP validation and discovery ....................................................................... 105
4.3.3 Minor allele frequencies .............................................................................. 105
4.4 Genotyping validated SNPs ............................................................................. 108
4.4.1 Quality control - assessing genotyping accuracy ........................................ 109
4.4.2 Quality control – Mendelian inheritance and Hardy-Weinberg equilibrium ...... 109
4.5 The haplotype structure of \textit{DYNC1H1} ............................................................... 110
4.6 Linkage disequilibrium patterns ...................................................................... 112
4.7 Phase II - SNP tagging .................................................................................... 114
4.7.1 Assessing tSNPs – “SNP-dropping” ............................................................. 117
4.7.2 Restriction endonuclease assays for rs2251644 and rs941793 ..................... 119
4.8 Phase III - Testing in sporadic ALS cases and controls ................................... 121
4.8.1 Association study samples ......................................................................... 121
4.8.2 Tagging SNP rs2251644 and rs941973 genotyping in cases and controls ...... 121
4.8.2.1 Association study genotyping accuracy ................................................... 122
4.8.2.2 Hardy-Weinberg disequilibrium in control samples ......................... 122
4.8.2.3 A sampling bias? .................................................................................. 123
4.8.2.4 Hardy-Weinberg equilibrium at genes independent of \textit{DYNC1H1} .......... 123
4.8.3 A comparison of tSNP genotype frequencies between cases and controls ..... 123
4.8.4 Comparison of tSNP haplotypes between cases and controls .................................................. 125

4.9 Evolutionary analysis of the DYNCIHI locus ........................................................................ 126
4.9.1 SNP ascertainment bias and impact on measurements of selection .................................. 126
4.9.2 DYNCIHI chimpanzee genotyping ................................................................................ 127
4.9.3 Genetic variation of DYNCIHI in additional populations .................................................. 128
  4.9.3.1 Allele frequency comparisons ..................................................................................... 128
  4.9.3.2 Genetic differentiation ............................................................................................... 129
4.9.4 Haplotype and LD comparisons between three worldwide populations ......................... 130
4.9.5 Natural selection at DYNCIHI? ....................................................................................... 134

4.10 Discussion ............................................................................................................................ 134
4.10.1 DYNCIHI association study caveats .............................................................................. 134
  4.10.1.1 An under-powered study? ......................................................................................... 134
  4.10.1.2 Changes in study design? .......................................................................................... 135
  4.10.1.2.1 Two-stage SNP ascertainment ....................................................................... 135
  4.10.1.2.2 What is the "right" SNP density and allele frequency? ..................................... 136
  4.10.1.2.3 Tag SNP selection ............................................................................................... 137
4.10.2 Case sample heterogeneity ............................................................................................ 138
4.10.3 The availability of HapMap data .................................................................................. 138

5 Genetic analysis of the cytoplasmic dynein subunit families ................................................. 140
5.1 Introduction ......................................................................................................................... 140
5.1.1 Beyond the cytoplasmic dynein heavy chain – the need to consider pathways ............. 140
5.1.2 The cytoplasmic dynein subunits and a need for clarity .................................................. 141
5.1.3 Standardising nomenclature – why agreeing a name for each gene is important ........... 141

5.2 Clarifying mouse and human cytoplasmic dynein subunit nomenclature, genominc locations and accession numbers ................................................................. 142
5.2.1 Collating human and mouse cytoplasmic dynein subunit synonyms ............................. 143
5.2.2 The many names of cytoplasmic dynein 1 heavy chain ................................................ 144
5.2.3 Clarifying subunit mapping positions and identifying paralogs .................................... 148
  5.2.3.1 Example 1: Cytoplasmic dynein light chain ............................................................ 149
  5.2.3.2 Example 2: Cytoplasmic dynein 1 heavy chain ......................................................... 153
5.2.4 Cytoplasmic dynein subunit accession numbers .............................................................. 154

5.3 Identifying cytoplasmic dynein orthologs .......................................................................... 154
5.3.1 Cytoplasmic dynein heavy chain gene family (DYN1H1, DYN2H1) .............................. 154
5.3.2 Cytoplasmic dynein intermediate chain gene family (DYN111, DYN112) ................. 155
5.3.3 Cytoplasmic dynein light intermediate chain family (DYN1L11, DYN1L12, DYN2L11) .... 156
5.3.4 Cytoplasmic dynein light chain Tctex1-family (DYNLT1, DYNLT3) ............................ 156
5.3.5 Cytoplasmic dynein light chain Roadblock family (DYNLRB1, DYNLRB2) .............. 158
6.10 Discussion....................................................................................................................... 197

6.11 Codon 129 and kuru susceptibility............................................................................ 197

6.11.1 Codon 129 independent susceptibility? .............................................................. 198
6.11.2 PRNP and kuru-mediated signatures of selection ........................................... 199
6.11.3 Additional susceptibility loci - PRNP G127V ..................................................... 200
6.11.4 Future study of PNG and kuru samples ............................................................. 200

6.11.4.1 The effect of whole-genome amplification of low quality material on whole- genome genotyping platforms ................................................................. 201
6.11.4.2 Ascertainment bias ......................................................................................... 201

7 PRNP copy number polymorphisms and prion disease susceptibility ................... 202

7.1 Introduction................................................................................................................... 202
7.1.1 Copy number polymorphisms in neurodegenerative disease ......................... 202
7.1.2 Variability in PRNP expression and disease susceptibility .............................. 203
7.1.3 Copy number polymorphisms in sporadic CJD? .............................................. 204
7.1.4 PRNP copy number hypothesis .......................................................................... 204

7.2 Quantitative Real Time PCR probe and primer design for PRNP. .................... 204

7.3 Validating qPCR probe efficiencies.......................................................................... 205

7.4 Copy number variation in elderly Fore females .................................................... 207

7.5 Copy number variation in sporadic CJD samples .................................................. 209

7.6 Discussion....................................................................................................................... 210

7.6.1 Hypothesis 1 - PRNP deletion mediates protection against kuru ......... 210
7.6.2 Hypothesis 2 - PRNP duplication mediates susceptibility of codon 129 heterozygotes to sCJD ................................................................. 211

7.6.3 Limitations of this approach and future work ...................................................... 211
7.6.3.1 PRNP gene coverage ................................................................................. 211
7.6.3.2 PRNP copy number and kuru .................................................................... 212
7.6.3.3 A test for somatic mutation ...................................................................... 212
7.6.3.4 A test for other diseases associated with PRNP .................................. 212

8 General discussion.......................................................................................................... 213

8.1 Evolving tools and techniques................................................................................... 213

8.2 DYN1IHI and ALS ...................................................................................................... 213

8.2.1 Is there a role for DYN1IHI in ALS? ................................................................. 214
8.2.1.1 The dynein-dynactin complex and ALS .................................................... 215
8.2.2 Are association studies the right approach for complex neurodegenerative diseases? ................................................................. 216

8.2.2.1 Sample size ......................................................................................... 217
8.2.2.2 Statistical analyses .............................................................................. 217
8.2.2.3 Rare diseases ..................................................................................... 218
8.3 Identifying kuru susceptibility loci ................................................................. 218

8.3.1 Signatures of selection at PRNP ................................................................. 220

8.3.1.1 Evolutionary studies require good epidemiological evidence .......... 220

8.3.1.2 Was evidence for selection seen at PRNP? ........................................... 221

8.3.1.3 PRNP selection study caveats ................................................................. 221

8.3.2 G127V is an additional PRNP susceptibility factor ................................. 222

8.3.3 Is selection a valid method for mapping loci in all neurodegenerative diseases?.. 223

8.4 Future directions for complex genetic analyses of neurodegenerative disease.. 224

9 References ............................................................................................................. 226

10 Appendices .......................................................................................................... 267

11 Publications ......................................................................................................... 269
List of figures

Figure 1.1 Allelic variants showing Mendelian inheritance discovered from 1988 to 2003........................21
Figure 1.2 The threshold liability model..............................................................................................24
Figure 1.3 The spectrum of human disease..........................................................................................24
Figure 1.4 Aggressive tagging of variation linked to six loci...............................................................29
Figure 1.5 Association study sample size predictions............................................................................30
Figure 1.6 Case-control selection using the threshold liability model................................................31
Figure 1.7 SNP submissions to dbSNP from over 5 years from 2000..................................................34
Figure 1.8 The effects of selection on the distribution of genetic variation........................................37
Figure 1.9 BSE and vCJD cases in the UK...........................................................................................55
Figure 2.1 Overview of the QIAamp 96 DNA Blood Mini kit procedure.............................................70
Figure 2.2 An overview of the Affymetrix GeneChip 250k NspI protocol...........................................76
Figure 2.3 The principle of haplotype tagging.......................................................................................81
Figure 3.1 Scheme representing the reverse mapping of transcribed sequences onto genomic sequence 87
Figure 3.2 Fine-scale localisation of DYNC1H1 on chromosome 14 (Builds 33 to 36.1).....................88
Figure 3.3 Exon sizes across the DYNC1H1 genomic locus................................................................89
Figure 3.4 Consensus sequences at DYNC1H1 splice junctions..........................................................90
Figure 3.5 DYNC1H1 UTR sequences and conserved motifs...............................................................91
Figure 3.6 Exon 8 A/G synonymous SNP............................................................................................96
Figure 3.7 Intron 13 A/G SNP............................................................................................................97
Figure 4.1 DYNC1H1 association study power and related sample size for varying $D^t$ values...........104
Figure 4.2 SNP rs13749 electropherograms and the redundancy of the threonine codon.....................106
Figure 4.3 DYNC1H1 SNP map.......................................................................................................108
Figure 4.4 Phase-ambiguous SNPs in CEPH trios.............................................................................111
Figure 4.5 Linkage disequilibrium ($D^t$) between the 16 SNPs spanning DYNC1H1........................112
Figure 4.6 Linkage disequilibrium ($r^2$) between the 16 SNPs spanning DYNC1H1..........................113
Figure 4.7 Average locus haplotype $r^2$ performance of all tSNP sets sizes from $H=1$ to $H=5$..........115
Figure 4.8 Performance of "best" tSNP sets of increasing size $H$ chosen using two criteria..............116
Figure 4.9 Optimal tSNP sets of size $H$ and their performance against all loci...............................117
Figure 4.10 SNP dropping performance for "best" tSNP sets of varying size.....................................118
Figure 4.11 Restriction digests for tSNPs rs2251644 and rs941793.....................................................120
Figure 4.12 A comparison of DYNC1H1 SNP allele frequencies between 3 populations..................129
Figure 4.13 Global and pairwise $F_{ST}$ comparisons........................................................................130
Figure 4.14 Linkage disequilibrium decay across DYNC1H1 in 3 populations................................133
Figure 4.15 Site frequency spectra for empirical data with varying ascertainment bias.......................136
Figure 4.16 Linkage disequilibrium ($r^2$) across DYNC1H1 identified using HapMap data............139
Figure 5.1 The mammalian cytoplasmic dynein complexes.............................................................141
Figure 5.2 Mouse and human DYNC1H1 synonyms from the Entrez database at NCBI.....................145
Figure 5.3 Schematic representation of the putative DYNLL1 locus at 12q24.31...............................152
Figure 5.4 In silico translation of artificially spliced DYNLL1 alignments...........................................152
Figure 5.5  Cladogram of the relationships between human dynein heavy chains and the hypothesised heavy chain FLJ46675 ................................................................. 153
Figure 5.6  Protein-based phylogenies of the cytoplasmic dynein heavy chain, intermediate chain and light intermediate chain families ................................................................. 157
Figure 5.7  Cytoplasmic dynein light chain (LC8) family protein alignments ........................................ 158
Figure 5.8  Protein-based phylogenies of the cytoplasmic dynein light chain family ................................ 159
Figure 6.1  Distribution of PRNP codon 129 genotypes by age in 147 kuru cases .................................. 165
Figure 6.2  Sex specific distribution of PRNP codon 129 genotypes by age at collection in 146 kuru cases .................................................................................................................. 166
Figure 6.3  Observed and expected PRNP codon 129 genotype frequencies for elderly Fore women repeatedly exposed to kuru .......................................................................................... 167
Figure 6.4  Schematic map of the Eastern Highlands of Papua New Guinea .......................................... 168
Figure 6.5  Heterozygosity at PRNP codon 129 stratified by sex and age ............................................. 172
Figure 6.6  Worldwide PRNP codon 129 valine allele frequency .......................................................... 175
Figure 6.7  Variation of PRNP codon 129 valine allele frequency worldwide ....................................... 176
Figure 6.8  PRNP codon 129 valine frequencies across the Eastern Highlands ...................................... 177
Figure 6.9  Microsatellite diversity upstream and downstream of the PRNP codon 129M allele .......... 179
Figure 6.10 Microsatellite diversity upstream and downstream of the PRNP codon 129V allele ........... 179
Figure 6.11 Microsatellite diversity upstream of the PRNP the codon 129M allele ............................... 180
Figure 6.12 Microsatellite diversity upstream of the PRNP codon 129V allele ................................. 181
Figure 6.13 An increasing cline in codon 129 valine frequency within Papua New Guinea ................ 183
Figure 6.14 The kuru region divided into three zones of increasing exposure ..................................... 184
Figure 6.15 Size of 127V-linked haplotype compared with the same haplotype on 127G alleles ............ 187
Figure 6.16 PCR of 7 South Fore multiple kuru exposure samples and 7 kuru samples ....................... 189
Figure 6.17 Fragmentation of ligated PCR products ............................................................................. 190
Figure 6.18 Dynamic Model scatter plot for calling genotypes .......................................................... 191
Figure 6.19 GeneChip NspI hybridisation array images for sample PDG7470 ....................................... 191
Figure 6.20 SNP gains and data reliability at various rank score thresholds ....................................... 192
Figure 6.21 Pairwise LD in 7 elderly multiple kuru-exposed samples at different thresholds .............. 194
Figure 6.22 The effect of sample size on genome-wide pairwise LD .................................................. 194
Figure 6.23 Linkage disequilibrium decay over distance ..................................................................... 195
Figure 6.24 Inflation of highly correlated ($r^2$>0.8) LD comparisons in small sized samples ............. 196
Figure 6.25 Decay of linkage disequilibrium of highly correlated ($r^2$>0.8) alleles with distance .......... 197
Figure 7.1 PRNP and ß-Actin quantitative real time PCR probes and primers ..................................... 205
Figure 7.2 Relative efficiencies for PRNP and PRNP_{129M} quantitative PCR probes ..................... 206
Figure 7.3 Average $\Delta C_T$ differences between PRNP codon 129 MM and MV CEPH samples ........ 206
Figure 7.4 PRNP 129M copy number analysis of 25 Fore women over 50 years old repeatedly exposed to kuru .................................................................................................................. 208
Figure 7.5 PRNP copy number analysis of Fore women over 50 years old repeatedly exposed to kuru 208
Figure 7.6 PRNP copy number analysis of sporadic CJD samples .................................................... 209

13
List of tables

Table 1.1 Mendelian genes identified with mutations in non-Mendelian diseases .................22
Table 1.2 Reports of selection in the human lineage .................................................................35
Table 1.3 Familial ALS genes and loci ......................................................................................43
Table 1.4 ALS susceptibility loci ...............................................................................................48
Table 1.5 Incubation times following intracerebral inoculation ...............................................58
Table 3.1 DYNCIHI genomic organisation ..............................................................................94
Table 3.2 Summary of DYNCIHI exon 8 mutation screen, in motor neuron disease patients and controls .........................................................................................................................97
Table 3.3 Summary of DYNCIHI mutation screening of exons 13, 14 and intron 14 in motor neuron disease patients and controls .................................................................98
Table 4.1 Marker information for sequence tagged sites flanking DYNCIHI .........................103
Table 4.2 SNPs identified and validated in a discovery panel of 16 CEPH individuals ..........107
Table 4.3 SNP frequency comparison between 32 and 128 chromosome discovery panels ....109
Table 4.4 The haplotype structure of DYNCIHI in CEPH .........................................................110
Table 4.5 Genotype frequencies of tSNPs rs2251644 and rs941793 in sporadic ALS cases and matched controls .........................................................................121
Table 4.6 DYNCIHI and VEGF control genotypes in Hardy-Weinberg equilibrium ..........125
Table 4.7 DYNCIHI tSNP haplotype frequencies in sporadic ALS cases and matched controls .................126
Table 4.8 Ancestral chimpanzee and common human DYNCIHI haplotypes .................127
Table 4.9 Northern European, Japanese, Cameroonian and chimpanzee DYNCIHI haplotypes 132
Table 5.1 NCBI LocusLink human and mouse genes containing the root symbol ‘dhc’ .......143
Table 5.2 Human and mouse cytoplasmic dynein genes and map positions .........................148
Table 5.3 DYNLL1 megaBLAST alignments against the human genome .........................151
Table 6.1 Hardy-Weinberg analysis of the Eastern Highland linguistic groups ...............171
Table 6.2 Genotypes of kuru patients and age-stratified healthy population controls .......185
Table 8.1 Consistent associations with complex disease .........................................................216
Table 8.2 Effect of differing statistical significance levels on sample size ..........................217

Appendix 1. Full genotypes for 60 SNPs across DYNCIHI ....................................................267
Appendix 2. DYNCIHI informative primers ..............................................................268
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>μg</td>
<td>Micro grammes</td>
</tr>
<tr>
<td>μl</td>
<td>Micro litres</td>
</tr>
<tr>
<td>AD</td>
<td>Autosomal dominant</td>
</tr>
<tr>
<td>AD</td>
<td>Alzheimer's disease</td>
</tr>
<tr>
<td>ALS</td>
<td>Amyotrophic lateral sclerosis</td>
</tr>
<tr>
<td>ALS2</td>
<td>Alsin</td>
</tr>
<tr>
<td>AMD</td>
<td>Age-related macular degeneration</td>
</tr>
<tr>
<td>ANG</td>
<td>Angiogenin</td>
</tr>
<tr>
<td>APEX</td>
<td>Apurinic endonuclease</td>
</tr>
<tr>
<td>APOE</td>
<td>Apolipoprotein E</td>
</tr>
<tr>
<td>APP</td>
<td>Amyloid precursor protein</td>
</tr>
<tr>
<td>AR</td>
<td>Autosomal recessive</td>
</tr>
<tr>
<td>AR</td>
<td>Androgen receptor</td>
</tr>
<tr>
<td>AT3</td>
<td>Ataxin 3</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine triphosphate</td>
</tr>
<tr>
<td>BAC</td>
<td>Bacterial artificial chromosome</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>BSA</td>
<td>Bovine serum albumin</td>
</tr>
<tr>
<td>BSE</td>
<td>Bovine spongiform encephalopathy</td>
</tr>
<tr>
<td>CCA</td>
<td>Congenital contractural arachnodactyly</td>
</tr>
<tr>
<td>CCR5</td>
<td>Chemokine (C-C motif) receptor 5</td>
</tr>
<tr>
<td>CDCV</td>
<td>Common-disease/common-variant</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CEPH</td>
<td>Centre d'étude du polymorphisme humain</td>
</tr>
<tr>
<td>CF</td>
<td>Cystic fibrosis</td>
</tr>
<tr>
<td>CFTR</td>
<td>Cystic fibrosis transmembrane regulator</td>
</tr>
<tr>
<td>CHMP2B</td>
<td>Chromatin modifying protein/Charged multivesicular body protein 2B</td>
</tr>
<tr>
<td>CJD</td>
<td>Creutzfeldt-Jakob disease</td>
</tr>
<tr>
<td>cM</td>
<td>CentiMorgan</td>
</tr>
<tr>
<td>CNP</td>
<td>Copy number polymorphism</td>
</tr>
<tr>
<td>CNS</td>
<td>Central nervous system</td>
</tr>
<tr>
<td>CNTF</td>
<td>Ciliary neurotrophic factor</td>
</tr>
<tr>
<td>COX1</td>
<td>Cyclooxygenase 1</td>
</tr>
<tr>
<td>Cra1</td>
<td>Cramping 1</td>
</tr>
<tr>
<td>CT</td>
<td>Threshold cycle</td>
</tr>
<tr>
<td>CYP2D6</td>
<td>Cytochrome p450, subfamily IID, polypeptide 6</td>
</tr>
<tr>
<td>DAPI</td>
<td>4',6-Diamidino-2-phenylindole</td>
</tr>
<tr>
<td>dATP</td>
<td>Deoxyadenosine triphosphate</td>
</tr>
<tr>
<td>DCTN1</td>
<td>Dynactin 1</td>
</tr>
<tr>
<td>dCTP</td>
<td>Deoxycytosine triphosphate</td>
</tr>
<tr>
<td>ddH20</td>
<td>Purified deionised water</td>
</tr>
<tr>
<td>df</td>
<td>Degrees of freedom</td>
</tr>
<tr>
<td>dGTP</td>
<td>Deoxyguanosine triphosphate</td>
</tr>
<tr>
<td>DLDH</td>
<td>Dementia lacking distinct histopathological features</td>
</tr>
<tr>
<td>DM</td>
<td>Dynamic mapping</td>
</tr>
<tr>
<td>DMSO</td>
<td>Dimethyl sulphoxide</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>Deoxynucleotide triphosphate</td>
</tr>
<tr>
<td>DPE</td>
<td>Downstream promoter element</td>
</tr>
<tr>
<td>DS</td>
<td>Down syndrome</td>
</tr>
<tr>
<td>dTTP</td>
<td>Deoxythymidine triphosphate</td>
</tr>
<tr>
<td>DYNC1H1</td>
<td>Cytoplasmic 1 dynein heavy chain 1</td>
</tr>
<tr>
<td>EAAT2</td>
<td>Excitatory amino acid transporter 2</td>
</tr>
<tr>
<td>EATDT</td>
<td>Extended allelic transmission disequilibrium test</td>
</tr>
<tr>
<td>EBV</td>
<td>Epstein bar virus</td>
</tr>
<tr>
<td>ECACC</td>
<td>European collection of cell cultures</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>EGF</td>
<td>Epidermal growth factor</td>
</tr>
<tr>
<td>EGFR</td>
<td>Epidermal growth factor receptor</td>
</tr>
</tbody>
</table>
El Exposure index
EM Expectation maximisation
ENCODE Encyclopaedia Of DNA Elements
ER Endoplasmic reticulum
EST Expressed sequence tag
EtBr Ethidium bromide
EtOH Ethanol
FALS Familial Amyotrophic Lateral Sclerosis
FAM 6-carboxy-fluorescein
fcJD Familial Creutzfeldt-Jakob disease
FDR False discovery rate
FFI Fatal familial insomnia
FISH Fluorescent in situ hybridization technique
FLD Frontal lobe degeneration of non-Alzheimer type
FOX2 Forkhead box P2
FTD Frontotemporal dementia
G6PD Glucose 6 phosphate dehydrogenase
GFP Green fluorescent protein
GRR Genotype relative risk
GSS Gerstmann-Straussler-Scheinker
GTP Guanosine-Tri-Phosphate
h^2 Heritability
HapMap International Haplotype Map project
Hbb Haemoglobin beta
HD Hungtington's disease
HEX Hexachloro-6-carboxyfluorescein
HEXA Hexosaminidase A
HFE Haemochromatosis
HGNC Human Genome Nomenclature Committee
HIV Human immunodeficiency virus
HKA Hudson-Kreitman-Aguadé
HKA Hudson–Kreitman–Aguade test
HLA Human leukocyte antigen
HSP Hereditary spastic paraplegia
htSNP Haplotype tagging SNP
HUGO Human Genome Organisation
HWE Hardy-Weinberg equilibrium
IARS2 Mitochondrial isoleucine tRNA synthetase
IBD Inflammatory bowel disease
IBM Inclusion body myositis
IFT Intraflagellar transport
Inr Transcription initiator
kb Kilobases
kDa Kilo Daltons
kV Kilo volts
LCT Lasctase
LD Linkage disequilibrium
LIF Leukaemia inhibitory factor
LINE Long interspersed nuclear element
LIS1 Lissencephaly 1
Loa Legs at odd angles
LOD Log-of-odds
LRH Long-range haplotype
M Molar
MAF Minor allele frequency
MAO-B Monoamine oxidase B
MAPT Microtubule-associated protein tau
MgSO4 Magnesium sulphate
MHC Major Histocompatibility Complex
Mito Mitochondrial DNA deletions
MMP3 Matrix metalloproteinase-3
MND Motor Neuron Disease
mol Moles
MRC Medical Research Council
MRCA Most recent common ancestor
MRI Magnetic resonance imaging
mRNA Messenger ribonucleic acid
MT Microtubule
mtDNA Mitochondrial DNA
MVB Multivesicular body
MYCN Myelocytomatosis viral related oncogene, neuroblastoma derived
NaAce Sodium acetate
NAIP Neuronal apoptosis inhibitory protein
NCBI National Centre for Biotechnology Information
ND2 Subunit 2 of mitochondrial NADH dehydrogenase
NEFH Neurofilament, heavy chain
NFT Neurofibril tangles
NGF Nerve growth factor
NPC Niemann-Pick type C
nt Nucleotide
OA Ocular albinism
OMIM Online Mendelian Inheritance in Man
OPN1LW Opsin 1 long wave
OPRI Octapeptide repeat insertion
ORF Open reading frame
PBS Phosphate buffered saline
PCR Polymerase chain reaction
PD Parkinson's disease
PDR Polymorphism discovery resource
Pers. comm. Personal communication
PET Positron emission tomography
PFA Paraformaldehyde
PKU Phenylketonuria
PLP Proteolipid protein
PMD Pelizaeus-Merzbacher disease
PNG Papua New Guinea
PNGIMR Papua New Guinea Institute of Medical Research
PRKN Parkin
PRND Dopple: downstream prion-like gene
PRNP Prion protein gene
PrP Prion protein
PrP<sup>C</sup> Prion protein (normal cellular isoform)
PRPH Peripherin
PrP<sup>sc</sup> Prion protein (scrapie isoform)
PSEN1 Presenilin-1
PSP Progressive supranuclear palsy
QTL Quantitative trait loci
QTN Quantitative trait nucleotide
rcf Relative centrifugal force
RFLPs Restriction fragment length polymorphisms
RML Rocky Mountain Laboratories
RNA Ribonucleic acid
RNAi RNA interference
rRNA Ribosomal RNA
RT-PCR Reverse transcriptase-polymerase chain reaction
SALS Sporadic amyotrophic lateral sclerosis
SBMA Spinobulbar muscular atrophy
SCA Spinocerebellar ataxia
sCJD Sporadic CJD
SCN9A Sodium channel, voltage-gated, type IX, alpha
SCNA α-synuclein
SDS Sodium dodecyl sulfate
SEM Standard error of mean
SETX Senataxin
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINE</td>
<td>Short interspersed repeat element</td>
</tr>
<tr>
<td>SMA</td>
<td>Spinal muscular atrophy</td>
</tr>
<tr>
<td>SMN1/2</td>
<td>Survival of motor neuron 1/2</td>
</tr>
<tr>
<td>SNCA</td>
<td>Alpha synuclein</td>
</tr>
<tr>
<td>SNCG</td>
<td>Persyn</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SOD1</td>
<td>Cu/Zn superoxide dismutase 1</td>
</tr>
<tr>
<td>SOD2</td>
<td>Manganese superoxide dismutase</td>
</tr>
<tr>
<td>SSC</td>
<td>Sodium chloride sodium citrate</td>
</tr>
<tr>
<td>STR</td>
<td>Single tandem repeat</td>
</tr>
<tr>
<td>STS</td>
<td>Sequenced tagged site</td>
</tr>
<tr>
<td>TBE</td>
<td>Tris-borate EDTA</td>
</tr>
<tr>
<td>TDT</td>
<td>Transmission disequilibrium test</td>
</tr>
<tr>
<td>TET</td>
<td>Tetrachloro-6-carboxy-fluorescein</td>
</tr>
<tr>
<td>Tm</td>
<td>Melting temperature</td>
</tr>
<tr>
<td>TNFSF</td>
<td>Tumour necrosis factor superfamily</td>
</tr>
<tr>
<td>Tris</td>
<td>2,3-dibromopropyl phosphate</td>
</tr>
<tr>
<td>TrkA</td>
<td>Tyrosine kinase</td>
</tr>
<tr>
<td>tSNP</td>
<td>Tagging SNP</td>
</tr>
<tr>
<td>TSS</td>
<td>Translational start site</td>
</tr>
<tr>
<td>UCSC</td>
<td>University California Santa Cruz</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>USA</td>
<td>United States of America</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>UV</td>
<td>Ultraviolet</td>
</tr>
<tr>
<td>VAPB</td>
<td>Vesicle-associated membrane protein-associated protein B</td>
</tr>
<tr>
<td>vCJD</td>
<td>Variant CJD</td>
</tr>
<tr>
<td>VCP</td>
<td>Valosin-containing protein</td>
</tr>
<tr>
<td>VDR</td>
<td>Vitamin D receptor</td>
</tr>
<tr>
<td>VEGF</td>
<td>Vascular endothelial growth factor</td>
</tr>
<tr>
<td>WGA</td>
<td>Whole genome association</td>
</tr>
<tr>
<td>YAC</td>
<td>Yeast artificial chromosome</td>
</tr>
<tr>
<td>(\lambda_r)</td>
<td>Relative risk</td>
</tr>
<tr>
<td>(\lambda_s)</td>
<td>Sibling relative risk</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Neurodegenerative diseases

Neurodegenerative diseases are a range of fatal disorders in which the disease pathogenesis results in the progressive degeneration of the central and/or the peripheral nervous systems. These diseases, which include Parkinson's disease (PD), Alzheimer's disease (AD), frontotemporal dementia (FTD), amyotrophic lateral sclerosis (ALS) and prion diseases, currently affect approximately 2% of the population in the developed world (Hardy et al., 2006a) and show a propensity to occur with increasing age. As average life expectancy in the developed and developing world continues to increase, so too will the worldwide prevalence of neurodegenerative diseases (Kondo, 1996): conservative models of population growth predict that by 2025, over 1 billion people will be aged greater than 60 years old. This shift will have profound implications on public health and therefore our ability to dissect the aetiology of neurodegenerative disease must match this challenge. Understanding the genetic aetiology of neurodegenerative diseases will improve disease diagnosis, assist genetic counselling in familial cases, inform on an individual's susceptibility to disease, provide targets for therapeutics and inform future public health policy. Until recently, our ability to identify the genetic determinants of neurodegenerative disease has been restricted to rare Mendelian forms. These gene discoveries have illuminated only a small fraction of the causes of more common, apparently sporadic, complex forms of disease. However, recent developments in novel and rapidly advancing study designs, experimental and analysis techniques have begun to shed light on complex neurodegenerative diseases.

1.1.1 Common features of neurodegenerative diseases

In the last decade, several converging lines of investigation have revealed common epidemiological patterns, disease aetiologies and pathogenic mechanisms, underlying neurodegenerative diseases. These shared features suggest that mutual genetic mechanisms may also exist amongst these diseases which may be illuminated using similar techniques. Common features include the abnormal function of the following: protein aggregation and deposition (Skovronsky et al., 2006; Taylor et al., 2002), mitochondrial or oxidative phosphorylation activity (Manfredi et al., 2000; Schon et al., 2003), axonal transport (Roy et al., 2005), endosomal and endocytic function (Bronfman et al., 2007; Nixon, 2005) and the ubiquitin-proteosome system (Ciechanover et al., 2003; Petrucelli et al., 2004; Ross et al., 2004).

In addition many neurodegenerative diseases share common features of genetic epidemiology such as the existence of both familial and sporadic or idiopathic forms of disease. At the same time, familial forms are often rare and sporadic forms of neurodegenerative diseases are often common. The rare exception to this observation are the triplet repeat disorders such as
Huntington’s disease and the spinocerebellar ataxias, of which the overwhelming majority are familial, or apparently sporadic but due to cryptic familial inheritance, non-paternity (Davis et al., 1994) or de novo repeat expansions (Bozza et al., 1995; Davis et al., 1994; Durr et al., 1995; Futamura et al., 1998; Mandich et al., 1996; Myers et al., 1993). In addition, common to all neurodegenerative diseases is the familial aggregation of disease, which may imply a genetic component to even sporadic disease.

1.1.2 Related genetic mechanisms

Beyond the clinical, epidemiological and pathological features of neurodegenerative diseases, different neurodegenerative disorders may also share a common genetic aetiology: the genetic determinant of one neurodegenerative disease may be a causal or risk factor in another disease. A commonly cited example is ε4 allele of the apolipoprotein E gene (APOE), which is a susceptibility factor for several diseases including: age of onset and dementia in PD (Feldman et al., 2006; Huang et al., 2006; Li et al., 2002; Li et al., 2004; Parsian et al., 2002; Tang et al., 2002a), age of onset in ALS (Li et al., 2004) and age of onset in FTD (Boccardi et al., 2004; Borroni et al., 2005). In addition, the gene charged multivesicular body protein 2B (CHMP2B), which has been shown to be causative for FTD in both a Danish family and an isolated FTD individual (Skibinski et al., 2005) has also been identified in two cases of ALS (Parkinson et al., 2006). Examples such as these support the view that neurodegenerative diseases, such as FTD and ALS, may have common genetic aetiologies. In addition familial clustering of diseases such as ALS and Creutzfeldt-Jacob disease (CJD) with other neurodegenerative diseases (Majoor-Krakauer et al., 1994; van Duijn et al., 1998) is indicative of a common cause to these diseases which can identify risk factors (Riemenschneider et al., 2004).

1.2 Gene identification in Mendelian forms of neurodegenerative diseases

To date, the overwhelming majority of genes causal for human disease have been identified through linkage analysis. These analyses have commonly required large pedigrees which display clear Mendelian inheritance of a phenotype. Essentially, linkage analysis relies on the co-segregation of a disease causing allele and adjacent DNA markers with the disease phenotype, within a family. Chromosomal segments that do not influence disease segregate randomly. A DNA segment that carries the disease-causing mutation will be shared among affected family members more often than would be predicted by chance and a likelihood score (log-of-odds ratio; LOD) is applied (LOD>3.6 is generally considered the criterion indicating linkage between a genome segment and disease).
Figure 1.1 Allelic variants showing Mendelian inheritance discovered from 1988 to 2003
Graph depicting the number of pathogenic mutations causing monogenic disorders, recorded in OMIM over 15 years (Modified from Antonarakis et al., 2006).

As Figure 1.1 illustrates, the number of single genes responsible for Mendelian disorders, identified and catalogued in the Online Mendelian Inheritance in Man (OMIM) database increased rapidly between 1988 and 2001. The key to success of linkage analysis has been the near-complete penetrance of simple traits; affected individuals are easily identified and always possess a genetic mutation, which in turn has permitted the collection of the large families required for linkage analysis. However, despite the advances linkage studies have provided in our understanding of the genetic basis of human disease, familial diseases continue to represent the minority of cases in almost all forms of disease. In fact, the majority of human diseases do not demonstrate the characteristics of simple monogenic Mendelian forms and until recently, the accepted dogma has regarded these non-Mendelian forms of disease as sporadic with possibly no genetic component. However, there is a large body of evidence that non-Mendelian diseases have a genetic component and are complex, i.e. are influenced by multiple genes, which interact with each other and with the environment.

1.3 The genetic component of non-Mendelian diseases
Diseases that do not demonstrate Mendelian inheritance may still possess a genetic component.

1.3.1 Familial genetics inform susceptibility loci in sporadic disease
Rare monogenic familial forms of several neurodegenerative diseases have informed our understanding of the genetic aetiology of sporadic disease. Causative genes identified in familial studies of AD, PD, ALS, FTD and CJD for example, have been found to be mutated in a minority of non-Mendelian cases, indicating that at least some sporadic cases have a genetic component (Table 1.1).
### Table 1.1 Mendelian genes identified with mutations in non-Mendelian diseases

Examples of genes from four Mendelian forms of neurodegenerative diseases that have been implicated in complex forms of the same disease. Maximum of two loci are shown for Alzheimer’s disease (AD), Parkinson’s disease (PD), amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). References are not exhaustive.

#### 1.3.2 Familial clustering of disease

The recurrence of an apparently sporadic disease in a family more frequently than could be predicted by population prevalence also provides evidence of a genetic component to non-Mendelian disease. This familial clustering of disease, can be measured by the relative recurrence risk ($R_{\text{rel}}$) - the risk to relative $R$ of an affected proband compared with the population risk. Familial clustering can be interpreted in several ways: disease can be due to shared genetic susceptibility, shared environmental exposure(s), or both. However, familial clustering may be due to other factors such as incomplete penetrance, as has been shown in ALS (Robberecht et al., 1996), or due to the transmission of environmental factors, such as viruses, within families which can mimic or confound true genetic susceptibilities (Szmuness et al., 1973).

Another measure of the contribution of inherited factors to disease risk is heritability ($h^2$). Estimated from family or twin studies, heritability signifies the fraction of the population variation that can be explained by genetic factors working together in an additive fashion. For diseases such as late onset Alzheimer’s disease, heritability may be between 58% and 79% (Gatz et al., 2006).

Twin studies are of particular importance in the estimation of heritability in complex diseases (reviewed in MacGregor et al., 2000) as they exploit the unique degree of genetic sharing among the two types of twin pair: (i) monozygotic twins, who share a common set of genes, and (ii) dizygotic twins, who share on average ~50% of their genes. In addition, the use of twins can control for non-genetic factors, as they share aspects of their early and later environment and these features allow the population-level variation of traits and diseases to be separated into genetic, shared environmental and random environmental components. Conversely, for twins...
reared in different environments such as in adoption studies, variances due to environment and gene-environment interactions can also be assessed.

1.4 Complex diseases

Our understanding of complex diseases (also known as complex traits) is founded within the fascinating history of genetics and genetic ideas. In the interest of brevity, the full history of modern genetics and complex diseases is not discussed here (but an excellent account can be found in Strachan et al., 1999). However, to understand what complex diseases are and why the underlying genes have been difficult to elucidate, a brief review of complex genetics is required.

Historically, there have been two traditions within human genetics. The first followed the principles of heredity established by Gregor Mendel who identified that physical traits were inherited discretely. Under this tradition all traits are considered dichotomous (such as polydactyly – either you have an extra finger or you don’t) - the off-spring of a set of parents inherit characteristics from one parent or the other and not a blend of the two. This premise also formed the basis of a saltation (or non-gradual) theory of evolution, in which sudden phenotypic changes are seen from one generation to the next. The second tradition was based on the study of variation, such as that carried out by Francis Galton in the late 19th century who observed that human characteristics, such as weight, height and reaction time were quantitative, continuously variable characters. These continuous traits could not be reconciled using Mendelian genetics and did not explain gradual genetic variation (i.e. gradual evolution) seen in many organisms described in the theory of modern synthesis.

In 1918, both traditions were aligned by the mathematician and geneticist Ronald Aylmer Fisher. Fisher proposed that continuous traits are in fact governed by a number of loci, which are inherited in a Mendelian fashion, and that alleles at each locus contribute quantitatively to the overall phenotype. Any variable character that depends on the additive action of a large number of individually small independent causes will be distributed in a normal (Gaussian) distribution in the population, and under this polygenic theory of inheritance, as the number of quantitative trait loci (QTL) increase, the distribution looks increasingly like a Gaussian curve.

However, Fisher’s polygenic theory of inheritance could not account for dichotomous characteristics which are not inherited within families. This problem was solved in 1981 by Douglas Scott Falconer who extended the polygenic theory to discontinuous non-Mendelian characters by postulating that (i) an underlying polygenic liability to disease exists within a population and it is this which is continuously distributed and not phenotype and (ii) diseases manifest when the cumulative diseases susceptibility burden exceeds a threshold of liability. By inference, individuals whose genetic liability does not exceed the threshold value do not develop disease (Figure 1.2).
Figure 1.2 The threshold liability model
The model shows the distribution for liability to disease within a population. Individuals to the right of the threshold develop disease whereas those to the left are healthy.

For example, this can be illustrated by the developmental abnormality, cleft lip and palate. Every embryo has a certain susceptibility to cleft lip and palate which follows a Gaussian distribution in the population. During early development the embryonic palatal shelves must become horizontal and fuse together within a specific developmental window of time. The speed at which the plates meet and fuse is unimportant, as long as they meet before a critical developmental stage. If they fuse at or before the critical point then a normal palate forms, and if they do not fuse then a cleft palate results and thus, there is a natural threshold imposed on a continuous trait which is transformed into a dichotomous phenotype.

Figure 1.3 The spectrum of human disease
Few characters are purely Mendelian, purely polygenic or purely environmental. Most depend on some mix of major and minor genetic determinants, together with environmental influences. The mix of factors determining any given character could be represented by a point located somewhere within the triangle. (From Strachan et al., 1999)

The threshold liability model also accounts for familial clustering in non-Mendelian disease because liability genes will be more common amongst relatives and therefore all members of the family will be closer to the liability threshold. Not all family members will develop disease as several loci and environmental factors will contribute to phenotype. The challenge for human genetics now is to identify those genes that increase liability to develop disease (susceptibility genes) and those that modify, enhance or reduce a phenotype (modifier genes). In Mendelian diseases, the action of these susceptibility and modifier loci may manifest as reduced
penetrance, meaning that the disease has a reduced likelihood of occurring in individuals who carry the risk genotype. Susceptibility to human disease can therefore be considered a spectrum of genetic and environmental risk, which together can explain Mendelian disease, quantitative traits and complex diseases (Figure 1.3).

1.5 Gene identification strategies in complex diseases

1.5.1 Linkage analysis based approaches

Traditional linkage based approaches have worked well for Mendelian disorders (Antonarakis et al., 2006) and have been carried out successfully on a few complex diseases such as schizophrenia (Stefansson et al., 2002) and type 1 diabetes (Nistico et al., 1996). However, for most complex, polygenic traits, linkage analyses of extended pedigrees, siblings or parent and child trios, have achieved only limited success: Altmuller and colleagues found that in a review of 101 genome-wide linkage studies in 31 complex human diseases, over two-thirds of the studies showed no significance (Altmuller et al., 2001). Where genes have been discovered using linkage analysis they usually explain only a small fraction of the overall heritability of the disease. For example, variants known to affect the risk of inflammatory bowel disease (IBD) together explain a sibling risk of approximately two-fold, compared with a total excess risk of 30-fold (Daly et al., 2004), which indicates that there are additional genes in IBD to be discovered. The lack of success can be attributed to several reasons: (i) linkage is an indirect statistical test, relying on distortions of Mendelian inheritance ratios to infer the nearby location of a disease-causing mutation. When the genetic effect is very large (as in Mendelian disorders), this indirect signal is sufficient. For QTLs with only a small magnitude of effect, the power of linkage may be inadequate. (ii) Large sample sizes can compensate for low magnitude of effect, however assembling large families, each containing multiple affected individuals, is difficult particularly when the disease is rare, has an advanced age of onset and high mortality (as is the case for many neurodegenerative diseases). (iii) In addition, genetic factors such as pleiotropy (one gene which yields multiple phenotypes) and epistasis (the non-additive effect of multiple genes) provide extra complications.

1.5.2 Association studies

In their simplest form, association studies compare the frequency of alleles or genotypes of a particular variant between disease cases and controls. This may be done directly, where a candidate variant, often putatively functional, is tested for enrichment or depletion in disease cases compared to normal controls. For example, resequencing the entirety of a candidate gene in patients and controls, and testing variants such as single nucleotide polymorphisms (SNPs) in cases and controls is a direct test of association as every nucleotide is tested for a potential association with disease. Direct tests are hypothesis-based, with genes and candidate variants
selected for further study either by virtue their location in a region of linkage, or on the basis of other evidence that they might affect disease risk (Carlson et al., 2004a).

In contrast, indirect studies test markers (commonly SNPs) for disease association under the assumption that the marker itself is not causal but that it may be linked by allelic association or linkage disequilibrium (LD) to the true causal variant. Therefore, if a risk polymorphism exists it will either be genotyped directly or be in strong LD with one of the genotyped markers. The benefit of such indirect association studies, also called LD mapping, are that they do not require prior determination of which marker might be functionally important.

1.5.2.1 Linkage disequilibrium

Linkage disequilibrium (LD) has become an important consideration in most association studies for complex diseases. LD describes the non-random correlation between alleles at a pair of genetic markers (commonly SNPs). In contrast to linkage, which describes the association of loci on a chromosome with limited recombination between them, LD describes a situation in which some combinations of alleles occur more or less frequently in a population than would be expected if they were segregating randomly (i.e. in disequilibrium). Whereas linkage analysis relies on informative recombination events within a pedigree, LD utilises informative meioses of an extended pedigree of the complete human population. As a result, the interval shared through identity by decent within a population is much smaller than that shared by linkage in a pedigree, which can aid in the fine mapping of disease loci.

The measure \( D \) has been widely used to quantify LD. Consider two adjacent loci — A and B, with two alleles (A,a and B,b) at each locus — the observed frequency of the haplotype that consists of alleles A and B is represented by \( P_{AB} \). Assuming independent assortment of alleles at the two loci, the expected haplotype frequency is \( P_A \times P_B \) (where \( P_A \) is the frequency of allele A and \( P_B \) is the frequency of allele B). From Equation 1.1 it is clear that the primary determinant of LD is recombination which, in addition to recurrent mutation and evolution, serves to erode LD.

\[
\text{Equation 1.1} \quad D = P_{AB} - (P_A \times P_B)
\]

The utility of LD in association studies relies on the fact there is redundancy in the information gained from typing multiple SNPs in LD: for two SNPs in perfect LD, only one SNP would need to be typed to inform on the genotype of the other. This method of selecting a minimal set of markers to represent the underlying genetic variation is known as tagging. Additionally, SNPs may also be in LD with unknown SNPs that are not genotyped. If the unknown SNP is a causal variant for disease susceptibility, the known genotyped SNP could be used as a proxy to test for an indirect association with disease. The known SNP therefore is a tagging SNP (tSNP) for unknown genetic variation.
Equation 1.2 \[ D' = \frac{D}{D_{\text{max}}} \]

Equation 1.3 \[ r^2 = \frac{D'}{P_A P_B P_a P_b} \]

Although the LD measure $D$ provides a useful explanation as to the extent of LD between two loci, the measures $D'$ and $r^2$ are more commonly used in association studies (Devlin et al., 1995). $D'$ varies between 0 and 1, is calculated by normalising $D$ by its maximum possible value, $D_{\text{max}}$. $D' = 1$ only if two SNPs have not been separated by recombination. $r^2$ is more commonly used in the context of association studies and LD mapping and can be calculated by normalising $D'$ by the product of all four allele frequencies. $r^2$ is simply the squared correlation coefficient between the two loci. Perfect LD ($r^2=1$) indicates that markers have not been separated by recombination and have the same allele frequencies. An attractive property of $r^2$ is that it directly relates to sample size and power. For a study to detect an association between disease and a marker locus, the sample size must be increased by roughly $n/r^2$ (where $n$ is the number of samples in the study) to have the same power as actually genotyping the susceptibility locus itself (Kruglyak, 1999; Pritchard et al., 2001).

1.5.2.2 Linkage disequilibrium patterns in the human genome

The pattern of LD throughout the human genome has been extensively studied and has been of particular importance in association studies as it impacts on both the efficiency and power of these studies. There are several salient features of LD in the human genome. The first is that LD is inversely correlated with distance. This aspect of LD is intuitive as recombination rates are generally higher between markers at increasing distances and therefore, as LD is eroded by recombination, LD decay is seen over distance (Hartl et al., 2007). However, LD has been shown to be highly variable, extending between a few to several hundred kilobases (kb) (Dawson et al., 2002; Kruglyak, 1999).

Secondly, whereas LD was thought to decay uniformly over distance, several authors have provided evidence to that islands of high LD exist, separated by highly recombining regions (Daly et al., 2001; Rioux et al., 2001). These haplotype blocks of LD, punctuated by hotspots of recombination have been witnessed in multiple studies (Gabriel et al., 2002; International HapMap Consortium, 2003; Jeffreys et al., 2001) however more recent studies have suggested that LD at finer-resolution is more variable than previously thought (Ke et al., 2004). This atomistic picture of LD has been crucial in the design and efficiency of association studies (see below).

Thirdly, LD is heterogeneous amongst populations. Early studies by both Reich and colleagues and Gabriel and colleagues identified that population samples from Africa have markedly less LD than Europeans and Asians (Gabriel et al., 2002; Reich et al., 2001a). In contrast, younger or isolated populations have been shown to have the highest LD, suggesting that within different
populations LD is a reflection of that population’s history (i.e. demographic processes influence LD patterns) (Bonnen et al., 2002; Dunning et al., 2000; Kidd et al., 2000). This elevated LD can assist in the mapping of disease genes, as less haplotype diversity is seen and studies require fewer makers. In addition, the genetic architecture in population isolates can also be utilised as groups that have undergone a population bottleneck tend to have a distinct set of common disease alleles and tend to share more environmental features than outbred populations (Peltonen et al., 2000). Taken together, these studies have informed the design of association studies. By identification of the underlying LD structure across a candidate gene or even across the genome, a small number of SNPs could be used to define a large proportion of genetic diversity and be tested for association with disease.

1.5.2.3 Tagging SNPs

The efficacy of using a subset of SNPs to define the majority of variation was first demonstrated by Johnson and colleagues, who found that up to 6 SNPs could define the majority of haplotype diversity in a selection of gene (Johnson et al., 2001). In all, 122 SNPs, spanning 9 genes were reduced to just 34 haplotype tagging SNPs (htSNPs). More recently, tagging strategies moved towards maximising pairwise $r^2$ as this measure is more informative with respect to association studies (Carlson et al., 2004b). Further efficacy has been gained by the observation that tSNP haplotypes (Figure 1.4) in turn show redundancy (Goldstein et al., 2003a; Weale et al., 2003). In a study by Weale and colleagues, investigating the sodium channel gene SCN1A, implicated in epilepsy, this aggressive tagging technique provided a 110-fold saving in genotyping effort. In summary, based on the pattern of LD within a genomic region of interest, a minimal set of tSNPs can be chosen to tag known and unknown variation, providing association studies with efficiencies of both cost and effort.
Figure 1.4 Aggressive tagging of variation linked to six loci
Six SNP loci are shown with their corresponding alleles. A. Loci 1, 3 and 4 tag SNPs 2, 5 and 6 respectively with a minimum correlation coefficient $r > 0.85$. B. Redundancy of tSNP haplotypes allows the AG haplotype (red) at tSNPs 1 and 3 to predict the A allele at locus 6.

1.5.3 Association study design
The design of association studies and the technology available to conduct them has changed dramatically since the work presented in this thesis was undertaken. These changes and their effect on study design are examined in later chapters. However, there are some fundamental elements of association study design that are discussed below and in addition, advances in the resources available for future studies are discussed.

1.5.3.1 Study power and sample size
The power of an association study is the statistical probability that the study will detect a true association if one is present. There are several factors that affect study power including sample size, the prevalence and magnitude of effect of the risk factor, and the strength of linkage disequilibrium between the marker and causal variant. Sample size is probably the most common influential factor in association studies, especially in those of rare neurodegenerative diseases where achieving large cohorts would be difficult.
Figure 1.5 Association study sample size predictions

The sample size required for 80% power to detect an association of various genotype relative risks (GRR) with susceptibility alleles of increasing frequency. Dashed line using linkage methodology and solid line using association methodology. (From Risch, 2000).

The effect on samples size due to magnitude of effect (genotype relative risk; GRR) and allele frequency can be witnessed through modelling studies. As Figure 1.5 illustrates association studies can have greater power over linkage when GRR<2 and large sample sizes are required when GRR is small.

1.5.3.2 Sample selection

There are several design considerations related to sample choice and ascertainment that can lead to either spurious association or reduced study power. Population stratification is an important consideration for association studies as allele and haplotype frequencies can differ considerably between populations, resulting in both false positives and negatives associations. This bias can be a potential problem if, for example, disease prevalence varies between populations (i.e. between ethnic groups) as this may result in the over-representation of a subgroup in the case cohort and the more frequent polymorphism in that subgroup will tend to be more associated with disease despite not influencing it (Ardlie et al., 2002; Cardon et al., 2003). Stratification may be overcome by modifying the study design to include family based controls, as is seen in the transmission disequilibrium test (Spielman et al., 1993); typing multiple unlinked genomic control loci to detect stratification and correct for inflation in the test statistic (Devlin et al., 1999; Devlin et al., 2004); or by simply ensuring that cases and control cohorts are matched for ethnicity and environmental covariates (which for most studies can be easily achieved).

Power to detect association may be reduced by inadequate selection of case and control samples. The inadequate selection of case samples may be due to phenotypic heterogeneity or clinical misdiagnosis. The latter may be reduced by ensuring that all cases are diagnosed based on a set of universally adopted diagnostic criteria, such as the El Escorial criteria in ALS.
(Brooks, 1994), and the latter may be eliminated by considering intermediate phenotypes (Gottesman et al., 2003). Intermediate disease phenotypes are likely to have fewer extraneous environmental influences compared to those at the endpoint of disease. To overcome phenotypic heterogeneity they are required to be more predictive of disease or disease progression, than disease end-points, which are often used as clinical phenotype. For example, in age-related macular degeneration (AMD), a complex neurodegenerative disease affecting the retinal pigment epithelium, intermediate phenotypes such as photoreceptor segment turnover and macular pigment levels have been identified (Chamberlain et al., 2006).

![Threshold Value Diagram](image)

**Figure 1.6 Case-control selection using the threshold liability model**

Individuals with liability to disease greater than the threshold value will develop disease and be assigned to the case cohort. Control individuals are absent of disease but may either have an extremely low liability to disease (A) or be approaching the threshold value (B).

There are also several issues with the selection of control samples for case-control association studies of complex diseases. The first can be illustrated using the liability threshold model (Figure 1.6). Generally, the selection of a control cohort is conducted by selecting individuals absent for disease phenotype. However, in complex diseases, where disease may only manifest once the threshold of liability is reached, any individual under the threshold may be selected as a control. If a proportion of the control cohort is comprised of individuals just below the threshold level of liability, this may reduce the power of an association study to detect an association. By choosing controls from the left-hand tail of the distribution (i.e. hyper controls), who have a lower population risk of disease, this power to detect an association may be increased. In most studies where the defined phenotype is the presence/absence of disease or disease end-point, it will impossible to place the control cohort on this distribution, however quantitative intermediate phenotypes (Gottesman et al., 2003) or other indicators of disease risk that vary normally could be used.

1.5.3.3 Candidate gene and whole genome approaches

To date, the majority of case-control association studies for complex diseases have been undertaken on candidate genes. This has largely reflected the limited availability of polymorphism data and the high cost of genotyping for polymorphism ascertainment, making genome-wide methodologies in complex diseases unfeasible for most researchers. Candidate
genes have largely been ascertained in a number of ways: through known causation in Mendelian forms of disease, which are predicted to predispose to complex forms; through regions of linkage from sib-pair studies; through biological plausibility based on phenotypes; model organism mutants, such as mouse models of human disease, or the action of a gene in a pathway relevant to disease pathogenesis. Critically, these methods have relied on prior assumptions. In contrast, genome-wide studies can be undertaken with no prior assumptions as to the location or function of a pathogenic variant; However, these methods, which test up to 500,000 SNPs in one experiment and are therefore likely to detect false positives, also bear a heavy multiple testing penalty.

1.5.3.4 Allelic architecture of complex disease

For most complex diseases the underlying genetic variation remains unknown. The number of relevant disease variants that exist is unknown as are their frequencies in the population: genetic variation could be rare (alleles with frequencies <1% in the population), as is true for most single-gene disorders, it could be more common (frequencies >1% in the population). The frequency spectrum of the alleles for complex diseases is important to consider, because the allele frequencies of variants that predispose to disease and the strength of their phenotypic effects directly relate to the statistical power of genetic association studies, and therefore their likelihood of success.

These underlying allele frequencies are largely unknown because the causal variants remain largely unknown, however theoretical and empirical studies suggest that for complex diseases, some of the causal genetic variants may be common (Chakravarti, 1999; Lander et al., 1994; Reich et al., 2001b). The common disease/common variant (CDCV) hypothesis proposes, that common diseases are a result of common variants (Reich et al., 2001b). Under this model, disease susceptibility is suggested to result from the joint action of several common variants, and unrelated affected individuals share a significant proportion of disease alleles. There are several arguments to support this theory: (i) common diseases are generally not as evolutionarily disadvantageous as single-gene disorders, which often cause early death or markedly decreased reproductive capability; (ii) variants that cause single-gene disorders are highly penetrant, whereas multiple variants are required to cause common diseases and thus, the impact of selective pressure is diluted for the variants for complex traits; (iii) the majority of monogenic diseases are rare, whereas most polygenic diseases are common- population genetic arguments predict that for common diseases, some of the causal genetic variation should have a high frequency in the population, due to the demographic history of the human population (Reich et al., 2001b); and (iv) empirical evidence has suggested that common variants do contribute to the risk of common diseases (Lohmueller et al., 2003).
However, evidence from monogenic disorders suggests that the responsible variants are generally rare, so could the variants that cause common disease not also be predominantly rare? The alternative to CDCV is the disease heterogeneity hypothesis (or multiple rare-variant hypothesis), in which disease susceptibility is due to distinct genetic variants in different individuals and disease-susceptibility alleles have low population frequencies <1% (Smith et al., 2002). This allelic heterogeneity is a potential problem for association studies, unlike the CDCV hypothesis which supports the mapping of loci through association (Pritchard et al., 2002), and may be a more pervasive phenomenon in common disease (Pritchard, 2001). So, what are the prospects for neurodegenerative diseases? Successes in mapping the susceptibility APOE ε4 locus associated with increased risk to Alzheimer disease (Corder et al., 1993) has been possible due to high allele frequencies, ranging from 5%-41% in various populations (Fullerton et al., 2000) and large effect size. However, identification of risk factors with smaller effect sizes at lower frequencies will require much larger samples and possibly more sensitive analytic techniques.

As discussed later, allelic heterogeneity and the CDCV hypothesis may be a surmountable problems for complex loci mapping and even some successes have been seen. For example, the gene NOD2 has been successfully associated with Crohn’s disease despite the presence of moderate allelic heterogeneity (Hugot et al., 2001) as has CAPN10, a susceptibility locus for type 2 diabetes (Horikawa et al., 2000). In addition populations with a greater degree of common variation such as isolated populations can be used to identify rare variants in outbred populations. For example, the Finnish population descends from two waves of immigration, 200 years ago from the southern Indo-Europeans and 4000 years ago from eastern Uralic speakers. The population expanded from ~50,000 in the 12th century to the present day ~5,000,000. Relative homogeneity of culture and lifestyle has also been noted. Approximately 30 recessive diseases are enriched in Finland, whilst rare diseases found elsewhere in the world are absent. Both allelic and non-allelic heterogeneity are reduced which serve to increase association study power. Thus isolated populations share some of the advantages of inbred mouse lines and provide greater power in association studies to detect rare alleles (Lee et al., 2001b).

1.5.3.5 Evolving resources for association studies in complex diseases

The association study data presented in this thesis serve as a reminder of the infancy of our knowledge of human genetic variation and ability to map complex traits. In the years since this work began, several aspects of study design, theory and the tools available to conduct these studies have matured. These advances are discussed in detail later and contrasted with our understanding at the outset of this work. Possibly the most significant difference however has been the availability of genomic variation and LD data, and advances in SNP genotyping technology.
An important consideration in association study design has been the construction of LD maps. Due to costs and feasibility issues, LD maps have historically been determined for candidate genes, as SNPs have had to be located, genotyped and LD assessed. From the year 2000, the volume of freely available SNP information in the database dbSNP has increased (Figure 1.7). dbSNP database now contains over 9.7 million SNPs (as of September 2006) obviating much of the need to ascertain SNPs in a candidate region.

![Figure 1.7 SNP submissions to dbSNP from over 5 years from 2000](image)

**Figure 1.7 SNP submissions to dbSNP from over 5 years from 2000**

Graph shows cumulative number of non-redundant SNPs, each mapping to a single location in the genome (solid line), the number of SNPs validated by genotyping (dotted line) and double-hit status (dashed line). Years are divided into quarters (Q1–Q4). (From The International HapMap Consortium, 2005)

Part of the exponential increase in SNP submissions since 2002 has been due to the launch of The International Haplotype Map project (HapMap) (International HapMap Consortium, 2003). This international collaboration undertook an extensive SNP discovery effort in 3 populations of African, Asian, and European ancestry to construct a dense 5kb SNP map and ascertain the underlying pattern of LD and therefore haplotype structure. This project has contributed vastly to our understanding of human genetic variation (The International HapMap Consortium, 2005) and has provided a resource for researchers choosing to undertake association studies. The availability of HapMap data (www.hapmap.org/) has obviated the need for investigators to have to ascertain SNPs and construct LD maps, saving considerable time and money. In addition, the advent of new genotyping technologies over the past seven years has seen genotyping costs reduced from $0.50 per genotype to $0.001 per genotype, which has had an effect on the number of SNPs and the number of samples that can be screened – ultimately increasing study power.

### 1.6 Evolutionary analyses

The search for new methods to identify genes involved in complex diseases has led several authors to propose the use of natural selection to identify disease susceptibility loci (Jorde *et al.*, 2001; Nielsen, 2001). The rationale behind this approach is two fold: (i) The previous
identification of genes implicated in both Mendelian and complex diseases and other complex phenotypes, show patterns of genetic variation characteristic of selection (referred to from here as signatures of selection), see Table 1.2 – over 90 different loci have been proposed as possible targets for selection to date (Sabeti et al., 2006). (ii) Genetic variants that may have been under evolutionary selection are by definition functional and thus maybe more likely to contribute to disease susceptibility today.

<table>
<thead>
<tr>
<th>Disease or phenotype</th>
<th>Gene</th>
<th>Mode of selection</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malaria resistance</td>
<td>G6PD</td>
<td>Positive</td>
<td>(Sabeti et al., 2002)</td>
</tr>
<tr>
<td></td>
<td>TNFSF</td>
<td>Positive</td>
<td>(Rockman et al., 2004)</td>
</tr>
<tr>
<td></td>
<td>FOXP2</td>
<td>Positive</td>
<td>(Enard et al., 2002; Zhang et al., 2002)</td>
</tr>
<tr>
<td>Heart disease risk</td>
<td>MMP3</td>
<td>Positive</td>
<td>(Tishkoff et al., 2007)</td>
</tr>
<tr>
<td></td>
<td>SCA2</td>
<td>Positive</td>
<td>(Yu et al., 2005)</td>
</tr>
<tr>
<td>Lactase deficiency</td>
<td>LCT</td>
<td>Positive &amp; balancing</td>
<td>(Wood et al., 2005)</td>
</tr>
<tr>
<td>Autosomal dominant</td>
<td>SCA2</td>
<td>Positive</td>
<td>(Bamshad et al., 2002; Wooding et al., 2005)</td>
</tr>
<tr>
<td>spinocerebellata ataxia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Putative viral resistance</td>
<td>CCR5</td>
<td>Positive &amp; balancing</td>
<td>(Meyer et al., 2006)</td>
</tr>
<tr>
<td>Sickle cell/malaria resistance</td>
<td>HBB</td>
<td>Positive &amp; balancing</td>
<td>(Mead et al., 2003)</td>
</tr>
<tr>
<td>Immune recognition</td>
<td>HLA</td>
<td>Balancing</td>
<td>(Verrelli et al., 2004)</td>
</tr>
<tr>
<td>Kuru</td>
<td>PRNP</td>
<td>Balancing</td>
<td>(Verrelli et al., 2004)</td>
</tr>
<tr>
<td>Trichromatic colour</td>
<td>OPN1LW</td>
<td>Purifying</td>
<td>(Verrelli et al., 2004)</td>
</tr>
</tbody>
</table>

Table 1.2 Reports of selection in the human lineage

Ten examples of genes with selection in the human lineage since the divergence of humans and chimpanzees. References given represent recent studies only and are not exhaustive. (Modified from Sabeti et al., 2006).

1.6.1 Genetic variation is shaped by several forces

The genetic variation of human populations today, which may lead to disease susceptibility, comprises a composite of ancestral influences including mutation, natural selection and population history, also known as demography (i.e. migration, expansion, isolation etc).

1.6.2 Measuring genetic variation and disease mapping by natural selection

Our increased understanding of human genetic variation, largely due to the sequencing of the human genome and the completion of the HapMap project sequencing has augmented our ability to detect neutral variation and deviations from neutrality. The neutral theory of molecular evolution (Kimura, 1968), is perhaps the best place to begin this discussion as it is has been integral to recent studies of natural selection. The neutral theory posits that the majority of polymorphisms that arise through random mutation have no appreciable effect on fitness (i.e. are neutral) and that these polymorphisms vary randomly in frequency over time (genetic drift) and are eliminated or fixed in populations as a consequence of the stochastic effects genetic drift. Specifically, the neutral theory can be used to make explicit and quantitative predictions about the amount, structure, and patterns of sequence variation expected under neutrality, and serves as a null hypothesis by which to evaluate the evidence for or against selection in empirical data.
The pattern of neutral variation in the human genome can be summarised using the site frequency spectrum which represents the distribution of derived* allele frequencies in a population (see Figure 1.8C). Data from the Encyclopaedia Of DNA Elements (ENCODE)* regions of the HapMap project have illustrated the exponential decay in SNPs of increasing frequency, which illustrates that the human genome is composed of a majority of young alleles. The importance of this spectrum to population genetics is that several tests of selection are based on deviations from this expected allele frequency distribution including Tajima’s D, Fu and Li’s D and Fu and Li’s F. In addition, there are several selection statistics that are independent of the site frequency spectrum but which rely instead on comparisons of polymorphisms within and/or between species such as Macdonald-Kreitman, Hudson-Kreitman-Aguadé (HKA) test and dN/dS tests. A comprehensive description of these tests and their relative merits is beyond the scope of this thesis but several excellent reviews can be found (Bamshad et al., 2003; Biswas et al., 2006; Kreitman, 2000; Sabeti et al., 2006).

1.6.2.1 Natural selection

Natural selection occurs when a new mutation results in differential reproductive success (fitness). In humans, it has been estimated that ~4 new amino-acid-altering mutations arise per diploid genome per generation (Eyre-Walker et al., 1999); these mutations can be broadly categorized as positive (i.e. enhances fitness), negative (i.e. reduces fitness) and balancing (i.e. enhances fitness only in heterozygous state). Accordingly, each form of selection can characteristically increase, decrease or have an intermediate effect on the mutant allele frequency, under positive, negative and balancing pressure respectively, and produce a characteristic signature of selection in the surrounding neutral variation.

Negative selection is the most pervasive form of selection as most mutations that cause protein coding changes are deleterious. These mutations are selected against and are likely to be lost. As this type of selection conserves amino acid sequence it is often called purifying selection. In the presence of strong purifying selection, nucleotide diversity at tightly linked sites is also reduced (Charlesworth et al., 1993), although this background selection is also dependent on local mutation and recombination rates.

Positive selection occurs when a new mutation enhances the fitness and is likely to become fixed within a population. The time taken for the new mutation to fixate is dependent on the mode of action of the new mutation: a dominant allele fixates faster than a recessive one. In addition, the type and the magnitude of the selective pressure influence time to fixation. If a

* New mutations create a polymorphic site – therefore derived alleles are younger than ancestral alleles

† ENCODE represents ten 500kb regions that have been fully resequenced in 48 unrelated samples to identify all variation sampled within the region
mutation increases in frequency, tightly linked neutral variation can be dragged along as well. This genetic hitchhiking can eliminate variation not linked to the advantageous mutation, resulting in a selective sweep.

Balancing selection occurs when the heterozygous phenotype has a greater fitness than either homozygote. The effect of balancing selection is to maintain both alleles at equilibrium frequency. There are few examples of balancing selection in human populations. Probably the most well known is sickle cell anaemia, in which the sickle cell mutation in the \textit{HBB} gene offers resistance to malaria in the heterozygous state but causes severe and fatal anaemia in the homozygous state. The \textit{G6PD} gene, which can cause haemolytic disease, has also been shown to have undergone balancing selection possibly as it is associated with protection against malaria in the heterozygous state. In addition, the Major Histocompatibility Complex (MHC) on chromosome 6 has been shown to have been under balancing selection which has been proposed to help maintain a wider repertoire for non-self-peptide recognition in an immune response. Most recently, the prion gene (\textit{PRNP}), known to be associated with risk of human prion disease, has been shown to have undergone balancing selection worldwide. All three forms of selection have distinct effects on genetic variation surrounding a mutation.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1_8.png}
\caption{The effects of selection on the distribution of genetic variation}
\textbf{(a)} The genealogies of three genes with 20 polymorphic sites that are typical of loci under positive selection (top), balancing selection (middle) and no selection (bottom) are shown. Each circle represents a mutation, and the colour shows the final frequency of each mutation. 
\textbf{(b)} Each haplotype contains mutations that have accumulated on each lineage in the gene genealogy, assuming no recombination. \textbf{(c)} The site frequency spectrum of each gene. Positive selection (top) can result in a lower level of sequence diversity (\(\pi\)), an excess of low-frequency variants (red) and, consequently, a negative value of Tajima’s \(D\). Balancing selection (middle) can result in a higher level of sequence diversity (\(\pi\)), an excess of intermediate-frequency variants (purple) and, consequently, a positive value of Tajima’s \(D\). The diversity estimate and site frequency spectrum of a neutral locus (bottom) can be used for comparison. (From Bamshad et al., 2003)}
\end{figure}
1.6.2.2 Testing for signatures of selection

Sabeti and colleagues have recently published a proposal of five signatures of selection, into which all specific tests can be categorised (Sabeti et al., 2006), as summarised below:

(i) High proportion of function-altering mutations: genetic variants that alter protein function are usually deleterious and are thus less likely to become common or reach fixation (i.e. 100% frequency) than are mutations that have no functional effect on the protein (i.e. synonymous change). Positive selection over a prolonged period, however, can increase the fixation rate of beneficial function-altering mutations and such changes can be measured by comparison of DNA sequence between species. The increase can be detected by comparing the rate of non-synonymous changes with the rate of synonymous or other presumed neutral changes, by comparison with the rate in other lineages, or by comparison with intraspecies diversity. Statistical tests commonly used to detect this signature include the $K_a/K_s$, $d_N/d_S$ and MKA tests.

(ii) Reduction in genetic diversity: As an allele increases in population frequency, genetic hitchhiking alters the typical pattern of genetic variation in the region – leading to a selective sweep. In a complete selective sweep, the selected allele rises to fixation, bringing with it closely linked variants; eliminating diversity in the immediate vicinity and decreasing it in a larger region. Although new mutations eventually restore diversity, these appear slowly (because mutation is rare) and are initially at low frequency. Positive selection thus creates a signature consisting of a region of low overall diversity, with an excess of rare alleles. The opposite is true (but has been observed less frequently) for balancing selection, such as that of the prion gene $PRNP$ – an excess of high frequency is seen. Statistical tests commonly used to detect this signal include Tajima’s $D$, the HKA test, and Fu and Li’s $D^*$.  

(iii) High-frequency derived alleles: Derived alleles arise by new mutation and typically have lower allele frequencies than ancestral alleles. In a selective sweep, however, derived alleles linked to the beneficial allele can hitchhike to high frequency. As a result of an incomplete sweep or recombination, many of these derived alleles will not reach complete fixation, thus positive selection creates a signature of a region containing many high-frequency derived alleles. For example, the 10-kb region around the Duffy red cell antigen (FY) has an excess of high-frequency derived alleles in Africans and is therefore, thought to be the result of selection for malaria resistance (Escalante et al., 2005; Hamblin et al., 2002). A commonly used statistical test to detect this signal is Fay and Wu’s $H$.

(iv) Differences between populations: When geographically separate populations are subject to distinct environmental or cultural pressures, positive selection may change the frequency of an allele in one population but not in another. Large differences between populations in the selected allele frequency or in surrounding variation can therefore signal a locus that has
undergone positive selection. For example, the region surrounding the lactase \((LCT)\) locus demonstrates large population differentiation between Europeans and non-Europeans, reflecting strong selection for the lactase persistence allele in Europeans (Bersaglieri et al., 2004). A commonly used statistic for population differentiation includes \(F_{ST}\).

(v) **Long haplotypes:** Under positive selection, a selected allele may rise in prevalence rapidly enough that recombination cannot substantially erode LD with hitchhiking alleles on the ancestral chromosome. In these circumstances alleles have both high frequency (typical of old alleles) and long-range LD with other alleles (typical of young alleles). Thus, selective sweeps can produce a distinctive signature that would not be expected under neutral genetic drift. Long-range LD manifests as a long haplotype that has not been broken down by recombination. For example, the lactase persistence allele at the \(LCT\) locus lies on a haplotype that is common (~77%) in Europeans but that extends undisrupted for more than 1 megabase (Mb) (Bersaglieri et al., 2004), much farther than is typical for an allele of that frequency. This signature can be detected with the long-range haplotype (LRH) test.

### 1.6.2.3 Confounding factors – demography and ascertainment

Robust inferences of natural selection from DNA sequence data are difficult because of the confounding effects of population demographic history, such as population bottlenecks, expansions and subdivisions, and ascertainment bias within the study design. Ascertainment bias results from the fact that low frequency SNPs can go undetected if the genotyping sample size is too small. As the probability that a SNP is identified in a limited sample is a function of the allele frequency, rare SNPs are more likely to go undiscovered compared with common SNPs. As a consequence the site frequency spectrum obtained will be different from that obtained under complete sampling (e.g. by resequencing the entire study sample). As a result, statistical attributes that rely on the site frequency spectrum - including Tajima's \(D\), \(F_{ST}\), and LD - will be affected.

Confounding due to population demography can mimic signatures of selection. For example, both positive selection and increases in population size lead to an excess of low-frequency alleles in a population relative to that expected under a neutral model. In addition population subdivision can lead to spurious departures from Hardy-Weinberg equilibrium (discussed below). Whilst population demography affects genetic variation across the entire genome, selection is much more likely to act in a defined region only. Therefore, a true signature of selection can be distinguished from population demography by examining unlinked loci or by determining an empirical genome-wide distribution for test statistics used. As genome-wide genotyping has become more feasible, a greater number of authors have attempted to use
genome-wide scans for tests of selection (Akey et al., 2002; Carlson et al., 2005; Kelley et al., 2006; Nielsen et al., 2005).

1.6.3 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium (HWE; otherwise known as Hardy-Weinberg law or principle) describes a mathematical model which has been routinely applied to population genetics and more recently, as a tool to identify disease loci. Independently discovered in 1908 by Wilhelm Weinberg, a German physician, and Godfrey Harold Hardy, a British mathematician, the Hardy-Weinberg principle states the frequencies of alleles in a population will remain the same regardless of the starting frequencies and that equilibrium genotypic frequencies will be established after one generation of random mating (i.e. genotype frequencies will be constant from generation to the next). HWE is found when the measured genotype frequencies match those predicted by allele frequencies using

\[
p^2 + 2pq + q^2 = 1
\]

(genotype frequency AA = \(p^2\), Aa = 2pq, aa = \(q^2\))

The principle is based on assumptions of random mating, large population size, equal sex distribution and the absence of migration, mutation and selection. Expected and observed genotype frequencies can be tested for significance using a \(\chi^2\) test and departures from HWE can imply that one of the assumptions listed above is incorrect. Departures from HWE have therefore been used as a quality control measure for large-scale genotyping and as a method for localising disease genes and loci under selection (Feder et al., 1996; Lee, 2003; Nielsen et al., 1998).

1.7 Amyotrophic lateral sclerosis

1.7.1 Clinical and pathological features of ALS

Amyotrophic lateral sclerosis (ALS), first described by the French physician and neurologist Jean-Martin Charcot in 1869, is characterised by the progressive degeneration of motor neurons from the motor cortex and corticospinal tract. Disease onset usually occurs focally in either limb, presenting as muscle weakness, or with bulbar/corticobulbar ALS, generally presenting as affected speech or swallowing. As the disease advances, symptoms of progressive weakness, atrophy, paralysis, spasticity and emotional lability are seen. Death usually occurs 2 to 5 years after disease onset and is commonly due to respiratory failure as a consequence of denervation of the respiratory muscles and diaphragm.

Pathologically, ALS is characterised by the loss of upper motor neurons in the motor cortex, degeneration of the corticospinal tract and loss of lower motor neurons in the brain stem and spinal cord. The selective degeneration of motor neurons only, generally spares the patient's sensory and cognitive functions. Motor neurons involved in ocular, sphincter and urethral
function are spared and these signs are often used to exclude other diagnoses. Histopathology of surviving lower motor neurons shows ubiquinated inclusions and marked axonal swelling.

In the absence of a diagnostic test, the clinical diagnosis of ALS is primarily one of exclusion. The early symptoms of ALS are often indistinguishable from similar disorders that affect motor neurons, and these are often confused for ALS. In the United Kingdom (UK), these disorders are termed motor neuron diseases (MNDs) – a blanket term referring to a whole spectrum of diseases which differentially affect the upper and lower motor neurons or both. In the United States, the term ALS is commonly used interchangeably with MND, where it is also known as Lou Gehrig’s disease, after the baseball player.

ALS is currently classified based on a set of diagnostic criteria set out by the World Federation of Neurology; the El Escorial Criteria (Brooks, 1994), and its recent revision, Airlie House Criteria (Miller et al., 1999). Clinical diagnosis of ALS using these criteria, is based on the presence of upper and lower motor neuron impairment, the detection of symptom progression over a limited period of time, and the exclusion of other conditions that may mimic ALS.

1.7.2 Epidemiology

ALS is currently the third most common neurodegenerative disease after Alzheimer’s and Parkinson’s disease and is largely a disease of midlife (although rare juvenile onset forms are the exception) with an average age of onset between the 5th and 6th decade of life. A slight sex bias exists with males more likely to be affected than females in a ration of 1.3:1–1.6:1 (Nelson, 1995). Estimates of the lifetime risk currently vary between 1 per 800-2000 (Cleveland et al., 2001; Shaw, 2005) and the annual incidence of ALS is ~1-2 per 100,000 which increases with age. Disease prevalence is ~2-4 per 100,000, which is kept low due to the rapid disease course.

ALS incidence is uniform throughout the world with the exception of several high incidence foci, such as those in the Western Pacific (the Kii peninsular of Japan and Guam in the Marianas), where a variant of ALS seen with Parkinsonism and dementia has been observed at an increased frequency (Arnold et al., 1953; Kimura, 1961). Although the vast majority of ALS cases are sporadic (known as sporadic ALS; SALS) with no apparent familial recurrence of the disease, ~10% of all ALS cases and ALS- related syndromes are inherited (known as familial ALS; FALS).

1.7.3 The Mendelian genetic basis of ALS

As with many other diseases, the greatest advance in our understanding of the genetics of ALS has come from familial forms of the disease, which show clear Mendelian inheritance. Given

---

* Lou "The Iron Horse" Gehrig is still regarded today as one of the greatest baseball players of all time. Of his many records, Lou Gehrig played in the most number of consecutive games over 15 years from 1925 to 1939 – a streak which ended when he developed ALS and subsequently died two years later.
that until recently, gene identification strategies have largely relied on observing Mendelian inheritance, within large families with highly penetrant mutations, it is not surprising that these genetic loci are the most robustly replicated genetic causes of ALS to date. However, FALS accounts for only a small proportion of all ALS cases and to date at least 12 genetic loci have been identified (Table 1.3).

The first ALS-associated gene to be identified was copper-zinc superoxide dismutase 1 (SOD1), on chromosome 21q22.1 (Rosen et al., 1993) which accounts for the greatest number (~20%) of familial ALS cases. To date, over 100 disease causing point mutations in SOD1, spanning all five exons, have been identified (Andersen et al., 2003) (see www.alsod.iop.kcl.ac.uk/als), the vast majority of which are heterozygous missense mutations (insertions and deletions are rarely seen). Almost all mutations are dominantly inherited with the exception of recessive inheritance such as the D90A mutation (Andersen et al., 1995).
<table>
<thead>
<tr>
<th>ALS disease type</th>
<th>Onset</th>
<th>Inheritance</th>
<th>Gene</th>
<th>Locus</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mendelian genes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALS1</td>
<td>Adult</td>
<td>AD, AR (D90A)</td>
<td>SOD1</td>
<td>21q22.1</td>
<td>(Al Chalabi et al., 1998; Rosen et al., 1993)</td>
</tr>
<tr>
<td>ALS2</td>
<td>Juvenile</td>
<td>AR</td>
<td>Alsin</td>
<td>2q33</td>
<td>(Hadano et al., 2001; Yang et al., 2001)</td>
</tr>
<tr>
<td>ALS4</td>
<td>Juvenile</td>
<td>AD</td>
<td>SETX</td>
<td>9q34</td>
<td>(Blair et al., 2000; Chen et al., 2004)</td>
</tr>
<tr>
<td>ALS8</td>
<td>Adult</td>
<td>AD</td>
<td>VAPB</td>
<td>20q13.33</td>
<td>(Nishimura et al., 2004b; Nishimura et al., 2004a)</td>
</tr>
<tr>
<td>Progressive LMN disease</td>
<td>Adult</td>
<td>AD</td>
<td>DCTN1*</td>
<td>2p13</td>
<td>(Munch et al., 2004; Munch et al., 2005; Puls et al., 2003)</td>
</tr>
<tr>
<td>ALS with dementia, Parkinsonism</td>
<td>Adult</td>
<td>AD</td>
<td>MAPT</td>
<td>17q21</td>
<td>(Clark et al., 1998; Hutton et al., 1998; Siddique et al., 1995)</td>
</tr>
<tr>
<td>Mendelian loci</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALS3</td>
<td>Adult</td>
<td>AD</td>
<td>?</td>
<td>18q21</td>
<td>(Hand et al., 2002)</td>
</tr>
<tr>
<td>ALS5</td>
<td>Juvenile</td>
<td>AR</td>
<td>?</td>
<td>15q15.1-q21.1</td>
<td>(Hentati et al., 1998)</td>
</tr>
<tr>
<td>ALS6</td>
<td>Adult</td>
<td>AD</td>
<td>?</td>
<td>16q12</td>
<td>(Abalkhail et al., 2003; Ruddy et al., 2003; Sapp et al., 2003)</td>
</tr>
<tr>
<td>ALS7</td>
<td>Adult</td>
<td>AD</td>
<td>?</td>
<td>20p1-p13</td>
<td>(Sapp et al., 2003)</td>
</tr>
<tr>
<td>ALS-FTD</td>
<td>Adult</td>
<td>AD</td>
<td>?</td>
<td>9q21-22</td>
<td>(Hosler et al., 2000)</td>
</tr>
<tr>
<td>ALS X</td>
<td>Adult</td>
<td>AD</td>
<td>?</td>
<td>9p13.2-21.3</td>
<td>(Morita et al., 2006; Vance et al., 2006)</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>AD</td>
<td>?</td>
<td>17q</td>
<td>(Wilhelmsen et al., 2004)</td>
</tr>
<tr>
<td></td>
<td>Adult</td>
<td>AD</td>
<td>?</td>
<td>Xp11-q12</td>
<td>(Siddique et al., 1998a)</td>
</tr>
<tr>
<td>Mitochondrial genes and misc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALS-M</td>
<td>Maternal;</td>
<td>Cox1</td>
<td>mtDNA</td>
<td>Single case (Siddique et al., 1998a)</td>
<td></td>
</tr>
<tr>
<td>ALS-M</td>
<td>Maternal</td>
<td>IARS2</td>
<td>mtDNA</td>
<td>Single case (Borthwick et al., 2006)</td>
<td></td>
</tr>
<tr>
<td>ALS-M</td>
<td>Maternal</td>
<td>CCO1</td>
<td>mtDNA</td>
<td>Single case (Comi et al., 1998)</td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td>NFH</td>
<td>Single case (Al Chalabi et al., 1999)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.3 Familial ALS genes and loci

Autosomal dominant (AD); amyotrophic lateral sclerosis (ALS); autosomal recessive (AR); alsin (ALS2); cyclooxygenase 1 (COX1); frontotemporal dementia (FTD); mitochondrial isoleucine tRNA synthetase (IARS2); mitochondrial DNA (mtDNA); senataxin (SETX); superoxide dismutase 1 (SOD1); vesicle-associated membrane protein-associated protein B (VAPB); and unknown gene (?)
The pathogenesis of SOD1 mutations is unclear. The normal 153-amino acid SOD1 protein functions as a copper and zinc containing homodimer, which detoxifies superoxide (produced during oxidative phosphorylation by mitochondria) to oxygen and hydrogen peroxide, thus preventing oxidative damage. SOD1 is ubiquitously expressed and predominantly located in the cell cytoplasm. Most SOD1 mutations affect correct subunit folding and dimerisation, although the resulting effect on SOD1 activity varies (e.g. G93A mutations have almost no effect on activity whilst H46R mutations inactivate SOD1). No clear correlation has yet been identified between mutant SOD1 activity, including copper affinity, protein half-life and ability to scavenge superoxide, and disease onset or duration. Patients with SOD1 mutations have normal protein levels suggesting that pathogenesis is not due to haploinsufficiency, however, SOD1 may exert a dominant negative effect or toxic gain of function.

Clinically, SOD1 FALS is indistinguishable from SALS although onset is ~10 years earlier and there is 90% penetrance by age 70. The mean age of onset is consistently 46 to 47 years regardless of the underlying SOD1 mutation. Mean duration is 3-6 years, however, exceptions such as the aggressive A4V mutation which has a short disease duration of 1.4 years (Cudkowicz et al., 1997) and G37R or G41D which have longer survival prognoses (Orrell et al., 1997b). Initially, SOD1 FALS cases present with predominantly lower motor neuron signs. Bulbar onset is unusual and is usually associated with a later-onset cases.

There are few common pathological features of SOD1-mediated ALS such as corticospinal and anterior horn cell loss but generally pathology varies with different mutations. SOD1-mediated FALS clinical phenotype including age of onset, severity and survival all vary and the wide range of mutations in SOD1 cannot account for this variability. In addition clinical variation can be observed in members of the same family (Andersen et al., 1995), suggesting that additional environmental or genetic factors may modify phenotype. That some SOD1 FALS mutations may appear more complex than previously thought is illustrated by the incomplete penetrance of the D90A mutation (discussed below) which is recessive on some genetic backgrounds and dominant on others (Al Chalabi et al., 1998).

1.7.4 The complex genetic basis of ALS

The biggest challenge for ALS research is currently dissecting the aetiology of sporadic ALS which accounts for ~90% of all ALS cases. As Simpson and Al-Chalabi have noted, the assumption that sporadic disease implies a disease with no genetic cause is no longer valid.
(Simpson et al., 2006). Indeed, clinical, epidemiological and genetic studies of SALS cases have all revealed a genetic component.

1.7.4.1 A genetic component to sporadic ALS

Perhaps the simplest example of a genetic component to SALS is the discovery of SOD1 mutations in 2 to 7% of SALS cases (Andersen et al., 1995; Jackson et al., 1997; Jones et al., 1994a; Jones et al., 1994b), indicating that at least some sporadic cases of ALS are genetic, although it is unclear how many are de novo mutations. Several lines of evidence implicate a genetic component to SALS:

Heritability A UK twin study, investigating concordance between mono- and dizygotic twins, has provided additional support for a genetic component to SALS. The authors demonstrated heritability of ALS risk to be between 38 and 85% - essentially that between 38% and 85% of variation in ALS is attributable to inherited (i.e. genetic) factors (Graham et al., 1997). In addition estimated sibling relative risk (As), under a set of assumptions regarding family history and inheritance patterns, is between 20 and 50 (i.e. the sibling of an affected individual is 20 to 50 times more likely to develop disease) (Simpson et al., 2006).

Ethnic differences It is widely recognized that there is variation within and between human populations with respect to their susceptibility to many diseases (Burchard et al., 2003). Investigation of variation in ALS incidence by ethnicity has been a source of much epidemiological research and controversy. In these studies, ethnicity can be regarded as a proxy to describe population genetic variation or variation in genetic background. Epidemiological data from worldwide populations analysing rates of ALS incidence and prevalence, migration and mortality have shown that the rates are consistently lower in African, Asian and Hispanic ethnicities that in Caucasians see for example (Dean et al., 1993; Elian et al., 1993; Noonan et al., 2005). However these studies often suffer from ascertainment and reporting bias. Cronin and colleagues have recently reviewed 61 published articles examining ethnic variation in the incidence of ALS and based on standardised measures of incidence, they concluded that African, Asian and Hispanic ethnicities did show a lower overall incidence of ALS compared to Caucasians (Cronin et al., 2007). It is entirely plausible that these differences may be due to differential environmental exposures too.
Genetic background has been postulated to influence several familial forms of ALS. For example, the D90A mutation of SOD1 is responsible for a recessively inherited form of ALS in families from the Torne Valley, in Sweden and Finland (Andersen et al., 1995; Andersen et al., 1996; Andersen et al., 1997). In addition, families from southern Sweden, UK, France and Belgium, heterozygous for the D90A mutation have also been reported with a more aggressive and variable ALS phenotype, and a few individuals with no family history of ALS have also been described (Andersen et al., 1995; Robberecht et al., 1996, Jackson et al., 1997; Khoris et al., 1997). Remarkably, all families carrying the D90A mutation have been shown to have descended from a single ancient founder ~895 generations ago, with the recessive allele becoming established in Torne Valley during the settlement and isolation of the area ~63 generations ago (Al Chalabi et al., 1998; Parton et al., 2002). Importantly, a region of up to 265kb around SOD1 is shared by the recessive Torne Valley kindreds which is postulated to contain a tightly linked cis-acting protective factor, such that two copies of D90A are required for ALS to develop. No such linked protective factor has yet been identified.

In addition the failure of replication of ALS association studies conducted in different worldwide populations may be due to differences in genetic background, they may also be due to different environmental factors, e.g. VEGF (Van Vught et al., 2005).

1.7.5 Susceptibility genes in sporadic ALS

Based on the liability threshold model, susceptibility to ALS can be considered to be a spectrum of genetic and environmental risk: (i) at one end of the spectrum are the well studied single genes, with large effect size, that cause fully penetrant, autosomal dominant FALS and (ii) at the opposite end are the multiple genes of small effect that may interact with the environment to cause apparently SALS. Until recently, linkage studies have only been able to illuminate the genes of large effect responsible for one side of the liability spectrum but have lacked power in elucidating those genes of small effect. Genes of small effect have, until recently, been investigated in ALS using candidate gene association studies, often informed by those genes identified from linkage studies. However, the availability of cheaper genotyping and denser marker sets is changing how these susceptibility genes will be discovered. Table 1.2 summarises the candidate genes which have been investigated in ALS as modifiers of risk or phenotype. Rather than list every single one of these in the text, I will discuss some of the most studied.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Reason for Investigation</th>
<th>Significance</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALAD</td>
<td>d-Aminolevulinic Acid Dehydratase</td>
<td>Lead exposure associated with ALS ALAD is involved in haem synthesis in erythrocytes</td>
<td>No association</td>
<td>(Kamel et al., 2002; Kamel et al., 2003)</td>
</tr>
<tr>
<td>ALS2</td>
<td>Alsin</td>
<td>Causes autosomal recessive juvenile ALS (ALS2).</td>
<td>No association</td>
<td>(Hadano et al., 2001; Hand et al., 2003; Hentati et al., 1994; Hosler et al., 1990; Yang et al., 2001)</td>
</tr>
<tr>
<td>ANG</td>
<td>Angiogenin</td>
<td>ANG is functionally similar to VEGF</td>
<td>Association found in Scottish and Irish populations</td>
<td>(Greenway et al., 2004; Hayward et al., 1999)</td>
</tr>
<tr>
<td>APEX</td>
<td>Apurinic endonuclease</td>
<td>Defective DNA repair hypothesis of ALS etiology</td>
<td>May have small role but not major factor</td>
<td>(Hayward et al., 1999)</td>
</tr>
<tr>
<td>APOE</td>
<td>Apolipoprotein E</td>
<td>Implicated in several other neurodegenerative disorders</td>
<td>Not associated with susceptibility. May be associated with age of onset, presentation and survival</td>
<td>(Al Chalabi et al., 1998; Droy et al., 2001; Moulard et al., 1996, Mui et al., 1995; Siddique et al., 1998b)</td>
</tr>
<tr>
<td>AR</td>
<td>Androgen receptor</td>
<td>Causes Kennedy spinal and bulbar muscular atrophy.</td>
<td>No association</td>
<td>(Garofalo et al., 1993)</td>
</tr>
<tr>
<td>CCS</td>
<td>Copper chaperone for superoxide dismutase</td>
<td>Gene responsible for copper insertion into SOD1</td>
<td>No association</td>
<td>(Silhatharoglu et al., 2002)</td>
</tr>
<tr>
<td>CNTF</td>
<td>Ciliary neurotrophic factor</td>
<td>Mice lacking CNTF develop mild, progressive motor neuron loss</td>
<td>Contradictory results</td>
<td>(Al Chalabi et al., 2003; DeChiara et al., 1995; Giess et al., 2002, Masu et al., 1993; Orrell et al., 1995; Takahashi et al., 1994)</td>
</tr>
<tr>
<td>CYP2D6</td>
<td>Cytochrome P450, subfamily III, polypeptide 6</td>
<td>Hypothesized poor metabolism of xenobiotics as risk factor</td>
<td>One reported association, not replicable</td>
<td>(Nicholl et al., 1999; Siddons et al., 1996)</td>
</tr>
<tr>
<td>DCTN1</td>
<td>Dynactin</td>
<td>Disruption of dynein/dynactin complex produces motor neuron disease phenotype in mice</td>
<td>One reported association, not yet replicated</td>
<td>(Munch et al., 2005)</td>
</tr>
<tr>
<td>DYNCH1</td>
<td>Dynein heavy chain</td>
<td>Mutations in Dnch1 result in progressive motor neuron degeneration in heterozygous mice, homozgyotes also have Lewy-like inclusion bodies,</td>
<td>No association</td>
<td>(Ahmad-Annuar et al., 2003; Hafezparast et al., 2003; Munch et al., 2004; Shah et al., 2006)</td>
</tr>
<tr>
<td>EAAT2</td>
<td>Excitatory amino acid transporter 2</td>
<td>Excitotoxicity hypothesized to result in motor neuron degeneration</td>
<td>No association</td>
<td>(Aoki et al., 1996; Flowers et al., 2001; Honig et al., 2000; Jackson et al., 1999; Lin et al., 1998; Meyer et al., 1998; Meyer et al., 1999)</td>
</tr>
<tr>
<td>HexA</td>
<td>Hexosaminidase A</td>
<td>HexA deficiency causes accumulation of ganglioside GM2 leads to neurodegeneration causing wide spectrum of neurological diseases</td>
<td>Oxidative stress is hypothesized to be implicated in neurodegeneration and misregulation of iron induces oxidative stress. Also, abnormal iron levels found in spinal cords of ALS patients</td>
<td>(Drory et al., 2003)</td>
</tr>
<tr>
<td>HFE</td>
<td>Haemochromatosis</td>
<td>Oxidative stress is hypothesized to be implicated in neurodegeneration and misregulation of iron induces oxidative stress. Also, abnormal iron levels found in spinal cords of ALS patients</td>
<td>Contradictory results</td>
<td>(Wang et al., 2004; Yen et al., 2004)</td>
</tr>
<tr>
<td>LIF</td>
<td>Leukaemia inhibitory factor</td>
<td>LIF is same cytokine family as CNTF, involved in motor neuron survival</td>
<td>One reported association, not yet replicated</td>
<td>(Giess et al., 2000; Meyer et al., 1995)</td>
</tr>
<tr>
<td>LOX</td>
<td>Lysyl oxidase</td>
<td>LOX is a copper containing enzyme and copper-induced cytotoxicity is a hypothetical mechanism of motor neuron degeneration</td>
<td>No association</td>
<td>(Chioza et al., 2001)</td>
</tr>
<tr>
<td>MAO-B</td>
<td>Monoamine oxidase B</td>
<td>MAO-B generates free radicals and these are implicated in neuronal damage</td>
<td>Association reported but p-values were weak</td>
<td>(Orru et al., 1999)</td>
</tr>
<tr>
<td>MAPT</td>
<td>Microtubule-associated protein tau</td>
<td>MAPT involved in other neurodegenerative diseases Guam variant of ALS has containing tau aggregates</td>
<td>Association with age at onset, not yet replicated</td>
<td>(Kowalska et al., 2003; Poorkaj et al., 2001)</td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Reason for Investigation</td>
<td>Significance</td>
<td>References</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Mito</td>
<td>Mitochondrial DNA deletions</td>
<td>Accumulation of mitochondrial DNA mutations associated with aging development of degenerative diseases; In ALS, abnormal mitochondria are often found in spinal motor neurons</td>
<td>Associations reported, but all studies very small</td>
<td>(Dhaliwal et al., 2000; Gajewski et al., 2003; Mawrin et al., 2003; Swerdlow et al., 1998; Wiedemann et al., 2002) (Mawrin et al., 2004; Ro et al., 2003)</td>
</tr>
<tr>
<td>NAIP</td>
<td>Neuronal apoptosis inhibitory protein Subunit 2 of mitochondrial NADH dehydrogenase</td>
<td>NAIP involved in related disease, spinal muscular atrophy</td>
<td>No association. Mutations in NAIP now considered specific for SMA</td>
<td>(Jackson et al., 1996; Kunst et al., 2000; Orrell et al., 1997a; Parboosingh et al., 1999)</td>
</tr>
<tr>
<td>ND2</td>
<td>Expression of ND2 found in Alzheimer's brains</td>
<td>No association</td>
<td>Significant association, tail domain deletions present in 1% ALS patients, not in controls. Findings have replicated in several studies Few mutations found, not a common cause of ALS</td>
<td>(Lin et al., 1992)</td>
</tr>
<tr>
<td>NEFH</td>
<td>Neurofilament, heavy chain</td>
<td>Neurofilament accumulation is a hallmark of ALS</td>
<td>New familial gene, not yet tested in sporadics</td>
<td>(Al Chalabi et al., 1999; Julien et al., 1995; Meyer et al., 1995; Rooke et al., 1996; Tomkins et al., 1998; Vechio et al., 1996)</td>
</tr>
<tr>
<td>PRPH</td>
<td>Peripherin</td>
<td>Transgenic mice over-expressing peripherin develop motor neuron degeneration</td>
<td>Weak association found, small study, not replicated</td>
<td>(Gros-Louis et al., 2004; Leung et al., 2004)</td>
</tr>
<tr>
<td>PSEN1</td>
<td>Presenilin-1</td>
<td>PSEN1 involved in apoptosis, a postulated mechanism for neuronal death</td>
<td>Association with lower motor neuron disease found in one small study, not yet replicated</td>
<td>(Panas et al., 2000)</td>
</tr>
<tr>
<td>PVR</td>
<td>Poliovirus receptor</td>
<td>Poliovirus attacks motor neurons selectively; enteroirna nucleic acids found in spinal cord of ALS patients</td>
<td>New familial gene, not yet tested in sporadics</td>
<td>(Saunderson et al., 2004)</td>
</tr>
<tr>
<td>SETX</td>
<td>Senataxin</td>
<td>Mutations in SETX cause autosomal dominant juvenile ALS (ALS4)</td>
<td>SMN protein levels associated with sporadic ALS</td>
<td>(Chen et al., 2004)</td>
</tr>
<tr>
<td>SMN1/2</td>
<td>Survival of motor neuron 1/2</td>
<td>Deletions and mutations of SMN genes cause spinal muscular atrophy</td>
<td>One case reported of young onset, slowly progressive upper and lower motor neuron syndrome with spastin mutation. No association with paraplegin</td>
<td>(Cocria et al., 2002b; Cocria et al., 2002a; Veldink et al., 2001; Veldink et al., 2005)</td>
</tr>
<tr>
<td>Spastin and paraplegin</td>
<td>Hereditary spastic paraparesis</td>
<td>Mutations in spastin and paraplegin are the most common causes of hereditary spastic paraparesis</td>
<td>Persyn is a member of the synuclein family, γ-synuclein. Mutations in α-synuclein found in Parkinson’s disease</td>
<td>(McDermott et al., 2003; Meyer et al., 2005)</td>
</tr>
<tr>
<td>SNCG</td>
<td>Persyn</td>
<td>Persyn is a member of the synuclein family, γ-synuclein. Mutations in α-synuclein found in Parkinson's disease</td>
<td>No association</td>
<td>(Flowers et al., 1999)</td>
</tr>
<tr>
<td>SOD1</td>
<td>Cu/Zn superoxide dismutase 1</td>
<td>Mutations in SOD1 account for 20% of familial ALS.</td>
<td>2–7% sporadic cases</td>
<td>(Andersen et al., 1997; Johnston et al., 2006; Jones et al., 1994b; Rosen et al., 1993; Siddique et al., 1991)</td>
</tr>
<tr>
<td>SOD2</td>
<td>Manganese superoxide dismutase</td>
<td>SOD2 is a related protein of SOD1, found in mitochondria</td>
<td>No association</td>
<td>(Tomkins et al., 2001)</td>
</tr>
<tr>
<td>VAPB</td>
<td>Vesicle-associated membrane protein-associated protein B</td>
<td>Mutations in VAPB cause an autosomal dominant, slowly progressive ALS (ALS8)</td>
<td>New familial gene, not yet tested in sporadics</td>
<td>(Nishimura et al., 2004b)</td>
</tr>
<tr>
<td>VDR</td>
<td>Vitamin D receptor</td>
<td>Lead exposure associated with ALS. Vitamin D can affect lead absorption and distribution</td>
<td>Not significant</td>
<td>(Kamel et al., 2002; Kamel et al., 2003)</td>
</tr>
<tr>
<td>VEGF</td>
<td>Vascular endothelial growth factor</td>
<td>Vascular endothelial growth factor</td>
<td>Association in some populations</td>
<td>(Lambrechts et al., 2003; Oosthuyse et al., 2001; Terry et al., 2004; Van Vught et al., 2005)</td>
</tr>
</tbody>
</table>

Table 1.4 ALS susceptibility loci
(Modified from Simpson et al., 2006)
1.7.6  **DYNC1H1 as a candidate susceptibility locus**

The cytoplasmic 1 dynein heavy chain 1 (DYNC1H1) has been implicated as a potential causative gene for motor neuron degeneration and possibly for ALS. The first evidence implicating DYNC1H1 as a locus for investigation in ALS came from mouse models of motor neuron degeneration generated by mutagenesis screen. The heterozygous 'Legs at odd angles' (Loa) mouse and allelic 'Cramping I' (Cral) mouse both displayed progressive locomotor defects. Although this did not affect lifespan, homozygous mice died within 24 hours of birth (Hafezparast et al., 2003). Histopathology, of the Loa mice showed significant anterior horn cell loss and deposition of ubiquitin, SOD1 and neurofilament in the remaining cells which also contained Lewy-like bodies.

1.7.6.1  **The cytoplasmic dyneins and dynactin**

The dyneins are large multi-subunit protein complexes that undertake a wide range of roles within the cell. They are microtubule minus-end-directed molecular motors that can be divided into two classes based on function: axonemal and cytoplasmic dyneins (reviewed in Gibbons, 1995; Vallee et al., 2004). Axonemal dyneins are responsible for the movement of cilia and flagella and cytoplasmic dyneins are involved in a range of microtubule-associated functions within the cell. Two cytoplasmic dynein complexes with distinct cellular functions exist, termed cytoplasmic dynein 1 and cytoplasmic dynein 2. Cytoplasmic dynein 2 is involved in intraflagellar transport (IFT), a process required for the assembly of cilia and flagella (reviewed in Cole, 2003). Cytoplasmic dynein 1 is the more abundant complex and is involved in a greater number of functions such as diverse as spindle-pole organization and nuclear migration during mitosis and the positioning and functioning of cellular organelles and compartments, including the endoplasmic reticulum (ER), Golgi apparatus and the nucleus. In addition cytoplasmic dynein 1 also provides the minus-end-directed transport of vesicles along microtubules, including endosomes and lysosomes, and retrograde axonal transport in neurons.

The cytoplasmic dynein 1 complex also interacts with other proteins such as LIS1, which is thought to modulate the enzymatic activity of DYN1H1 (Mesngon et al., 2006), and a second multimer, dynactin, which is itself comprised of at least seven different proteins including p22, p50 and p150 and possesses three functional domains: microtubule-binding, dynein-binding, and cargo binding (reviewed in Schroer, 2004). These domains allow dynactin to function both as an adaptor, associating the dynein motor with different cargoes (Holleran et al., 1998; Karki et al., 1999).
1998; Karki et al., 1999), and as an enhancer of dynein motor processivity is enhanced two- to four-fold over the distance travelled by a single dynein molecule alone (Culver-Hanlon et al., 2006; King et al., 2000).

1.7.6.2 Cytoplasmic dynein 1 heavy chain subunit and function in neurons

The cytoplasmic dynein 1 complex comprises several subunits: two large heavy chain polypeptides (~530 kDa each) which homodimerise to form the core of the complex and associated intermediate (~74 kDa), light intermediate (~33-59 kDa), and light chain polypeptides (~10-14 kDa) (see Figure 4.1). In the interest of brevity and as this thesis is largely focused on the dynein heavy chain, only the heavy chain will be discussed however several excellent reviews exist for the interested reader (see for example Pfister et al., 2005b).

The cytoplasmic dynein-dynactin complex performs multiple cellular roles and are essential protein complexes in higher eukaryotes; both fly and mouse knockouts are lethal in embryogenesis (Gepner et al., 1996; Harada et al., 1998). Within post-mitotic neurons cytoplasmic dynein is responsible for several essential functions including maintaining ER to Golgi traffic, vesicle transport, mRNA localisation and retrograde axonal transport. In motor neurons in particular, which possess axons that can exceed 1m in length and which form multiple synapses with other neurons and muscle cells, correct and efficient retrograde transport is essential. Cytoplasmic dynein utilises the axonal cytoskeleton within motor neurons to transport cargo such as organelles and structural and signalling proteins and neurotrophic factors from the axon termini to the perikaryon. Indeed, defects in the machinery that drives retrograde transport may inhibit neurotrophic factor signalling, and therefore lead to neuronal degeneration (Holzbaur, 2004). In support of this hypothesis, inhibition of retrograde transport in postnatal motor neurons by over-expression of a dynactin subunit results in motor neuron loss and muscle atrophy (LaMonte et al., 2002).

1.7.6.3 Cytoplasmic dynein and human motor neuron degeneration

The cytoplasmic dynein-dynactin complex has been implicated in several instances of human motor neuron degeneration. Puls and colleagues identified a G59S mutation in the p150Glued subunit of dynactin (DCTN1) in a family with a slowly progressive autosomal dominant form of motor neuron disease (Puls et al., 2003; Puls et al., 2005). Affected individuals with this mutation developed symptoms in early adulthood including: vocal fold paralysis causing breathing difficulty; progressive facial weakness, and weakness and atrophy in distal-limb muscles. The
G59S mutation resides in a highly conserved domain that interacts directly with microtubules and the microtubule plus end protein EB1 (Ligon et al., 2005) leading to decreased microtubule binding and enhanced dynein and dynactin aggregation (Levy et al., 2006), thus contributing to the degeneration of motor neurons.

Mutations in DCTN1 have also been identified in three FALS cases and one SALS case in a study of 250 ALS patients and 150 unrelated controls (Munch et al., 2004). In this study Munch and colleagues investigated all DCTN1 exons and identified a T1249I change in the SALS case and a M571T change in one of the FALS cases. The two remaining FALS cases were related, from the same kindred and possessed an autosomal dominant R785W mutation. Interestingly, two unaffected individuals from the same kindred also possessed the R785W mutation, suggesting that this mutation may display incomplete penetrance or that this mutation may not be responsible for disease and it is instead an unknown autosomal recessive gene defect or that DCTN1 represents a genetic risk factor, rather than a causative factor for ALS.
1.8 Human prion diseases

Prion diseases are a group of human and animal neurodegenerative conditions that have in common a key role for the prion protein in their pathogenesis. The archetypal prion disease, sheep scrapie, was first reported in the 18th century (McGowan, 1922) (the name is derived from the scratching of fence posts by diseased, ataxic animals), and since then several additional animal and human prion diseases have been identified. There are three aetiological categories of human prion disease: inherited, acquired (transmitted between humans or animals) and sporadic. Inherited prion diseases include Gerstmann-Sträussler-Scheinker (GSS) disease, fatal familial insomnia (FFI) and familial Creutzfeldt-Jakob disease (fCJD); acquired prion diseases include iatrogenic CJD, variant CJD (vCJD) and kuru; and sporadic disease is sporadic CJD (sCJD).

1.8.1 Aetiology of prion diseases

A key feature of the prion diseases is their transmissibility amongst and between species (Chandler L, 1961; Cuillé et al., 1936; Gajdusek et al., 1967). The discovery of the aetiology of prion diseases is a fascinating story (Collinge, 2005) and is based largely on the observations of disease transmissibility and the search for, and identification of, the transmissible species. A single protein, the prion protein (PrP), seen to copurify with infectivity has led to the development of a 'protein-only' hypothesis of prion disease. Under this hypothesis, the infectious agent of prion diseases is comprised of abnormal isoforms of PrP and these disease associated isoforms act as templates to promote the conversion of normal cellular PrP (PrP\textsuperscript{C}) to the pathological state PrP\textsuperscript{Sc} (Prusiner, 1982).

1.8.1.1 PrP and PRNP

The normal prion protein is a large protein 33-35 kDa whose sequence is highly conserved in most mammals (Wopfner et al., 1999). PrP\textsuperscript{C} is widely expressed in most adult tissues (Manson et al., 1992) however highest expression levels are seen in the central nervous system. The precise cellular function of PrP\textsuperscript{C} is still not known, however, as a glycosyl phosphatidyl inositol-anchored cell-surface glycoprotein, it has been speculated that PrP\textsuperscript{C} may have a role in cell adhesion or signalling processes.

Human and mouse genetics have made major contributions to prion disease research. Following the identification of PrP as the transmissible agent, partial sequencing of the protein allowed for the prediction of coding nucleotides and eventually to the cloning of the cognate human (PRNP) the and mouse (Prnp) prion protein genes on chromosomes 20 and 2 respectively (Chesebro et
al., 1985; Oesch et al., 1985; Robakis et al., 1986). Strong supportive evidence for the central role of PrP in prion disease has been provided by (i) the linkage of scrapie incubation time loci in the mouse to Prnp (Carlson et al., 1986) and (ii) the identification of mutations in PRNP linked to the inherited human prion diseases fCJD (Owen et al., 1989) and GSS (Hsiao et al., 1989).

The human PRNP gene on chromosome 20 comprises two exons (Puckett et al., 1991) and spans ~16kb. The open reading frame (ORF) of 759 nucleotides is entirely contained within the larger second exon. In the mouse, Prnp is located in the syntenic region of chromosome 2 and comprises three exons (with an additional short second exon compared to PRNP), with the ORF located in exon 3. The ORFs of mammalian prion genes are well conserved, generally exhibiting ~90% similarity. Several regions of human PrP are relevant to disease. The N-terminal domain (codons 51–91) encodes a 5-mer repeat region consisting of a nonapeptide followed by four identical octapeptides. Alterations in the number of repeats are found as polymorphisms and pathogenic mutations, but no point mutations or common SNPs have been found in this region. The C-terminal domain of PrP contains known point mutations, causing inherited prion disease, and a common coding polymorphism at codon 129 of PRNP (A385G) between methionine and valine has a critical role in susceptibility and modification of prion disease (discussed later).

### 1.8.2 Inherited prion diseases

Approximately 15% of all human prion diseases are inherited as autosomal dominant disorders and all can be accounted for by mutation of PRNP, of which over 30 different mutations have been described to date (Mead, 2006). Three types of pathogenic PRNP mutation exist in inherited prion disease: (i) point mutations leading to an amino-acid substitution or leading to (ii) a premature stop codon and (iii) insertion of additional octapeptide repeats (OPRI). With the exception of some OPRI cases and C-terminal point mutations, all mutations are fully penetrant.

Inherited prion disease is generally associated with an earlier age of onset and longer duration of illness than sporadic forms. However, the phenotype is highly variable ranging from prolonged slowly progressive dementia over 20 years to an aggressive disease indistinguishable from sporadic CJD. The spectrum of inherited prion diseases has been historically encapsulated by three clinical categories: GSS, FFI and fCJD, with features including slow progression of ataxia followed by later onset dementia in GSS; refractory insomnia, hallucinations, dysautonomia and motor signs in FFI; and rapidly progressive dementia, with myoclonus in fCJD. The phenotypic variability of inherited prion diseases is clearly illustrated by a large British 6-OPRI mutation.
family, where affected individuals show a range of pathological phenotypes from CJD to a slowly progressive dementia without characteristic neuropathology (Collinge et al., 1992). Due to this clinical variability, it has been difficult to estimate the incidence of inherited prion disease and PRNP mutation is usually sought as confirmation of disease.

1.8.3 Sporadic prion disease

Over 85% of human prion disease cases worldwide are sporadic. In the UK, approximately 50 cases of sporadic CJD are seen annually (which equates to ~1 per million of the population), however this may be underestimated as autopsy studies have shown that 40% of all pathological cases of CJD go clinically undiagnosed (Bruton et al., 1995). Disease incidence is approximately equal worldwide between the sexes and with an apparently random distribution. Age is a dominant risk factor for sCJD. The incidence of sCJD in younger adults aged <40 years is low, but above this age incidence increases and is seen to peak in the sixth decade. At older ages however, incidence appears to decline (Will et al., 1998). There are several lines of evidence which indicate that the epidemiology of sCJD is inconsistent with a single major environmental influence, such as animal to human transmission. For example, cases of sCJD are well recognised in countries that have never reported sheep scrapie (Masters et al., 1979).

In general, CJD is characterised by the rapid onset of neurological degeneration. This initially presents as dementia, followed by the development of movement disorders such as tremor, spasticity and rigidity. A common feature in almost all people with CJD is myoclonus, and most also display abnormal electroencephalogram recordings. Once symptoms are detected, the disease course is extremely aggressive (especially when compared to other neurodegenerative diseases such as Alzheimer’s and Parkinson’s), with average disease duration of 7.6 months, and the vast majority of patients (70%) die within 6 months (Collinge, 2001). At disease end stage, patients sink into akinetic mutism, with death usually occurring due to systemic or pulmonary infection. In terms of neuropathology, CJD displays extensive spongiform degeneration and neuronal loss, coupled with gliosis and deposits of the prion protein within the brain.

1.8.4 Acquired CJD

The most recent human prion disease to have been identified emerged in 1996 following an epidemic of the bovine prion disease, bovine spongiform encephalopathy (BSE), and was recognised as a novel clinico-pathological variant of CJD (Will et al., 1998). To date there have been 161 deaths from definite or probable vCJD (Department of Health CJD statistics, 2 July
Several lines of evidence, including geographical and temporal coincidence and molecular studies, support a causative link between vCJD and exposure to BSE-infected bovine tissue (Bruce et al., 1997; Collinge et al., 1996; Hill et al., 1997).

vCJD exhibits marked differences in disease profile to those associated with classical CJD or the other human prion diseases. The age at onset ranges from 16 to 51 years (mean 29 years), and the clinical course is unusually prolonged (9–35 months, median 14 months) (Collinge, 2001). The initial symptoms are behavioural, with disorders of movement such as ataxia occurring weeks to months later. Dementia and myoclonus (usually the initial symptoms in classical CJD) occur towards the later stages of the disease. The neuropathology of vCJD is also in marked contrast to most other human prion diseases, with spongiform degeneration concentrated in the basal ganglia and thalamus, and the presence of so called florid plaques, similar to other transmitted prion diseases including kuru and scrapie.

Following the BSE epidemic, there has been widespread concern over an associated vCJD epidemic. Estimates based on current vCJD figures (see Figure 1.9) suggest that the total epidemic may be small (Ghani et al., 2003). However, dietary exposure of the UK population to BSE prions is known to have been widespread and the precise disease incubation period of vCJD is not yet known. Estimates of incubation period based on evidence from the acquired prion disease epidemic kuru, where incubation times in excess of 50 years have been recorded in the absence of a species barrier, suggest that additional vCJD cases may be seen in the future (Collinge et al., 2006). Additionally, as evidence from mouse studies have shown (discussed later) disease incubation may be under the control of multiple genetic loci and the vCJD patients
identified could represent a distinct genetic subpopulation possessing short incubation alleles to BSE prions, with the susceptibility \textit{PRNP} genotype. Therefore, a human BSE epidemic may be multiphasic, and recent estimates of the size of the vCJD epidemic based on uniform genetic susceptibility could be substantial underestimations (Collinge \textit{et al.}, 2006). Until such human modifier loci are identified and their gene frequencies in the population can be measured the size of a potential epidemic will be difficult to model.

1.8.5 Kuru

Kuru first came to the attention of Western medicine in 1957, after it reached epidemic proportions in a defined population of Eastern Highlands of Papua New Guinea (PNG), and provides our largest experience of an acquired prion disease for epidemiological study. At its peak, kuru was responsible for 200 deaths per year and the disease predominantly affected women (it was the main cause of female mortality) and children of both sexes, with only 2\% of cases in adult men (Collinge \textit{et al.}, 2006)

Kuru mainly affected the people of the Fore linguistic group and also their neighbours with whom they intermarried and was transmitted through endocannibalism, as these communities practised the ritual consumption of dead relatives as a mark of respect and mourning. The original acquisition of the disease is thought to have been by means of the consumption of a single person with sporadic prion disease, with subsequent propagation by cannibalism to epidemic proportions (Attenborough \textit{et al.}, 1992). The first recorded cases of kuru were reported to the north of the Fore between 1920 and 1925. The disease spread southwards though the Fore where, based on living accounts (recorded by M. Alpers and J. Whitfield pers. comm.), cannibalism was common although practices varied. Different body parts were consumed by specific participants – the brain and other high prion titre organs were consumed by women and children. However, boys older than 6–8 years participated little in mortuary feasting as they ceased residing with their mothers and instead lived with male members of the group. Men rarely participated in mortuary feasts as traditionally this was considered emasculating and was considered to weaken a male’s power. This is likely to explain the differential age and sex incidence. During the peak of the kuru epidemic, Papua New Guinea and the Eastern Highlands came under Australian administrative control and cannibalism was effectively prohibited. With the abrupt cessation of cannibalism, children born after the late 1950s have been free of kuru, indicating that maternal transmission was unlikely to have occurred.
1.8.5.1 Clinical and pathological characteristics

Clinically, kuru presents initially as progressive ataxia (kuru is the Fore word for shaking or trembling) until the patient is incapacitated, and cognitive impairment is also a feature. Towards the final stages of the disease, dementia is common and incontinence is frequently present. Death usually occurs within 3 years, often through starvation, infection or through falling into cooking fires. At autopsy, the neuropathological hallmarks have many similarities to those of CJD, including spongiform degeneration, vacuolated neurons and astrocytosis.

1.8.5.2 Genetics of kuru

Since 1967 when Gajdusek, Gibbs and Alpers demonstrated that kuru could be transmitted to chimpanzee - the first transmission of a human prion disease which confirmed that these human diseases belonged within the same category as the animal prion disease- kuru has continued to illuminate our understanding of prion disease (Gajdusek et al., 1967). As with other human prion diseases, the major genetic determinant of kuru susceptibility is PRNP. As discussed in detail in Chapter 6, heterozygosity at codon 129 is a major susceptibility factor for kuru incubation time. Other loci have also been investigated for association with kuru incubation time but no associations have yet been seen (Collinge et al., 2006). These include: the PRNP haplotype, which has previously been associated with sCJD (Mead et al., 2001); the prion-like gene Dopple (PRND) to which no association to prion disease has yet been seen (Mead et al., 2000); APOE and HLA-DQB7 (see below). In addition, Mead and colleagues have shown, based initially on work with kuru exposed PNG samples, that the prion gene has undergone significant balancing selection both within PNG and possibly even worldwide (Mead et al., 2003) – this is also discussed further in Chapter 6.

1.8.6 Evidence for human genetic susceptibility to prion disease

1.8.6.1 Information from animal studies - inbred mouse lines

Inbred mouse lines are an established model for prion disease genetics and have been useful in determining the genetic contribution to disease such as incubation time*. Following inoculation with a particular prion strain, mean incubation times can vary from 100 to over 500 days in different inbred mouse lines. There are many factors that can influence incubation time including prion strain type, level of host PrPC, dose of infectious material and route of administration

* In experimental transmissions to mice this is measured in days from the time of prion inoculation to the onset of clinical signs.
incubation times are shorter when inoculation is administered directly into the brain compared to peripheral or oral routes). When experimental conditions are kept constant, incubation times are highly reproducible with small standard deviations. Between mouse lines however, incubation times can vary substantially: incubation times in SM/J and MAI/Pas mice, inoculated with Chandler/RML mouse adapted scrapie, vary from 133±1 to 360±11 days respectively (Lloyd et al., 2004). The major determinant of incubation time is the inbred mouse line used which implies a strong genetic effect.

Inbred mouse lines can be separated into two characteristic incubation groups: short incubation time (100-200 days) and long incubation time (>255 days). Over a period of a decade from 1988, several authors used classical genetic studies to link the genetic determinant of prion incubation time in different mouse strains to loci on chromosome 2 tightly linked to PRNP, which were eventually shown to be PRNP alleles (for review see Lloyd et al., 2005). In all, three PRNP alleles have been identified which differ at amino acid 108 and 189 of the protein product: PRNP$^a$ (108-Leu, 189-Thr) and PRNP$^b$ (108-Phe, 189-Val) are the short and long incubation alleles respectively (Westaway et al., 1987) and PRNP$^c$ (108-Phe, 189-Thr) is an intermediate incubation time allele (255±12 days) found only in MAI/Pas mice (Lloyd et al., 2004). However, within a given PRNP genotype a range of incubation times are still seen (Table 1.5) suggesting that many genes, other than PRNP, also have a role in determining prion disease incubation time.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Incubation time (days) + SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>a allele mice</td>
<td></td>
</tr>
<tr>
<td>NZW/OlaHsd</td>
<td>108 ± 1 (n=38)</td>
</tr>
<tr>
<td>SJL/OlaHsd</td>
<td>122 ± 1 (n=37)</td>
</tr>
<tr>
<td>FVB/NHsd</td>
<td>131 ± 1 (n=33)</td>
</tr>
<tr>
<td>SM/J</td>
<td>133 ± 1 (n=47)</td>
</tr>
<tr>
<td>SWR/OlaHsd</td>
<td>135 ± 1 (n=36)</td>
</tr>
<tr>
<td>RIIS/J</td>
<td>135 ± 1 (n=34)</td>
</tr>
<tr>
<td>C57BL6/JOlaHsd</td>
<td>143 ± 1 (n=42)</td>
</tr>
<tr>
<td>b allele mice</td>
<td></td>
</tr>
<tr>
<td>JU/FaCt</td>
<td>313 ± 3 (n=24)</td>
</tr>
<tr>
<td>VM/Dk</td>
<td>300 ± 3</td>
</tr>
<tr>
<td>l/Ln</td>
<td>255 ± 14</td>
</tr>
<tr>
<td>c allele mice</td>
<td></td>
</tr>
<tr>
<td>MAI/Pas</td>
<td>360 ± 11 (n=20)</td>
</tr>
<tr>
<td>C57 MAI-PRNP</td>
<td>255 ± 12 (n=9)</td>
</tr>
</tbody>
</table>

Table 1.5 Incubation times following intracerebral inoculation

Mice inoculated with Chandler/RML mouse adapted scrapie (Modified from Lloyd et al., 2005)

The first non-PRNP susceptibility locus was identified using congenic mice by Kingsbury and colleagues in 1983 (Kingsbury et al., 1983) within the MHC on chromosome 17. The MHC was
specifically targeted because the lymphoid system is known to have a role in the peripheral early replication of prions. The locus was designated prion incubation determinant-1 \( (\text{Pid}-1) \).

Incubation time is a continuous or quantitative trait which can be used as a measurable phenotype in linkage studies. Several whole-genome linkage studies of F2-intercrosses\(^\dagger\) and backcrosses\(^\ddagger\) to map prion disease QTLs have been conducted. Of the five studies conducted to date, over 20 loci have been mapped on eight mouse chromosomes (Lloyd \textit{et al.}, 2001; Lloyd \textit{et al.}, 2002; Manolakou \textit{et al.}, 2001; Moreno \textit{et al.}, 2003; Stephenson \textit{et al.}, 2000). In the largest such study by Lloyd and colleagues, an F1 intercross between CAST/Ei and NZW/OlaHSd mice was challenged with RML prion strain. Consistent with the segregation of multiple prion incubation genes, the resulting F2 generation showed a larger standard deviation in incubation time than the F1 generation (26 days and 7 days, respectively). Three highly significant regions were mapped to chromosomes 2 (including \textit{PRNP}), 11 and 12, with suggestive linkage on chromosomes 6 and 7. It is difficult to directly compare data from all five QTL studies, especially as they all involve different experimental conditions, however some loci are represented in more than one study such as the loci on chromosome 11 which are seen in two studies (Lloyd \textit{et al.}, 2001; Stephenson \textit{et al.}, 2000). Loci on chromosome 2 have also been identified in more than one study, which used identical CAST x NZW F2 intercrosses but inoculated with different prion strains (Lloyd \textit{et al.}, 2001; Lloyd \textit{et al.}, 2002). These two regions show considerable overlap and may represent the same underlying genes. The incubation times seen in the F2 mice were opposite to those observed in the parental lines, suggesting that a \textit{PRNP} independent locus on chromosome 2 (or other loci) exists. Such discordance between incubation times and \textit{PRNP} genotypes has been noted previously by Carlson and colleagues\(^1\) whose work investigating \textit{Prni} (a locus tightly linked to \textit{PRNP} and which was at the time thought to be the major determinant of prion incubation time) identified a long incubation mouse with a short incubation time \textit{PRNP} allele (Carlson \textit{et al.}, 1986). Carlson speculated that this mouse was a rare recombinant between \textit{Prni} and \textit{PRNP}, which could not be verified as the animal died before it could be progeny tested. Together these results raise the possibility that a susceptibility locus tightly linked but separate from \textit{PRNP} may exist.

In summary, studies of inbred mouse lines have identified \textit{PRNP} as the major determinant of prion incubation time. Other genes are also involved and although QTL studies have identified

\(^{\dagger}\) Two heterozygous mice from the first filial generation (F1) are mated to produce offspring with genotypes in Mendelian ratios

\(^{\ddagger}\) A heterozygous mouse is mated with a homozygous mouse of a parental line
several large regions of linkage, including a QTL linked to PRNP but independent of its coding sequence. The quantitative trait nucleotide (QTN) responsible has yet to be characterised.

1.8.6.2 HLA

The human leukocyte antigen (HLA) locus has been investigated as a susceptibility locus in several neurodegenerative diseases including idiopathic Parkinson’s disease (Lampe et al., 2003) and Alzheimer’s (Lehmann et al., 2006; Zareparsi et al., 2002). In mouse, the HLA locus Pid-1 has been associated with prion disease incubation time (Kingsbury et al., 1983). In human prion disease, the HLA allele HLA-DQ7, has been shown to be associated with resistance to vCJD but not sCJD by Jackson and colleagues (Jackson et al., 2001). In their study, Jackson et al. had access to a small sample set, analysing just 50 vCJD and 26 sporadic CJD cases. Subsequent studies by other investigators failed to reproduce this association but were also limited by available sample size (Laplanche et al., 2003; Pepys et al., 2003). The HLA-DQ7 allele has been shown to vary significantly in different worldwide populations (Galvani et al., 2005) may have confounded these previous studies.

Despite the lack of replication, the potential association of HLA-DQ7 with vCJD is informative with respect to the distinct mechanism of pathogenesis compared to sporadic CJD. In vCJD peripheral prion replication, which occurs in the lymphoreticular system and is where the product of the HLA locus is likely to act, may be the greatest determinant of susceptibility (i.e. incubation time, progression etc.).

1.8.6.3 PRNP codon 129 polymorphism

A common polymorphism at codon 129 of the PRNP, encoding either a methionine or valine, is a strong susceptibility factor for all three aetiological categories of human prion disease. Methionine homozygotes comprise 37% of the UK population whereas valine homozygotes comprise 12% (Owen et al., 1989). In sCJD, patients are largely methionine or valine homozygotes (Laplanche et al., 1994; Palmer et al., 1991; Windl et al., 1996; Windl et al., 1999), however an excess of methionine homozygotes over valine has been seen in sCJD (Salvatore et al., 1994). In addition, codon 129 genotype has also been shown to correlate directly with phenotype. Windl and colleagues found that 90.5% of patients diagnosed with definite CJD were methionine homozygous compared to 1.9% valine homozygous. In addition an atypical sCJD cohort comprised 41.4% methionine and 14.8% valine homozygotes.
Susceptibility to iatrogenic growth hormone associated prion disease is mediated by valine homozygosity at codon 129 (Collinge et al., 1991). In familial 6-OPRI cases, the age of onset for methionine homozygotes is approximately a decade earlier than that for heterozygotes (Poulter et al., 1992). Incubation time in kuru is shortened for homozygotes of either allele (Cervenakova et al., 1998) and all cases of variant CJD to date have been methionine homozygotes (Collinge, 2005).

The only non-codon 129 PRNP susceptibility allele identified to date has been a rare lysine to glutamine polymorphism at codon 219 (E219K) in Japan (Shibuya et al., 1998). In this population, where both the codon 129 and 219 polymorphisms are rare, all sCJD patients have been found to be homozygous for either allele at codon 129 but lack the 219 polymorphism, suggesting a protective effect.

1.8.6.4 PRNP regulatory elements, PRNP expression and prion disease

PRNP mediated disease susceptibility is likely to extend beyond the coding sequence of the gene and also involve proximal cis-acting regulatory elements including the PRNP promoter and upstream enhancers but also long range elements such as locus control regions. The evidence for prion disease susceptibility influenced by variation in PRNP regulatory regions and thus affecting expression, is reviewed in Chapter 7.
1.9 Aims of this thesis

The aims of this thesis are to apply the techniques and analyses available to investigate complex diseases, towards identifying susceptibility loci for two neurodegenerative diseases, ALS and prion disease. More specifically, this thesis aims:

- to characterise the genomic arrangement of \textit{DYNC1H1}, a candidate gene for human ALS supported by multiple lines of evidence
- to conduct a mutational screen of human \textit{DYNC1H1} exons, homologous to those harbouring mutations in two mouse models of progressive neurodegeneration similar to ALS, in an attempt to determine if this gene explains cases of FALS or other motor neuron diseases
- to conduct a candidate gene association study of \textit{DYNC1H1} with sporadic ALS, ascertaining SNPs for investigation, determining the LD and haplotype structure to define a minimum set of tSNPs which would allow an economic and effective means to test known and unknown genetic variation in the gene for an association with disease
- to clarify the nomenclature and genetic relationships of the cytoplasmic dynein subunits, to assist future studies and help devise a standardised nomenclature system. Investigating the genetic relationships of the subunit families in humans and mouse will also ensure that all homologous members of the cytoplasmic dyneins have been identified.
- to undertake a comprehensive analysis of kuru and Papua New Guinea codon 129 polymorphism data.
- to identify signatures of selection imposed by the kuru epidemic, experienced by the people of the Eastern Highlands of PNG, at codon 129 of \textit{PRNP} as a paradigm for additional candidate gene or whole genome studies. These candidate genes may be identified from mice models of prion incubation time or cell-based studies of prion incubation.
- to develop a PNG sample set that will facilitate future investigations of candidate genes for signatures of selection.
- to assess the feasibility of conducting genome-wide genotyping on archived kuru DNA samples and identify optimal analysis parameters
- to assess if copy number variation of \textit{PRNP} is a pathogenic or protective mechanism in both sporadic CJD and kuru.
2 Materials and methods

2.1 Materials

2.1.1 General chemicals and reagents

<table>
<thead>
<tr>
<th>Chemical/Reagent</th>
<th>Supplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>10x TBE</td>
<td>Life Technologies</td>
</tr>
<tr>
<td>Absolute ethanol (100%)</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>Agarose (electrophoresis grade)</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>Biotinylated anti-streptavidin antibody</td>
<td>Vector Laboratories</td>
</tr>
<tr>
<td>Bromophenol blue</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>Denhardt's solution (50x)</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>Dimethyl sulphoxide (DMSO)</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>dNTPs</td>
<td>Promega</td>
</tr>
<tr>
<td>EDTA (0.5M, pH 8.0)</td>
<td>Ambion</td>
</tr>
<tr>
<td>Ethanol (96-100%)</td>
<td>BDH Chemicals</td>
</tr>
<tr>
<td>Ethidium bromide</td>
<td>Sigma</td>
</tr>
<tr>
<td>Glycerol</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>H₂O (molecular biology grade)</td>
<td>Cambrex</td>
</tr>
<tr>
<td>Herring sperm DNA</td>
<td>Promega</td>
</tr>
<tr>
<td>Human Cot-1 DNA</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>Hyperladder (I, IV and V)</td>
<td>Bioline</td>
</tr>
<tr>
<td>MES hydrate SigmaUltra</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>MES sodium salt</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>MGB probes</td>
<td>Applied Biosystems</td>
</tr>
<tr>
<td>Primers</td>
<td>Sigma-Genosys</td>
</tr>
<tr>
<td>Sodium acetate (NaAce)</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>Sodium dodecyl sulphate (SDS)</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>SSPE (20x)</td>
<td>Cambrex</td>
</tr>
<tr>
<td>Streptavidin, R-phycoerythrin conjugate (SAPE)</td>
<td>Invitrogen</td>
</tr>
<tr>
<td>TE buffer, reduced EDTA (10mM Tris HCl, 0.1mM EDTA, pH 8.0)</td>
<td>TEKnova</td>
</tr>
<tr>
<td>TMACL (5M)</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>Tween-20 (10%)</td>
<td>Pierce</td>
</tr>
</tbody>
</table>
2.1.2 Prepared solutions

<table>
<thead>
<tr>
<th>Loading Buffer</th>
<th>Sodium Dodecyl Sulphate (SDS), 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5ml 1M Tris-HCl pH7.6</td>
<td>100g SDS per 1ml autoclaved water</td>
</tr>
<tr>
<td>25ml glycerol</td>
<td>Stored at room temperature</td>
</tr>
<tr>
<td>0.5ml 10% SDS</td>
<td></td>
</tr>
<tr>
<td>0.05g bromophenol blue</td>
<td></td>
</tr>
<tr>
<td>Made up to 50ml with ddH₂O</td>
<td></td>
</tr>
<tr>
<td>Stored at room temperature</td>
<td></td>
</tr>
</tbody>
</table>

2.1.3 Commercial kits

| Affymetrix DNA Amplification Clean-Up Kit                                       | Affymetrix                                                               |
| Affymetrix GeneChip 500K Assay                                                 | Affymetrix                                                               |
| Better Buffer                                                                   | Microzone                                                                 |
| BigDye Terminator Ready Reaction kit                                            | Applied Biosystems                                                       |
| Clontech TITANIUM Taq buffer                                                   | Clontech                                                                  |
| Clontech TITANIUM Taq DNA polymerase                                           | Clontech                                                                  |
| DYEEnamic ET Dye Terminator kit                                                | Amersham Pharmacia Biotech                                               |
| G-C Melt                                                                        | Clontech                                                                  |
| GeneScan 2500 TAMRA Red                                                         | Applied Biosystems                                                       |
| HotStartTaq DNA Polymerase                                                     | Qiagen                                                                    |
| MegaBACE ET400-R Size Standard                                                 | Amersham Pharmacia Biotech                                               |
| MegaBACE Loading Solution                                                      | Amersham Pharmacia Biotech                                               |
| MegaBACE Long Read Matrix                                                      | Amersham Pharmacia Biotech                                               |
| MegaBACE LPA Buffer                                                            | Amersham Pharmacia Biotech                                               |
| MegaMix Blue                                                                   | Microzone                                                                 |
| MicroClean PCR purification kit                                                | Microzone                                                                 |
| QIAamp 96 DNA Blood Mini kit                                                   | Qiagen                                                                    |
| QIAamp DNA Blood Mini kit                                                      | Qiagen                                                                    |
| QuantiTect Probe PCR Master Mix                                                | Qiagen                                                                    |
2.1.4 Restriction enzymes and ligases

<table>
<thead>
<tr>
<th>Enzyme*</th>
<th>Buffer/conditions</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNaseI</td>
<td>-</td>
<td>New England Biolabs (NEB)</td>
</tr>
<tr>
<td>Hinfl</td>
<td>2</td>
<td>New England Biolabs (NEB)</td>
</tr>
<tr>
<td>NspI</td>
<td>2 + BSA</td>
<td>New England Biolabs (NEB)</td>
</tr>
<tr>
<td>PvuII</td>
<td>2</td>
<td>New England Biolabs (NEB)</td>
</tr>
<tr>
<td>T4 DNA Ligase</td>
<td>T4 DNA Ligase Buffer</td>
<td>New England Biolabs (NEB)</td>
</tr>
</tbody>
</table>

2.1.5 Equipment

8 Strip PCR tubes and caps, thin-walled  
96-well PCR plates  
ABI Prism 377 DNA Sequencer  
Allegra 25R centrifuge for 96-well plates  
Aluminium adhesive PCR foil roll  
DNA Engine Tetrad PTT-225 Peltier Thermal Cycler  
Electrophoresis power packs  
Electrophoresis tanks (Horizon 11.14)  
Eppendorf 5415R Microcentrifuge  
Eppendorf Comfort Thermomixer 1.5ml  
GeneChip Hybridization Fluidics Station 450  
GeneChip Hybridization Oven 640  
GeneChip Scanner 3000 7G  
Gilson pipettes  
Grant JB5 waterbath  
MegaBACE 1000 DNA Analysis System  
MicroAmp Fast Optical 96-well Reaction Plates  
MicroAmp Optical 96-well Reaction Plates  
MicroAmp Optical Adhesive Film  
Microtubes, hydophobic 0.65ml  
PCR seal film  
Pipette tips  
QIAvac 96 vacuum manifold  
Raven incubator  
Sigma 6K 15 plate rotor 2 x 96-well plate

* updated nomenclature used, see (Roberts et al., 2003).
Sigma 6K 15 refrigerated centrifuge  
Sigma Laborzentrifugen GmbH

Ultrospec 2000 UV visible spectrophotometer  
Pharmacia Biotech

Vacuum pump  
KNF Neuberger

2.1.6 Photography

Gel Doc EQ UV-transilluminator  
Bio-Rad Laboratories

Thermal paper for Mitsubishi video printer  
Bio-Rad Laboratories

2.1.7 Software and websites

Bioinformatic-Harvester  
http://harvester.embl.de/

BLAST  

ClustalW  
http://www.ebi.ac.uk/clustalw (Thompson et al., 1994)

Ensembl Fugu  
http://www.ensembl.org/Fugu_rubripes

Genome Browser  

GDAS  
GeneChip DNA Analysis Software (Affymetrix)

HapMap  
http://www.hapmap.org

HGNC  
http://www.genenames.org/

MGI  
http://www.informatics.jax.org/

MultiPipmaker  
http://pipmaker.bx.psu.edu/cgi-bin/multipipmaker

NCBI  

NEBcutter  

Nspl library files  
GeneChip Human Mapping 500K Set library files (Affymetrix)

PHASE  

PHYLIP  

PL-EM  
http://www.people.fas.harvard.edu/~junliu/plem (Qin et al., 2002)

PLINK  
http://pngu.mgh.harvard.edu/~purcell/plink/

Power Calculator  
http://pngu.mgh.harvard.edu/~purcell/gpc/ (Purcell et al., 2003)

Primer Express v2.0  
Applied Biosystems

Primer3  
http://frodo.wi.mit.edu/cgi-bin/primer3/primer3 www.cgi

PubMed  

RepeatMasker  
http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker

Sequence Analyzer  
Molecular Dynamics (Wu and King, 2003)
2.2 Samples

2.2.1 Healthy population DNA samples

DNA samples from healthy individuals were obtained from the following sources: 33 European trios were obtained from the Centre d'Etude du Polymorphisme Humain (CEPH) and 48 unrelated Japanese and 49 unrelated West African (Cameroonian) individuals, were obtained from the laboratory of Professor David B. Goldstein, Institute for Genome Sciences and Policy, Centre for Population Genomics and Pharmacogenetics, Duke University, North Carolina, USA. DNA samples were collected by the Goldstein laboratory with signed consent or were anonymous legacy collections provided by collaborators from other academic institutions.

480 UK White Caucasian control samples (plates HRC-1 to 5), were obtained from the European Collection of Cell Cultures (ECACC). A further 90 UK Caucasian control samples used for whole genome SNP genotyping were collected by Dr. Simon Mead, MRC Prion Unit, London, UK from the West End Blood Donor Centre, c/o Dr Jean Harrison, London, UK. North American control samples comprising of both unaffected spouses from Caucasian familial ALS pedigrees and other Caucasian individuals were collected with signed consent by Professor Robert H. Brown Jr., Day Neuromuscular Research Laboratory, Massachusetts General Hospital, Massachusetts, USA or provided to R.H. Brown by collaborators from other North American academic and clinical institutions.

Chimpanzee DNA was extracted from blood provided by the Institute of Zoology, London Zoo, Regents Park.

2.2.2 Patient/case DNA samples

Sporadic ALS DNA samples were obtained from patients seen by Professor R.H. Brown at Massachusetts General Hospital or provided, to R.H. Brown by collaborators from various academic and clinical institutions across North America. All sporadic ALS patient samples were obtained with informed consent. Patients were diagnosed by El Escorial criteria and had not been
screened for \textit{SOD1} mutations. The patient cohort comprised of 134 females and 147 males with an average age at diagnosis of 46 years (range 24 to 79 years). 34\% of patients were diagnosed with early onset in the lower extremities, 35\% with upper onset, 30\% with bulbar onset and site of onset in the remaining 2\% was unknown.

Motor neuron disease samples screened for \textit{DYNC1H1} mutations were index cases from pedigrees with familial history of disease. ALS samples were collected by Professors Pamela Shaw, Sheffield University, Sheffield, UK and Karen Morrison, Department of Neurology, University of Birmingham, Birmingham, UK and Dr Richard Orrell, Department of Clinical Neuroscience, University College London, London, UK. All ALS samples were Caucasian adults with definite or probable ALS, without detectable \textit{SOD1} mutation. Spinal muscular atrophy (SMA) samples were adult and juvenile onset cases without \textit{SMN} deletions provided by Professor Morrison. All SMA cases were Caucasian, except 2 Indian and 1 Arab case (K. Morrison pers. comm.). Hereditary Spastic Paraplegia (HSP) samples were British Caucasian index cases from 32 different families (20 recessive inheritance families and 12 dominant inheritance families).

Sporadic CJD samples were obtained from patients seen at the Institute of Neurology, Queen Square, London, UK, with signed consent. Prion disease was confirmed post-mortem by the histological identification of spongiform change, astrocytosis, neuronal loss and proteinase-resistant \textit{PrP}^\text{Sc} deposition in the brain. Of prion disease cases, sporadic CJD is a diagnosis of exclusion and diagnosis was made in the absence of histopathological and \textit{PrP}^\text{Sc} strain-type features of variant CJD, the absence of mutation of the \textit{PRNP} open reading frame seen in inherited prion disease and the absence of obvious exposure to infectious human prion material.

\subsection{Papua New Guinea Samples}

Papua New Guinean samples were obtained as frozen DNA and/or blood archives from the MRC Prion Unit, London, UK, or from the Papua New Guinea Institute of Medical Research (PNGIMR), Goroka, PNG. All studies were approved by the Papua New Guinea Medical Research Advisory Committee and by the local UK research ethics committees. The full participation in the project from the communities, which was critical with respect to the ethics and operation of the study, was established and maintained through discussions of the joint MRC-PNGIMR field team with village leaders, communities, families, and individuals. Field studies followed the principles and practice of the Papua New Guinea Institute of Medical Research.
2.3 Methods

2.3.1 DNA isolation from chimpanzee blood

DNA was extracted from whole blood using QIAamp DNA Blood Mini Kit (Qiagen) following its associated protocol. The basis of the procedure is salt deproteinisation and is summarised as follows: 200μl whole blood was added to 20μl QIAVEN Protease in a 1.5ml microcentrifuge tube – all centrifugation steps were carried out in a microcentrifuge (Eppendorf). The tube was gently agitated to ensure the proper mixing of the blood and enzyme. 200μl Buffer AL was added and the sample pulse-vortexed for 15 seconds to ensure efficient lysis of all cells. The protease and buffer mixture is required to lyse cells in the sample and degrade all proteins, including those bound to DNA. The buffer has a high salt content and optimum pH to precipitate protein out of solution and away from the DNA. The mixture was incubated in a waterbath at 56°C for 10 minutes and tube centrifuged briefly to remove liquid condensed on the lid. 200μl ethanol (96-100%) (BDH Chemicals) was added and the sample was again pulse-vortexed for 15 seconds and centrifuged briefly.

The sample was transferred into a QIAamp Spin Column placed in a 2ml collection tube, sealed and centrifuged at 6000 relative centrifugal force* (rcf), for 1 minute. The silica gel membrane within the spin column absorbs DNA but does not retain protein or other contaminants within the lysate. The addition of ethanol to the lysate optimises DNA binding to the membrane. The filtrate was discarded and the QIAamp Spin Column was placed in a clean 2ml collection tube and 500μl Buffer AW1 was added. Buffers AW1 and AW2 are both ethanol based wash buffers used to remove residual contaminants without affecting DNA binding to the spin column membrane. The QIAamp Spin Column was again centrifuged at 6000 rcf for 1 minute and the collection tube containing the filtrate was again discarded. The QIAamp Spin Column was placed in a clean 2ml centrifuge tube and 500μl Buffer AW2 was added. The column was centrifuged at 20,000 rcf for 3 minutes and then placed in a clean 1.5ml microcentrifuge tube and the collection tube containing filtrate was discarded. 200μl Buffer AE was added to the column to elute the DNA. The column was incubated at room temperature for 5 minutes and then centrifuged at 6000 rcf for 1 minute and the primary eluate containing the DNA was collected and stored at -20°C. The QIAamp Spin Column was again placed in a clean 1.5ml centrifuge tube for a second elution to increase the total DNA yield. Again, 200μl Buffer AE was added, incubated at room temperature for 5 minutes and centrifuged at 6000 rcf for 1 minute. The second elute was also stored at -20°C.

* rcf to rpm conversion: \( \text{rcf} = 1.12 \times r \times (\text{rpm}/1000)^2 \), where \( r \) is rotor radius and \( \text{rpm} \) revolutions per minute.
2.3.2 DNA isolation from human blood

DNA was extracted from whole blood using QIAamp 96 DNA Blood Mini Kit (Qiagen) following its associated protocol. The kit is identical to the QIAamp DNA Blood Mini Kit but spin columns are in a 96-well plate format for increased throughput and a modified protocol summarised as follows: 20µl QIAGEN Protease was aliquoted to each well of a 96-round-well block (Figure 2.1) and 200µl whole blood was then added. 200µl Buffer AL was added to each sample and the wells were sealed using caps for blocks and tubes. Care was taken not to wet the well-rims at this stage to prevent caps from loosening in later steps. The sample was then mixed thoroughly by holding the round-well block with both hands and shaking up and down vigorously for 15 seconds. The block was centrifuged briefly at 2200 rcf (all centrifugation steps were carried out using the Sigma 6K 15 refrigerated centrifuge set at 40°C, unless otherwise stated) and then incubated at 70°C in a dry incubator for 10 minutes. A large heat-resistant weight was placed on top of the block to prevent the caps from popping off during the incubation and the contents evaporating. After incubation, the round-well block was briefly centrifuged again at 2200 rcf. 200µl ethanol (96–100%) was then added to each well of the block and the wells sealed with new caps for blocks and tubes. The block was again vigorously mixed by shaking for 15 seconds, and briefly centrifuged at 2200 rcf.

A QIAamp 96 plate was placed on top of an S-Block (Figure 2.1) and both were marked for later identification. The contents of the round-well block (620µl per well) was added to the QIAamp 96 plate. Care was taken not to wet the rims of the wells to avoid aerosol formation during

Figure 2.1 Overview of the QIAamp 96 DNA Blood Mini kit procedure
Adapted from the QIAamp DNA Blood Mini Kit protocol (Qiagen)
centrifugation. The QIAamp 96 plate was sealed with an AirPore Tape sheet and was placed with the S-Block into the centrifuge rotor bucket and centrifuged at 5796 rcf for 4 minutes.

The S-block was emptied of lysate and 500μl Buffer AW1 added to each well of the QIAamp 96 plate. The plate was sealed with a new AirPore Tape sheet and centrifuged at 5796 rcf for 2 minutes. Again the S-Block was emptied and 500μl Buffer AW2 was added to each well of the QIAamp 96 plate. The plate and S-Block were centrifuged at 5796 rcf for 15 minutes, without AirPore Tape (which was omitted to allow the heat generated during centrifugation to evaporate residual ethanol from Buffer AW2 in the sample).

To elute the DNA a QIAamp 96 plate was placed on top of a rack of elution MicroTubes (Figure 2.1) and 200μl Buffer AE was added using a multichannel pipette. The plate was sealed with a new AirPore tape sheet and incubated for 5 minutes at room temperature. The plate and MicroTubes rack were then centrifuged at 5796 rcf for 4 minutes. A further 200μl Buffer AE was added and the plate centrifuged at 5796 rcf for another 4 minutes to obtain an extra 20% yield.

2.3.3 Determination of DNA concentration and purity
DNA concentration and purity was determined using spectrophotometric analysis, where necessary. An appropriate dilution of DNA was prepared and absorption readings were taken at 260nm and 280nm, using either a Ultrospec 2000 Ultra Violet (UV) visible spectrophotometer (Pharmacia Biotech) or a NanoDrop 1000 spectrophotometric analyser (NanoDrop Technologies). The following equations were used:

DNA purity = $A_{260}/A_{280}$ (an ideal $A_{260}/A_{280}$ ratio was between 1.8 to 2.0)

DNA concentration (μg/μl) = $A_{260} \times$ dilution factor $\times$ 50μg/ml

2.3.4 Primer design
Primers were designed for Polymerase Chain Reaction (PCR) amplification, genotyping and sequencing reactions. Genomic sequence for the region of interest was obtained from National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov). Amplicon sequence and flanking primer binding sequence was examined for DNA repeats using RepeatMasker (www.repeatmasker.org/cgi-bin/WEBRepeatMasker) and potential primer sequences were designed to avoid annealing to repetitive DNA. Primers for PCR and sequencing were designed using Primer 3 software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3 www.cgi) or designed by eye. All primers were made by Sigma-Genosys.
2.3.4.1 Real-time PCR primers and probe design

Real-time primers and probes were designed using Primer Express v2.0 (Applied Biosystems) and all real-time probes were designed as Minor Groove Binder (MGB)\(^*\). Primers and probes were designed based on the following principles:

- The template strand was selected to increase the cysteine (C) content of the probe rather than the guanine (G) content.
- Primers were designed to be as close to the probe as possible with an overall amplicon length of ~50-150bp.
- Minimum probe length was 13 nucleotides.
- Probe/primer G-C content was maintained at a 30–80% range.
- Probes with a guanine nucleotide on the 5' end were avoided.
- Primers were limited to no more than two G and C residues within the last five nucleotides at 3' end.
- Probes/primers containing runs of an identical nucleotides were avoided (especially four or more guanine’s)
- Probe melting temperatures (T\(_m\)) were between 65–67 °C.
- Primer T\(_m\)'s were between 58–60 °C.

Primers were made by Sigma-Genosys and probes by Applied Biosystems. Probes were divided into 100μl aliquots upon receipt and care was taken to avoid excessive freeze-thawing and exposure to ultra violet (UV) light – which promotes aberrant cleavage of the end-labelled fluorescent reporter dyes.

2.3.5 Polymerase chain reaction

All PCRs were carried out using the following protocol in 0.65ml microtubes (Scientific Specialities Inc), 8 well PCR strips (Scientific Specialities Inc) or 96-well plates, unless otherwise stated: A PCR master mix was prepared containing 1 x PCR Buffer (supplied with HotStarTaq kit), 200μM dNTPs, 1μM forward primer, 1μM reverse primer, 0.2U/μl HotStarTaq DNA Polymerase (Qiagen) and distilled water, made up to a final reaction volume of 20μl. The master mix was then divided into 18μl aliquots using a Gilson Distriman repeater-pipette and 1μl DNA (~100ng/μl concentration) was added to each tube/well with one well assigned as a non-template, water control.

Thermal cycling conditions on the DNA Engine Tetrad Thermal Cycler (MJ Research) were as follows: HotStarTaq polymerase activation at 95°C for 1 min and then 35 cycles of: DNA template

\(^*\) MGB probes use non-fluorescent quenchers, which allow a more precise measurement of reporter dye compared to conventional probes. They also bind the DNA minor groove, increasing their melting temperature (T\(_m\)) and allowing the use of shorter probes, with better allele discrimination properties.
denaturation at 94°C for 30 seconds, primer annealing at 60°C (or relevant temperature listed in table 3.1) for 30 seconds, extension 72°C for 1 minute, and a final extension step at 72°C for 5 minutes. After amplification, 4μl PCR product was mixed with 4μl loading buffer and electrophoresed on a 2% agarose gel; 2 grammes (g) agarose (Invitrogen) dissolved in 100ml 1x TBE (Life Technologies), containing 0.1μg/ml ethidium bromide (Sigma) and visualised by UV-transillumination. All agarose gels were 2% unless otherwise stated. 4μl of the appropriately sized DNA ladder of was used to estimate the specificity of PCR amplicons.

PCRs for genotyping microsatellites, a54000*, a108990*, D20S889, D20S482 and D20S97 were carried out in 15μl reactions containing 13.7μl MegaMix Blue (Microzone), 0.15μl forward primer (50pm/μl), 0.15μl reverse primer (50pm/μl) and 1μl template DNA (~50ng/μl). Cycling conditions on the Tetrad thermal cycler were polymerase activation step of 95°C for 3 minutes and then 34 cycles of: DNA template denaturation at 95°C for 30 seconds, primer annealing at 55°C (or relevant temperature) for 40 seconds, 72°C extension step for 45 seconds, and a final extension step of 72°C for 5 minutes and 15°C for 2 minutes. PCR products were electrophoresed as described above but without addition of loading buffer.

2.3.6 Purification of PCR amplicons for sequencing

PCR products were purified using microCLEAN (Microzone), to remove reaction buffers, enzymes, primer dimers and unincorporated primers and dNTPs, following the manufacturer’s protocol. Briefly, an equal volume of microCLEAN was added to a 96-well plate containing PCR products and incubated at room temperature for 5 minutes. The plate was centrifuged at 3000 rcf for 40 minutes and then inverted onto absorbent paper and pulsed at 100 rcf to discard the supernatant. The purified DNA pellet was then resuspended in 20μl sterile distilled water.

2.3.7 Sequencing of purified templates

Automated fluorescent sequencing was carried out with the BigDye Terminator Ready Reaction Kit (Applied Biosystems), DYEnamic ET Dye Terminator Kit (Amersham Pharmacia) and 1μM PCR primers. For the BigDye Terminator Ready Reaction Kit, a sequencing master-mix was made containing (per reaction) 1μl BigDye terminators, 5μl BetterBuffer (Microzone), 0.5μl forward/reverse primer (1μM), and 7μl sterile distilled water. A Gilson Distriman repeater-pipette was used to aliquot 13.5μl BigDye master-mix into wells of a skirted 96-well plate, to which 1.5μl purified PCR products were added, for a total reaction volume of 15μl. Cycling conditions were 96°C for 30 seconds, followed by 30 cycles of 50°C for 15 seconds and 60°C for 3 minutes and a final hold step of 15°C for 5 minutes.

* From (Mead, 2002)
The DYEnamic ET Dye Terminator Kit sequencing master-mix contained, per reaction, 2\(\mu\)l ET terminators, 6\(\mu\)l BetterBuffer and 4\(\mu\)l forward/reverse primer (1\(\mu\)M). 12\(\mu\)l ET terminators master-mix was aliquoted into wells of a skirted 96-well plate, to which 8\(\mu\)l purified PCR products were added, for a total reaction volume of 20\(\mu\)l. Cycling conditions were 25 cycles of 95°C for 20 seconds, 50°C for 15 seconds and 60°C for 1 minute, with a final hold step of 10°C for 10 minutes.

2.3.8 Post-reaction clean-up with ethanol/salt precipitation

Sequencing products were precipitated to remove reaction buffers and unincorporated terminators to avoid tall early peaks, often termed "terminator blobs". 55\(\mu\)l 100% absolute ethanol (Sigma-Aldrich) and 1\(\mu\)l 3M sodium acetate (Sigma-Aldrich) was added to 20\(\mu\)l reaction and the mixture was chilled on ice for 15 minutes. The plate was centrifuged at 3000 rcf for 45 minutes to pellet the DNA and the supernatant discarded by inverting onto absorbent tissue paper and pulse-centrifuging at 100 rcf for 1 minute. 150\(\mu\)l 70% ethanol was added to wash the pellets and the plate was centrifuged at 3000 rcf for 10 minutes. Again, the plate was inverted onto absorbent tissue paper, pulse-centrifuged at 100 rcf for 1 minute to remove the supernatant and the pellet was allowed to air dry for 10 minutes.

2.3.9 DNA sequencing on the MegaBACE 1000 DNA Analysis System

The automated MegaBACE 1000 DNA Analysis System (Amersham Pharmacia) was used to sequence DNA. The MegaBACE uses capillary array electrophoresis to perform fragment size separation of fluorescently labelled DNA samples with a confocal optical system to collect data. The manufacturer’s guidelines and protocols were followed unless otherwise stated. Precipitated DNA with fluorescently labelled terminators, was resuspended in 10\(\mu\)l MegaBACE Loading Buffer (Amersham Pharmacia), vortexed thoroughly to ensure the complete solublization of the DNA and then pulse-centrifuged to return the liquid to the bottom of the 96-plate wells.

MegaBACE Long Read Sequencing Matrix (Amersham Pharmacia) was used to pressure-fill the capillary array with sieving matrix to separate DNA fragments of varying sizes. Sequencing products were electrokinetically injected into the capillary array by a potential difference of 3kV for 40 seconds and electrophoresis was carried out in 1x LPA Buffer (Amersham Pharmacia) at 9kV for 100 minutes. Laser excitation of the fluorescently labelled samples yielded data as an electropherogram for each capillary/sample which were analyzed in forward and reverse directions using the Sequence Analyser v3.0 package (Molecular Dynamics), and inspected by eye.
2.3.10 SNP genotyping by restriction digestion

SNPs within DYNC1H1 were tested for alteration of restriction enzyme recognition site using WebCutter (http://www.firstmarket.com/cutter/cut2.html) and NEBcutter (http://tools.neb.com/NEBcutter2/index.php). For SNPs that modified restriction enzyme recognition sites, PCR primers were designed to amplify a region of approximately 500bp around the polymorphism. PCR was performed as previously described in a 20μl reaction volume of which 4μl was electrophoresed on 2% agarose gel to check quality.

Restriction digestion was performed in 0.65ml microtubes in a total 20μl reaction volume, using 16μl un-purified PCR product and a 4μl restriction master-mix based on 1μl buffer recommended by manufacturer and 1-10 units of restriction enzyme per well. The amount of restriction enzyme required per SNP was estimated by initially performing a restriction digest on a subset of samples with only 1 unit of enzyme per sample. If this led to partial digestion of the amplicon after 3 hours, the experiment was repeated by adding additional enzyme until complete digestion was achieved. Digested amplicons were electrophoresed on 2% agarose and visualised as described previously.

The efficacy of SNP genotyping by restriction digest was tested by assaying a panel 25 CEPH samples, with a priori genotypes distinguished by DNA sequencing, in a blind test randomised by a second operator. Genotyping results from the blind test were then matched to sequence verified genotypes by the second operator and an efficacy score as a percentage was calculated.

2.3.11 Microsatellite genotyping

PCRs for microsatellites D20S482, D20S97, D20S889 and 108991 were initially multiplexed (each forward primer was tagged with 6-FAM, HEX, TET, 6-FAM fluorophores respectively; Sigma-Genosys) for the same DNA sample by mixing PCR products in the ratio 1:2:2:6 respectively, in a separate 96-well plate. Multiplexed PCRs comprised a total volume of 22μl and were made up to a total volume of 90μl with dd.H2O. 45μl aliquots were stored at -20°C and the remaining 45μl were carried forward for clean-up. D20S97-HEX forward primer was eventually replaced by D20S97-6-FAM due to a low signal affected by spectral overlap* with microsatellite 108991 which had a similar sized PCR product. PCR clean-up was carried using ethanol/salt precipitation as describe above for post-sequencing reactions, with the following amendments: 45μl of ethanol was mixed with 1μl NaAce and aliquoted into 45μl multiplexed PCR sample. Pellets were resuspended in 10μl dd.H2O.

* spectral overlap refers to the emission spectrum of one fluorophore overlapping with that of another, The excitation of one fluorophore therefore causes an erroneous recording of both fluorophores.
Microsatellites were genotyped using the MegaBACE 1000 DNA Analysis System. A loading solution of 5.6µl MegaBACE Loading Solution, 0.4µl MegaBACE ET400-R size standard and 2µl dd.H₂O was aliquoted into a new 96-well plate and 2µl resuspended PCR product added. The plate was vortexed briefly at medium speed and centrifuged briefly at 3000 rcf. The MegaBACE 1000 DNA Analysis System was set up following the manufacturer’s guidelines and as described above. Multiplexed PCR products were electrokinetically injected into the capillary array by a potential difference of 3kV for 60 seconds and electrophoresis was carried out in 1 x LPA Buffer at 9kV for 70 minutes. Results were analysed using the Genetic Profiler v1.5 package (Molecular Dynamics).

2.3.12 Affymetrix GeneChip 250k NspI Assay

Whole-genome SNP genotyping was achieved using the NspI array of an Affymetrix GeneChip 500k kit. The manufacturer’s protocol for the Affymetrix GeneChip 250k NspI assay was followed and is summarised in Figure 2.2. A designated pre- and post- PCR clean area was defined for the following steps, to avoid cross-contamination of pre-PCR steps with PCR amplicons. Genomic DNA samples were diluted to a 50ng/µl working dilution with reduced EDTA TE buffer (0.1mM EDTA) (TEKnova).

![Diagram of Affymetrix GeneChip 250k NspI protocol]

Figure 2.2 An overview of the Affymetrix GeneChip 250k NspI protocol

2.3.12.1 NspI digestion

The DNA Engine Tetrad thermal cycler was preheated to 37°C in preparation for genomic DNA digestion. A genomic DNA digestion master mix was prepared on ice with 1x NE Buffer 2 (NEB), 1x BSA (NEB) and 0.5 Units NspI (NEB), made up to a total volume of 14.75µl per sample, with molecular biology grade H₂O (Cambrex). 5µl genomic DNA (50ng/µl) was added to each well of a 96-well plate with a Gilson DistriMan repeater-pipette and 14.75µl of the digestion master mix was added to each sample. The plate was then covered with a plate seal (Applied Biosystems), vortexed
at medium speed for 2 seconds and then centrifuged at 716 rcf for 1 minute. The plate was then placed on the thermal cycler on the following program: 37°C for 120 minutes, 65°C for 20 minutes and then held at 4°C. The digested DNA plate was stored in a -20°C freezer until required.

2.3.12.2 NspI adaptor ligation

The Tetrad thermal cycler was heated to 16°C in preparation for NspI adaptor ligation. A DNA ligation master mix of total volume 5.25μl per sample, was prepared on ice containing 1.5mM Affymetrix NspI adaptor (50mM), 1x T4 DNA ligase buffer (NEB), T4 DNA ligase (400U/μl) (NEB). 5.25μl ligation master mix was aliquoted into each digested DNA sample. The 96-well plate was covered with a new plate seal, vortexed at medium speed for 2 seconds and then centrifuged at 716 rcf for 1 minute. The plate was run in the thermal cycler at 16°C for 180 minutes, 70°C for 20 minutes and then held at 4°C. The adaptor ligated samples were stored in a -20°C freezer until required.

2.3.12.3 Adaptor-ligated fragment PCR and clean-up

The PCR master mix was prepared on ice to a final total volume of 270μl per sample, where each sample was amplified in triplicate in a volume of 90μl. For each 270μl triplicate reaction, 118.5μl molecular biology grade H2O was added with 1x Clontech TITANIUM Taq PCR buffer (Clontech), 1M G-C Melt (Clontech), 350μM each dNTP, 4.5μM PCR Primer 002 (Affymetrix) and 1x Clontech TITANIUM Taq DNA Polymerase (Clontech). The 25μl NspI adaptor-ligated DNA was diluted 1 in 4 by adding 75μl molecular biology grade H2O and 10μl was aliquoted to a 96-well plate in triplicate. 90μl PCR master mix was aliquoted to each adaptor-ligated DNA sample using a Gilson Distriman repeater-pipette. The plate was then covered with an adhesive plate seal, vortexed at medium speed for 2 seconds and then centrifuged at 716 rcf for 1 minute. The plate was then placed on a thermal cycler under the following conditions: 94°C for 3 minutes, 30 cycles of 94°C for 30 seconds, 60°C for 30 seconds and 68°C for 15 seconds, 68°C for 7 minutes and finally held at 4°C.

PCR clean-up was carried out using a Clean-Up Plate (Affymetrix) placed on a QIAvac 96 vacuum manifold (Qiagen) and attached to a vacuum pump (KNF Neuberger). 8μl 0.1M EDTA (Ambion) was added using a Gilson Distriman repeater-pipette to each PCR and the plate covered with a plate seal, vortexed at medium speed for 2 seconds and centrifuged at 716 rcf for 1 minute. The three PCRs for each sample were consolidated into a single well of the Clean-Up Plate and ~600mbar vacuum applied until the wells were completely dry. The PCR products were washed by adding 50μl of molecular biology grade H2O and dried by applying a ~600mbar vacuum. Two additional washes were applied. The Clean-Up Plate was removed from the vacuum manifold and any excess
liquid on the bottom of the plate removed using absorbent paper. 45μl RB Buffer (supplied with Affymetrix Clean-Up Kit) was added to each well and the plate covered with a PCR plate seal. The Clean-Up Plate was agitated at a moderate speed on a plate shaker (the plate was secured by strong adhesive tape to an Eppendorf Thermomixer usually used for mixing 1.5ml Eppendorf tubes), for 10 minutes at room temperature, to recover the PCR product. Recovered PCR products were pipetted to a new 96-well plate and yield quantified using spectrophotometric analysis of a 50-fold diluted sample. Approximately 90μg of purified DNA was transferred to a new 96-well plate and the total volume of each sample brought up to 45μl with RB Buffer (supplied with GeneChip kit).

2.3.12.4 Fragmentation

For the fragmentation step, the thermal cycler was pre-heated to 37°C. 5μl 10x Fragmentation Buffer was added to each sample on ice. The Fragmentation Reagent was diluted to 0.05U/μl in a total volume of 120μl (12μl Fragmentation Buffer and the remainder molecular biology grade H2O up to 120μl), on ice. 5μl of diluted Fragmentation Reagent was added to the sample plate containing Fragmentation Buffer. A plate seal was used to cover the plate and it was vortexed at medium speed for 2 seconds and then centrifuged briefly at 716 rcf at 4°C. The fragmentation plate was placed immediately in the pre-heated thermal cycler and the following cycling conditions were applied: 37°C for 35 minutes, 95°C for 15 minutes and finally the plate held at 4°C. 4μl fragmented PCR product was mixed with 4μl loading buffer and run on 4% agarose gel (made as described previously) at 120V for 45 minutes.

2.3.12.5 Labelling and target hybridisation

A labelling master mix was prepared on ice containing 1x Terminal deoxynucleotidyl Transferase (TdT) Buffer, 0.857mM GeneChip DNA Labelling Reagent, 1.5U/μl TdT, at a total volume of 19.5μl per sample, and vortexed at medium speed for 2 seconds. 19.5μl of the labelling master-mix was aliquoted into the fragmentation plate containing 50.5μl of product from the fragmentation step. The plate was vortexed at medium speed for 2 seconds and then centrifuged at 716 rcf for 1 minute. The plate was then cycled at 37°C for 4 hours, 95°C for 15 minutes and held at 4°C indefinitely. The plate was again centrifuged briefly at 1500 rcf.

All procedures carried out from this point forward were undertaken at the Institute of Child Health Gene-Array Centre, Institute of Child Health (ICH), London, UK, using solutions specified in Affymetrix protocol, prepared by staff at the centre (consult the Affymetrix manual for details on making up these solutions). A hybridisation cocktail master mix was prepared to a total volume of 190μl containing 0.056M 2-(N-Morpholino) ethanesulfonic acid (MES) (1.25M), 5% dimethyl sulfoxide (DMSO) (100%), 2.5x Denhardt's Solution (50x), 5.77mM EDTA (0.5M), 0.115mg/ul
Herring sperm DNA (10mg/ul), 1x Oligo Control Reagent, 11.5ug/ml Human Cot-1 DNA (1mg/ml), 0.0155% Tween-20 (3%) and 2.69M TMACL (5M).

Each 70µl labelled DNA sample was transferred from the plate to a 1.5ml Eppendorf tube and 190µl of hybridization cocktail master mix aliquoted into the same tube. The 260µl mixture was heated at 99°C in a pre-heated heat block (a stationary Eppendorf Thermomixer was used) for 10 minutes to denature the DNA and cooled on ice for a maximum of 10 seconds. The tubes were centrifuged at 400 rcf in a microcentrifuge and placed at 49°C for a minute. The solution was mixed by pipetting to ensure any precipitate was properly dissolved and then 200µl injected into each GeneChip array. Arrays were hybridised in Affymetrix Hybridisation Oven 640 at 49°C for 16 hours at 60 rpm.

2.3.12.6 Washing, staining and scanning

The Fluidics Station 450 was set-up according to manufacturer’s guidelines to run the GeneChip arrays. Buffers Wash A (non-stringent wash buffer), Wash B (stringent was buffer), Stain Buffer and 1x Array Holding Buffer were prepared by staff at the ICH Gene-Array Centre according to the Affymetrix protocol. Each array was filled with 250µl Array Holding Buffer. For each Fluidics module, 495µl Stain Buffer was mixed with 5µl Streptavidin Phycoerythrin (SAPE) (1mg/ml) in an Eppendorf tube and placed in sample holder 1; 495µl Stain Buffer mixed with 5µl biotinylated antibody (0.5mg/ml) was placed in sample holder 2 and 800µl Array Holding Buffer was placed in sample holder 3. The Fluidics wash/stain/scan protocol was run as described in the Affymetrix protocol. Briefly, the protocol involved two post hybridisation washes of the arrays, with non-stringent Wash Buffer A and stringent Wash Buffer B to remove excess labelling and hybridisation reagents. The arrays were then stained with SAPE solution. SAPE contains streptavidin molecules, which bind with high affinity to biotin and therefore will bind biotinylated DNA hybridised to the array, conjugated to the fluorophore phycoerythrin. A post-stain wash with Wash Buffer A removed excess SAPE and a second staining with the biotinylated anti-streptavidin antibody, followed by SAPE solution, was used to amplify the fluorescent signal. A final wash with Wash Buffer A removed any unbound SAPE and the arrays were then filled and held in Array Holding Buffer. Arrays were scanned on a GeneChip Scanner 3000 7G (Affymetrix) and the data stored for analysis using GeneChip DNA Analysis Software (Affymetrix).

2.3.12.7 Array data analysis

GeneChip array intensities were converted into genotype calls and call rates assessed using the Affymetrix GeneChip DNA Analysis Software (GDAS) at various rank score thresholds (see Chapter 6). Concordance/discordance data for replicate or pseudo-replicate chips were generated
using HelixTree (Golden Helix). Pairwise LD comparisons were conducted by importing data from Affymetrix Genotyping Analysis Software (GTYPE) into Haplovie.

### 2.3.13 Quantitative real-time PCR

Allele discrimination reactions for SNP genotyping were performed as follows. Initially 15\(\mu\)l volume reactions were used. A quantitative real-time PCR (qPCR) master-mix was prepared on ice consisting of 8\(\mu\)l QuantiTect Probe PCR Master Mix (2x) (Qiagen), 0.3\(\mu\)l forward primer (50pmol/\(\mu\)l), 0.3\(\mu\)l reverse primer (50pmol/\(\mu\)l), 0.65\(\mu\)l SNP probe allele A (5pmol/\(\mu\)l), 0.65\(\mu\)l SNP probe allele B (5pmol/\(\mu\)l) and 4.1\(\mu\)l dd.H2O. 14\(\mu\)l qPCR master-mix was aliquoted using a Gilson Distriman repeater-pipette into a MicroAmp Optical 96-well Reaction Plate on ice and 1\(\mu\)l genomic DNA (~50ng/\(\mu\)l) added. Work by colleagues within the MRC Prion Unit had suggested that efficient allelic discrimination could be achieved with total reaction volumes of 5\(\mu\)l and therefore in the latter part of my work the following small-volume qPCR master-mix was used, on ice: 2.5\(\mu\)l QuantiTect Probe PCR Master Mix (2x), 0.09\(\mu\)l forward primer (50pmol/\(\mu\)l), 0.09\(\mu\)l reverse primer (50pmol/\(\mu\)l), 0.2\(\mu\)l SNP probe allele A (5pmol/\(\mu\)l), 0.2\(\mu\)l SNP probe allele B (5pmol/\(\mu\)l) and 0.92\(\mu\)l dd.H2O. 4\(\mu\)l the qPCR master-mix was aliquoted into a MicroAmp Optical 96-well Reaction Plate on ice and 1\(\mu\)l genomic DNA (~50ng/\(\mu\)l) added.

Plates were sealed using MicroAmp Optical Adhesive Seals with particular attention paid to sealing around the edges of the plate to prevent evaporation. Plates were then vortexed briefly at medium speed for 2 seconds and pulse centrifuged at 3000 rcf. Plates were run on an ABI 7000 Sequence Detection System (Applied Biosystems), following the manufacturers protocol. A rubber evaporation mat supplied with the machine was used to prevent evaporation. Cycling conditions were: 95°C for 10 minutes, 40 cycles of 94°C for 15 seconds and 60°C for 1 minute.

Copy number assays were conducted in a similar manner to allelic discrimination assays but run in triplicate in 25\(\mu\)l reaction volumes, unless otherwise stated. The major differences in protocol are as follows: A qPCR master-mix was prepared containing 12.5\(\mu\)l QuantiTect Probe PCR Master Mix (2x), 1\(\mu\)l SNP PRNP probe (5pmol/\(\mu\)l), 1\(\mu\)l SNP \(\beta\)Actin probe allele B (5pmol/\(\mu\)l), 0.45\(\mu\)l PRNP forward primer (50pmol/\(\mu\)l), 0.45\(\mu\)l PRNP reverse primer, 0.1\(\mu\)l \(\beta\)Actin forward primer (50pmol/\(\mu\)l), 0.1\(\mu\)l \(\beta\)Actin reverse primer (50pmol/\(\mu\)l), and 7.4\(\mu\)l dd.H2O. The master-mix was aliquoted into a MicroAmp Optical 96-well Reaction Plate or a MicroAmp Fast Optical 96-well Reaction Plate as previously described, and run following manufacturers guidelines on an ABI 7000 Sequence Detection System or ABI 7500 Sequence Detection System, respectively. Data was analysed using ABI Sequence Detection Software v1.3.1 (Applied Biosystems).
2.3.14 Linkage disequilibrium analysis and “tagging” SNP selection

Linkage disequilibrium between SNPs was calculated and visualised using the freely available program, TagIT (Goldstein et al., 2003a; Weale et al., 2003) (http://popgen.biol.ucl.ac.uk/software.html). Methods outlined in the associated User’s Guide were followed and the supplied matrix converter was used to convert SNP genotypes into a binary matrix. Briefly, SNP genotypes were imported into the TagIT matrix converter and modified using the convention 1=major allele and 2=minor allele for each locus in the CEPH population and 0 for missing genotypes. All subsequent allele assignments in additional populations were identical to those in the CEPH to ensure that each allele was represented by the same binary number in all populations. For trio data, TagIT was used to identify deviations from Mendelian inheritance using the “checkM(M)” command, as a safeguard against genotyping errors. Genotypes at each locus were checked for deviations from HWE which, although may often be caused by gene mutation, gene migration, genetic drift, non-random mating or natural selection (Hosking et al., 2004), may also be caused by sampling or genotyping errors. Non-significant $\chi^2$ values ($\chi^2 > 3.841$, $p > 0.05$) indicated consistency with HWE.

Haplotypes were inferred directly for trio data using the “EMtrio” function, with EM algorithm parameters as default: haplotypes returned with an estimated minimum frequency (“limit” parameter) of 0.5% and with a frequency difference of less than $10^{-6}$ for any haplotype between one EM iteration and the next. Haplotypes were cropped to those with an estimated frequency $\geq$1% and the remaining haplotype frequencies were summed to one. Linkage disequilibrium patterns were detected and presented using the TagIT graphical output for $D'$ and $r^2$. The Fisher's Exact test $P$-values for 2x2 tables was invoked using the “EMpairP” command specific for EM estimated haplotypes and visualised in a graphical format to statistically correlate non-random association between SNPs.

![Figure 2.3 The principle of haplotype tagging](https://example.com/f23.png)

SNPs A,B,C and D are tSNPs for a genomic region. As shown none of the tSNPs has a pairwise $r^2$ against SNP X, however SNP X is perfectly predicted by the haplotype $r^2$ criterion or coefficient of determination. (Modified from Goldstein et al., 2003a).
A minimal set of SNPs, or “tagging” SNPs (tSNPs), sufficient to represent a level total genetic diversity in a gene region was identified using the following TagIT routine: Criterion 5 was used to assess the performance of tSNP sets as this criterion was recommended by the authors as the best to use to identify tSNP sets to be used in future association studies (see TagIT User’s Guide). TagIT uses a direct approach to identify a set of tSNPs which maintain high $r^2$ values directly with other (Goldstein et al., 2003a) SNPs or a set of haplotypes that do so, however selecting tSNPs based on their pairwise $r^2$ values with the tagged SNPs (Carlson et al., 2003) has well-documented limitations in that to achieve high $r^2$ values, the frequencies of the variants must be matched. Using haplotype $r'$, or the coefficient of determination resolved by a linear regression of tSNP haplotypes which predict the state of tagged SNPs, can overcome the limitation of pairwise analysis (Figure 2.3). This method also has shortcomings and has reportedly failed to represent SNPs even though the coefficient of determination with tSNPs (or the haplotypes they define) was 1 (Goldstein et al., 2003a). A minimum $r^2$ threshold was chosen at the 0.85 level, allowing any tSNP chosen to represent 85% of the detected and undetected variation.

Using crit=5 and an $r'$ threshold of 0.85, the “best” set size $H$ SNPs was determined so that the set $H+1$ did not greatly increase the performance of the “best” against excluded loci, as described in (Goldstein et al., 2003a). Briefly, the performance of a set of tSNPs against excluded loci can assess how well the set represent as yet undetected SNPs. If $K$ is the total number of SNPs identified in the region of interest, $H$ the tSNPs identified from $K$ and $A$ the set of all common SNPs in the region, the issue is how well $H$ represents $A$. The “excludes” function in TagIT invoked a sub-sampling procedure to determine this: SNPs are sequentially dropped out from the set $K$ (to give $K'$) and the performance of tSNPs derived from $K'$ is tested for ability to predict the state of each dropped SNP in turn. For large data sets, the “excludes” function was too computationally expensive and so alternatively, the “forward2Bx” function was used. tSNP sets were also evaluated for performance against known loci without excluding loci.

**2.3.15 Haplotype inference**

For unrelated individuals, haplotypes were identified in a phase unknown manner using either: (i) Partition Ligation–Expectation Maximisation (PLEM) (Qin et al., 2002) ([www.people.fas.harvard.edu/~junliu/plem/](http://www.people.fas.harvard.edu/~junliu/plem/)) run using parameters top=0, parsize=2 (Japanese data only) and 1 (West African Data only), buffer=n (population size) and number of rounds=20 and input data file prepared according to published documentation. (ii) SNPHAP (Clayton, 2004) via the online GENEPRISE interface at the Archimedes/Wellcome Trust server
PHASE (version 2.1) (Stephens et al., 2003) (www.stat.washington.edu/stephens/software.html) was used with the following command line parameters: -n, -f1, -S (random seed) and default values for number of iterations (100), thinning interval (1) and burn-in* (100). The random seed was required to initiate the random number generator and was changed for each subsequent run. Adhering to the author’s recommendation, each data set was run a minimum of five times to ensure that results were consistent. For obtaining more reliable recombination data the x100 flag was also added, which multiplied the number of iterations, thinning intervals and burn-ins by 100 fold.

Comparison of haplotype inference simulations using PL-EM, PHASE and SNPHAP has been shown to demonstrate a good concordance (Adkins, 2004) although PL-EM performs slightly better under certain circumstances (Goldstein, pers. comm.). Unless otherwise stated, haplotype inference was achieved using PL-EM data and using SNPHAP and PHASE to assign allelic states to loci unspecified by PL-EM.

2.3.16 Human and mouse cytoplasmic dyneins nomenclature, map positions and sequences


Human and mouse chromosomal locations were obtained from the literature and from the MGI and LocusLink databases. Nucleotide and protein sequences (prefix NM_ and NP_ respectively) were NCBI Reference Sequence (RefSeq) and UniProt/Swiss-Prot accession numbers (Wheeler et al., 2003) drawn from the primary sequence database GenBank using the Entrez query interface (www.ncbi.nlm.nih.gov/Entrez/index.html).

2.3.17 Human/mouse homology searches

Homology searches of human and mouse genes for paralogy and orthology were conducted using Position Specific Iterative BLAST (Altschul et al., 1997) at NCBI (www.ncbi.nlm.nih.gov/BLAST).

* A burn-in is the portion of the Markov Chain algorithm (implemented in the PHASE software) that is discarded before a stationary distribution is reached.
The PSI-BLAST program identifies families of related proteins using an iterative BLAST procedure (Altschul et al., 1997). In an initial search, a position specific scoring matrix (PSSM) is constructed from a multiple sequence alignment of the highest scoring hits. Subsequent iterations using the PSSM are performed in a new BLAST query to refine the profile and find additional related sequences. Nucleotide and protein sequences of known genes were used through PSI-BLAST to query the human and mouse non-redundant (nr) sequence databases at GenBank, using default parameters and the BLOSUM-62 substitution matrix, the most effective substitution matrix to identify new members of a protein family (Henikoff et al., 1993). Where protein isoforms were present, the longest sequence was used to search the databases. Homology searches to identify conserved non-coding sequences were conducted using MultiPipmaker (http://pipmaker.bx.psu.edu/cgi-bin/multipipmaker) (Schwartz et al., 2000).

2.3.18 Phylogenetic analyses of the cytoplasmic dynein genes

Homologous sequences were identified by searching the GenBank non-redundant protein database, with the human protein using PSI-BLAST with default parameters and the BLOSUM-62 substitution matrix. Searches of pufferfish sequence Takifugu rubripes were performed using the BLAST (TBLASTN) feature at the Ensembl Fugu Genome Browser (version release 2.0; www.ensembl.org/Fugu_rubripes), searching with human protein sequence against a translated nucleotide database.

Protein sequences were aligned for comparison across their full-lengths using the multiple sequence alignment program CLUSTALW (Thompson et al., 1994) (www.ebi.ac.uk/clustalw/) applying the GONNET250 matrix as default, allowing 250 accepted point mutations per 100 amino acids using scoring tables based on the PAM250 matrix (Gonnet et al., 1992).

Two different phylogenetic methods were used to analyse the dynein gene family alignments. Maximum-likelihood trees were inferred under the Jones, Taylor, and Thornton (JTT) empirical model of amino-acid substitution using PHYML (Version 2.4.3;(Guindon et al., 2003)), as was non-parametric bootstrapping* using 100 resampled alignments for each gene family. Bayesian analyses were performed using MrBayes (Version 3.0) (Ronquist et al., 2003), using the default Bayesian priors on tree topologies and branch lengths. Two different sets of analyses were performed for each gene family, the first allowing the Markov-chain Monte-Carlo algorithm to move between the 11 different amino-acid substitution models available in MrBayes, and another specifying the JTT

* Bootstrapping is a method for testing the reliability of a dataset. It involves the creation of pseudoreplicate datasets by resampling of the original. The frequency with which a given branch is found is recorded as the bootstrap proportion and can be used as a measure of the reliability of branches in the optimal tree.
model. The first analysis allowed the chain to take into account uncertainty in the substitution process. For all analyses performed, the posterior probability of the JTT model was at least 99%, confirming that this model best describes the evolution of the dynein sequences. Only results from the fixed-JTT model analyses were used.

For each analysis, three chains of 1,000,000 generations each were run, sampling parameters every 100 generations and discarding the first 100,000 generations as a burn-in period. Running these multiple independent chains allowed visual confirmation that the chains had reached a stationary state by ensuring that all three chains were moving around a region of similar likelihood. In all cases, the majority rule consensus of the posterior sample of tree topologies from all three Markov chains was used and trees drawn using TREEVIEW (Page, 1996) with posterior clade probabilities and maximum-likelihood bootstrap values shown for each clade on these trees.

2.3.19 PRNP codon 127 statistical analysis

Analysis of \( F_{ST} \) was performed with SPAGeDi. The age of the G127 polymorphism was estimated using the formula in (Risch et al., 1995), as corrected by (Colombo, 2000): for a given marker, the age in generations (\( g \)) is estimated from \( g = \log(\delta) / \log(1 - \theta) \) where, if \( p_D \) is the frequency of a specified allele on \( V \) carrying chromosomes and \( p_D \) the frequency on \( G \) carrying chromosomes, \( \delta = (p_D - p_N) / (1 - p_N) \). 13 micro-satellites were genotyped for this analysis: of these 4 were uninformative and one was excluded because of doubt over the genetic distance between this marker and PRNP. The median results for the remaining 8 markers was used as the point estimate of age, and provide confidence intervals based on 10000 bootstraps of the data (Efron, 1993).

2.3.20 Miscellaneous software

DNA sequence was translated into protein sequence in all six reading frames using the program Translate (http://us.expasy.org/tools/dna.html) at the online proteomics server, ExPASy. Genomic DNA, mRNA and protein sequences were retrieved from NCBI (www.ncbi.nlm.nih.gov/). HapMap data was downloaded from www.hapmap.org. Power calculations were calculated using Genetic Power Calculator at http://pngu.mgh.harvard.edu/~purcell/gpc/.
3 Mutation screening of DYN1H1

3.1 Introduction

The research detailed in this chapter investigates the candidate gene DYN1H1 for motor neuron disease associated variation. The paucity of information available on human DYN1H1 and the recent publication of high quality human genome sequence, dictated that the initial work required for identifying disease associated variants was to elucidate the genomic architecture of DYN1H1, using in silico methods. The localisation of DYN1H1 to chromosome 14q32.32 on human genome contig AL118558 and identification of a 78-exon structure is the first work presented in this chapter, and these data informed much of the later work in this chapter and subsequent chapters.

The 86.6kb locus of DYN1H1 would have been too costly to resequence in its entirety and so, this chapter concludes with a screen of candidate exons 8 and 13, homologous to those causally mutated in the Loa and Cral mouse models of late-onset neurodegeneration, and also intron 13 and exon 14. No variants were identified as significantly associated with motor neuron disease in the regions screened and so this chapter closes with a discussion on the significance of this result and shortcomings of this technique.

3.2 Determining the genomic structure of the DYN1H1 locus in silico

When this study was undertaken in April 2003, little or no published information was available on the genomic architecture of human DYN1H1. With emerging evidence for the importance of this gene as candidate for ALS and other motor neuron diseases, determining the genomic organisation of the gene was paramount. Following the publication of high-quality genomic sequence of human chromosome 14 in February 2003 (Heilig et al., 2003), it became possible to accurately determine the genetic architecture of DYN1H1 by in silico means: essentially, transcribed DYN1H1 sequences collated in sequence databases could be aligned back on to a genomic reference sequence to identify protein-coding sequences (Figure 3.1). The results of this otherwise straightforward experiment were extremely important, primarily for formulating a cost-effective strategy for screening the gene for an association with disease.
Figure 3.1 Scheme representing the reverse mapping of transcribed sequences onto genomic sequence

A. Transcription and intron splicing of protein coding genes. B. In silico identification of exonic sequence using cDNA homology and alignment to a human genome contig.

3.2.1 Information available on the genomic organisation of DYNCIHI in 2003

The chromosomal location of human DYNCIHI had been known since the publication of work from Narayan and colleagues in 1994, who used fluorescence in situ hybridisation (FISH) to locate the gene to the long arm of chromosome 14 (Narayan et al., 1994). Refinement of the chromosomal position to the cytogenetic band 14q32 was published in 2002 (Witherden et al., 2002), but further refinement of gene location, to the resolution of human genome assembly contig and exact sequence, would be needed before elucidating the genomic architecture of DYNCIHI could even be attempted.

Cursory information was available however, on the genomic organisation of DYNCIHI at two public sequence repositories and genome browsers, NCBI and Ensembl, but the data available were neither comprehensive nor conclusive. For example, comparing NCBI RefSeq entries for DYNCIHI coding sequence (Acc: NM_001376), between April 2003 (GI 29788997) and May 2006 (GI 94557306) illustrates the paucity of sequence information available in 2003, as compared to the current state of the database: A 443% increase in coding sequence length is seen between the two DYNCIHI RefSeq entries, increasing the available cDNA sequence from 3210 bp to 14229 bp (full-length).

The Ensembl database and genome browser (release 16; May 2003) did provide some useful information on predicted DYNCIHI genomic organisation. Ensembl is a stable genome database with the capability of annotating known genes and predicting novel genes against the human

*RefSeq or Reference Sequence, is a database that provides a biologically non-redundant collection of DNA, RNA, and protein sequences. Each RefSeq represents a single, naturally occurring molecule from a particular organism is a synthesis of information, not a piece of a primary research data in itself.*

†GI or GenInfo Identifier is a unique sequence identification number for a nucleotide sequence. Every time a change is made to a sequence, a new version of the sequence is produced and a new GI number is assigned. GI numbers are not changed after changes to sequence annotation.*
genome reference sequence (see Hubbard et al., 2002). The Ensembl gene annotation process is an automated one, based upon \textit{ab initio} gene predictions that are either supported or modified by comparison with experimental sequence evidence submitted to the database. The accuracy of these predictions are therefore a function of the completeness of the sequence database. Ensembl predicted exon/intron structure of the gene was available (May 2003; Acc: ENSG00000100839.1, current Acc: ENSG00000197102) and used for analysis.

3.2.2 Fine-scale localisation of \textit{DYNC1H1} to the human genome, chromosome 14 assembly

The initial step in determining the genomic architecture of \textit{DYNC1H1} was to localise the gene at the resolution of DNA sequence, on the Human Genome Project assembly (human genome build 33; April 2003). The genome assembly DNA sequence would provide a reference to compare transcribed sequences against and allow annotation of the genomic architecture of the gene.

At the time, no full-length human cDNA or overlapping cDNA clones had been published for \textit{DYNC1H1} but full-length mouse cDNA (Dynclhl Acc: AY004877) and human protein sequence (DYNC1H1 Acc: NP_001367) were both available to conservatively estimate the size of the human gene. The length of the mouse cDNA and reverse translation of the 4646 amino acid human protein both suggested that the size of the human gene was at least $\sim$14 kb. Assuming the human gene contained introns, the gene size was likely to be greater than 14 kb. The mouse \textit{Dynclhl} cDNA sequence was aligned against the human genome using the sequence alignment tool WU-BLAST v2.0, interfaced through Ensembl genome database and browser. The sequence aligned with an average of 92% homology to human bacterial artificial chromosome (BAC) clone AL118558, contained within the chromosome 14 'golden tile path' BAC RP11-1017G21 and RefSeq contig NT_026437, and the cytogenetic location of the gene was found to be 14q32.32 (Figure 3.2).

\textbf{Figure 3.2} Fine-scale localisation of \textit{DYNC1H1} on chromosome 14 (Builds 33 to 36.1)

\textit{DYNC1H1} was localised to clone AL118558, within the 'golden' tile path BAC RP11-1017G21, and contig NT_026437 at the margin of 14q32.31 and 14q32.32. Adapted from \url{www.ensembl.org/Homo_sapiens}. 

88
3.2.3 Identifying \textit{DYNC1H1} exons and introns

\textit{DYNC1H1} exons were identified by aligning transcribed gene sequences against human genome contig AL118558 and identifying regions of homology, which were then annotated against the genomic reference (Figure 3.1). In the absence of available full-length human cDNA sequence, full-length cDNA sequences from mouse (Acc: AY004877) and rat (Acc: D13896) were used. The mouse and rat transcribed sequences were specifically chosen as (i) both species are closely related to humans reducing the omission of exonic sequences due to species/sequence divergence and (ii) full-length sequences were available. Sequences were aligned against AL118558 using the web-based alignment program ‘BLAST 2 Sequences’. In addition, a number of partial human cDNA sequences (Acc: AB002323, AF234785, LOC196863 and U53530) and expressed sequence tags (ESTs) (Acc: AI418688, AI457261 AI478488, AI497769, AI652200, AI971265, AI991665, AW292121, BE673985, BE701697, BE701780, BE811371, BF062339, BF511560, BF511676, BF515276, BF940018) were aligned against AL118558 (not shown).

![Exon sizes across the DYNC1H1 genomic locus](image.png)

**Figure 3.3 Exon sizes across the DYNC1H1 genomic locus**

78 exons of varying sizes were identified by sequence alignment of transcribed human, mouse and rat dynein 1 heavy chain 1 sequences, against human chromosome 14 genome clone, AL118558. Exons are shown in genomic order and exons 1 and 78 include the 5' and 3' UTR sequences respectively. Median exon size was 165bp (broken red line).

In total, 78 putative exons were observed with an average 90% identity between the three species. The largest and smallest exons were 1077bp and 61bp in length (Figure 3.3) and median exon length was 165bp (mean 188bp), which conforms well with the mean exon size observed in the human genome of <200bp (excluding non-coding UTR sequences) (Sakharkar \textit{et al.}, 2004). All 78 exons had near identical alignments between the three species, which provided a high degree of confidence for the result. However, discordance between alignments were repeatedly observed at...
intron/exon boundaries, which were easily resolved by inspecting boundary sequences for conserved splice junction motifs* (Figure 3.4).

The *in silico* delineated gene architecture was compared to the predicted exon/intron structure available on Ensembl (accessed April 2003; search term “DNCH1”) and all exons corresponded extremely well (data not shown), except exons 1 and 78.

![Figure 3.4 Consensus sequences at DYNC1H1 splice junctions](image)

Relative frequencies are shown at each base pair, extending up to 6 nucleotides 5' and 3' of the splice junction (position 0). Positions with relative frequencies of 0 represent sites where each of the four nucleotides is equally frequent.

3.2.4 Identifying the 5' and 3' UTRs of exons 1 and 78

The discordance of the first and last exons identified above and Ensembl predictions was thought to be due to the accuracy of identifying the size and position of both 5' and 3' untranslated regions (UTRs). UTRs are regions which commonly encode functional elements such as transcription factor binding sites and cDNA stabilisation sites, and are essential for correct gene expression. Mutations in these regions (especially the 5' UTR) can affect gene dosage and therefore these regions may be important in disease process. Consequently, exons 1 and 78 were examined in detail to ensure that the UTRs were identified in their entirety.

Exon 1 contains the common **ATG** translational start site (TSS) at nucleotide 92,523 of the cDNA, which encodes the first amino acid – a methionine – of the DYNC1H1 protein. The methionine TSS is well represented within a conserved ribosomal binding motif, known as a Kozak consensus sequence (Figure 3.5), named after Marilyn Kozak who first described the sequence (Kozak, 1981; Kozak, 1984; Kozak, 1986). The 5' UTR of exon 1 was delimited to position 91,948 by aligning human EST BF511676 against the genomic sequence. The placement of the 5' UTR was supported by the identification of several common eukaryotic RNA polymerase II elements within the

---

* consensus sequences of up to 9 nucleotides in which AG/GT is commonly conserved at exon-intron boundaries and AG/G is commonly conserved at intron-exon boundaries; see (Mount, 1982).
sequence: a conserved transcriptional start site or transcriptional initiator (Inr) element was found at position 91,948 and a well conserved downstream promoter element (DPE) was seen at 91,971 nt. In addition two promoter elements, a TATA box motif and CAAT box motif, were seen at positions approximately 30 nt and 80 nt respectively, upstream of the transcription initiation site. The software UTRscan identified two cis-acting elements further upstream of the Inr, including an internal ribosome entry site (IRES), both of which are involved in cDNA stabilisation and translational control.

The 3' UTR of exon 78 was defined by the alignment of human EST AW249663 against AL118558. The translational stop codon (TAA) was seen at 178,392 nt and polyadenylation signal at 178,605 nt. The final verification of DYNC1H1 genomic architecture was achieved by translating concatenated exonic sequence, using TRANSLATE, into a polypeptide and comparing against protein sequence available on NCBI (NP_001367). Table 3.1 gives the final genomic location of DYNC1H1 exons and introns on AL118558.

Figure 3.5 DYNC1H1 UTR sequences and conserved motifs
Coding sequence shown in red and UTRs highlighted in yellow. 5' UTR extends to the transcription initiator (Inr; +1 position), position 91,948 on AL118558, encompassing the downstream promoter element (DPE) at 91,971 and translational start site (TSS) at 92,523.
<table>
<thead>
<tr>
<th>Exon/intron number</th>
<th>AL118558 pos (bp)</th>
<th>Size (bp)</th>
<th>5' Sequence</th>
<th>3' Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  5' UTR</td>
<td>91,948</td>
<td>92,778</td>
<td>TTCAATGAGAGAAACA</td>
<td>GCTCGACAGCCTCAAG</td>
</tr>
<tr>
<td>2     intron</td>
<td>92,779</td>
<td>103,542</td>
<td>GTCCGGGCGCGGAGGA</td>
<td>TATTATGATTGTAAG</td>
</tr>
<tr>
<td>3     exon</td>
<td>103,543</td>
<td>103,630</td>
<td>AGGACGTTGCGTGAGT</td>
<td>GGTGTTAACCAATAG</td>
</tr>
<tr>
<td>4     exon</td>
<td>103,631</td>
<td>107,149</td>
<td>GTGAGTGACATTACAT</td>
<td>TATTTTTCTTTCATAG</td>
</tr>
<tr>
<td>5     exon</td>
<td>107,150</td>
<td>107,549</td>
<td>CCTGCGCTTCAAGTA</td>
<td>TTTGGCAAGGACAGC</td>
</tr>
<tr>
<td>6     exon</td>
<td>107,549</td>
<td>108,705</td>
<td>GAAACACGTGTTGTTT</td>
<td>GAAACACGTGTTGTTT</td>
</tr>
<tr>
<td>7     exon</td>
<td>108,706</td>
<td>110,194</td>
<td>GATGCTGATACGTTTAA</td>
<td>ATTACTTTTTTCAAAG</td>
</tr>
<tr>
<td>8     exon</td>
<td>110,195</td>
<td>111,212</td>
<td>GTGACCAAACTGAGT</td>
<td>GTGACCAAACTGAGT</td>
</tr>
<tr>
<td>9     exon</td>
<td>111,213</td>
<td>113,517</td>
<td>GAAATTTTCTTTCATAG</td>
<td>GAAATTTTCTTTCATAG</td>
</tr>
<tr>
<td>10    intron</td>
<td>113,518</td>
<td>114,594</td>
<td>GTGACCGGCTTCCGA</td>
<td>AAATTCGAGAAGAAG</td>
</tr>
<tr>
<td>11    intron</td>
<td>114,595</td>
<td>115,283</td>
<td>GTATCGCTCTGAGTAGA</td>
<td>TATTCTTTCTTTCAG</td>
</tr>
<tr>
<td>12    intron</td>
<td>115,284</td>
<td>115,463</td>
<td>GTGAGGGTGTTTGGAGAG</td>
<td>GCTCGAGGAGACAG</td>
</tr>
<tr>
<td>13    intron</td>
<td>115,464</td>
<td>116,533</td>
<td>GTGAAAGACTCTTAA</td>
<td>GTTCTCTTCTTTCAG</td>
</tr>
<tr>
<td>14    intron</td>
<td>116,534</td>
<td>116,683</td>
<td>ATTGAAAGACTATTTTATG</td>
<td>GAGCCAAAGATCAAA</td>
</tr>
<tr>
<td>15    intron</td>
<td>116,684</td>
<td>119,504</td>
<td>GTGAGTGGTTTGGCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>16    intron</td>
<td>119,505</td>
<td>122,012</td>
<td>GAGCACTATGACATTAC</td>
<td>GAGCACTATGACATTAC</td>
</tr>
<tr>
<td>17    intron</td>
<td>121,015</td>
<td>122,155</td>
<td>GTGAAAGACTCTTAA</td>
<td>GTTCTCTTCTTTCAG</td>
</tr>
<tr>
<td>18    intron</td>
<td>122,156</td>
<td>122,503</td>
<td>GAGCACTCTACCTTAA</td>
<td>GTTCTCTTCTTTCAG</td>
</tr>
<tr>
<td>19    intron</td>
<td>122,504</td>
<td>122,680</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>20    intron</td>
<td>122,681</td>
<td>122,816</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>21    intron</td>
<td>122,817</td>
<td>123,011</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>22    intron</td>
<td>123,012</td>
<td>123,131</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>23    intron</td>
<td>123,132</td>
<td>124,865</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>24    intron</td>
<td>124,866</td>
<td>125,105</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>25    intron</td>
<td>125,106</td>
<td>127,819</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>26    intron</td>
<td>127,820</td>
<td>127,975</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>27    intron</td>
<td>127,976</td>
<td>128,116</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>28    intron</td>
<td>128,117</td>
<td>128,230</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>29    intron</td>
<td>128,231</td>
<td>128,784</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>30    intron</td>
<td>128,785</td>
<td>128,985</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>31    intron</td>
<td>128,986</td>
<td>129,185</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>32    intron</td>
<td>129,186</td>
<td>129,365</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>33    intron</td>
<td>129,366</td>
<td>129,512</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>34    intron</td>
<td>129,513</td>
<td>130,367</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>35    intron</td>
<td>130,368</td>
<td>130,534</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>36    intron</td>
<td>130,535</td>
<td>130,622</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>37    intron</td>
<td>130,623</td>
<td>130,796</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>38    intron</td>
<td>130,797</td>
<td>132,348</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>39    intron</td>
<td>132,349</td>
<td>132,514</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>40    intron</td>
<td>132,515</td>
<td>132,592</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>41    intron</td>
<td>132,593</td>
<td>132,781</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>42    intron</td>
<td>132,782</td>
<td>132,872</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>43    intron</td>
<td>132,873</td>
<td>133,067</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>44    intron</td>
<td>133,068</td>
<td>133,718</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>45    intron</td>
<td>133,719</td>
<td>134,001</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>46    intron</td>
<td>134,002</td>
<td>134,838</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>47    intron</td>
<td>134,839</td>
<td>134,939</td>
<td>GTGAGGCTTCTTTCATAG</td>
<td>GCTTTGTGGTGGCATAG</td>
</tr>
<tr>
<td>Exon/intron number</td>
<td>Start (bp)</td>
<td>End (bp)</td>
<td>Size (bp)</td>
<td>Sequence</td>
</tr>
<tr>
<td>--------------------</td>
<td>------------</td>
<td>----------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>29 intron</td>
<td>134,940</td>
<td>136,008</td>
<td>1,069</td>
<td>GTGAGACACTTATATG...CTCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>30 exon</td>
<td>136,009</td>
<td>136,168</td>
<td>160</td>
<td>GCAAATGGGCCCAGATCC...CGAAACTACGACACAGA</td>
</tr>
<tr>
<td>31 intron</td>
<td>136,169</td>
<td>138,106</td>
<td>1,505</td>
<td>GTGAAGACACTTCTTCCTA...TCATCTATTCTCCAG</td>
</tr>
<tr>
<td>32 exon</td>
<td>137,674</td>
<td>138,290</td>
<td>616</td>
<td>CCTTCCGGCCACATTATCG...CTGCCGCGCCAGAACAG</td>
</tr>
<tr>
<td>33 intron</td>
<td>137,918</td>
<td>138,570</td>
<td>1,052</td>
<td>GTAAAGTGACACTTACTT...TTTTCTCTCTCTCTAG</td>
</tr>
<tr>
<td>34 exon</td>
<td>138,571</td>
<td>139,544</td>
<td>974</td>
<td>CTTCCGAGAATGGAGAATG...AGCTTCCGACACAGA</td>
</tr>
<tr>
<td>35 intron</td>
<td>139,706</td>
<td>140,144</td>
<td>200</td>
<td>GTACGTCTCTCTTTTGA...ACAGTCTCCACTGAG</td>
</tr>
<tr>
<td>36 exon</td>
<td>140,145</td>
<td>142,935</td>
<td>2,789</td>
<td>GTGAGAATGACTTATTC...CTCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>37 intron</td>
<td>142,936</td>
<td>143,683</td>
<td>747</td>
<td>GTACGGCGAGCCTATCG...CTCTCTCTCTCTAG</td>
</tr>
<tr>
<td>38 exon</td>
<td>143,164</td>
<td>143,866</td>
<td>702</td>
<td>ATGAGGAGAGCTATCG...CTCTCTCTCTCTAG</td>
</tr>
<tr>
<td>39 intron</td>
<td>143,687</td>
<td>143,917</td>
<td>231</td>
<td>ACTCAAAGAGATGGAC...GAGCGGCTACATCCAG</td>
</tr>
<tr>
<td>40 exon</td>
<td>144,180</td>
<td>144,320</td>
<td>141</td>
<td>CGATATCTGTTTATAT...ATTATCAGATTCTACAG</td>
</tr>
<tr>
<td>41 intron</td>
<td>144,321</td>
<td>144,596</td>
<td>275</td>
<td>GTGAGACGTGACATG...CATCTGCTGCTGATGAG</td>
</tr>
<tr>
<td>42 exon</td>
<td>144,497</td>
<td>144,830</td>
<td>234</td>
<td>GTGCCATACGCTGAGA...TGCTCTGCTGCTGAG</td>
</tr>
<tr>
<td>43 intron</td>
<td>144,831</td>
<td>144,918</td>
<td>88</td>
<td>GTAAAAAGAGCCCGAGA...GCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>44 exon</td>
<td>144,919</td>
<td>145,125</td>
<td>207</td>
<td>GTTGGTGCTCTCAACAG...TCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>45 intron</td>
<td>145,126</td>
<td>145,213</td>
<td>88</td>
<td>GTTTTTTCTATCTCCA...TGTCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>46 exon</td>
<td>145,214</td>
<td>145,335</td>
<td>122</td>
<td>GTGAGGAGACCTGCGAG...GCCCTCTCTCTCCAG</td>
</tr>
<tr>
<td>47 intron</td>
<td>145,336</td>
<td>146,281</td>
<td>946</td>
<td>GTAAACAGACTCGTCTC...GCTGCTGCTGCTGAG</td>
</tr>
<tr>
<td>48 exon</td>
<td>146,282</td>
<td>146,447</td>
<td>166</td>
<td>GTTCTGGCCCTACTGT...TACACTGCTGCTGAG</td>
</tr>
<tr>
<td>49 intron</td>
<td>146,448</td>
<td>147,723</td>
<td>2,275</td>
<td>GTACGAGCTAGTTTCTT...TTTCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>50 exon</td>
<td>147,724</td>
<td>147,887</td>
<td>164</td>
<td>GAGAGCTCTCTCTCTCC...TCTCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>51 intron</td>
<td>147,888</td>
<td>150,581</td>
<td>2,694</td>
<td>GTGCCATAGCAAGTAGA...CTGGCAATGCAAG</td>
</tr>
<tr>
<td>52 exon</td>
<td>150,582</td>
<td>150,711</td>
<td>130</td>
<td>ACTCTGTAAGGATGAG...AAGCTGCTGCTGAG</td>
</tr>
<tr>
<td>53 intron</td>
<td>150,712</td>
<td>154,404</td>
<td>3,693</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>54 exon</td>
<td>154,405</td>
<td>154,538</td>
<td>134</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>55 intron</td>
<td>154,539</td>
<td>155,004</td>
<td>466</td>
<td>GTTGGCCTCTCTCTCC...TACACTGCTGCTGAG</td>
</tr>
<tr>
<td>56 exon</td>
<td>155,005</td>
<td>155,119</td>
<td>115</td>
<td>ATATTCCGCTCAACAC...GCTGCTGCTGCTGAG</td>
</tr>
<tr>
<td>57 intron</td>
<td>155,120</td>
<td>155,213</td>
<td>94</td>
<td>GTGCCGCTGCTGCTGCT...GCTGCTGCTGCTGAG</td>
</tr>
<tr>
<td>58 exon</td>
<td>155,214</td>
<td>155,375</td>
<td>162</td>
<td>GTGCCATAGCAAGTAGA...CTGGCAATGCAAG</td>
</tr>
<tr>
<td>59 intron</td>
<td>155,376</td>
<td>155,444</td>
<td>74</td>
<td>GTACGGCGAGCCTATCG...GCTGCTGCTGCTGAG</td>
</tr>
<tr>
<td>60 exon</td>
<td>155,450</td>
<td>155,664</td>
<td>215</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>61 intron</td>
<td>155,665</td>
<td>155,767</td>
<td>103</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>62 exon</td>
<td>155,767</td>
<td>155,972</td>
<td>205</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>63 intron</td>
<td>155,973</td>
<td>157,369</td>
<td>1,397</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>64 exon</td>
<td>157,370</td>
<td>157,543</td>
<td>174</td>
<td>GCGAATGCTGCTGCT...GAGACAGTGACAG</td>
</tr>
<tr>
<td>65 intron</td>
<td>157,544</td>
<td>157,649</td>
<td>106</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>66 exon</td>
<td>157,650</td>
<td>157,769</td>
<td>120</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>67 intron</td>
<td>157,770</td>
<td>157,992</td>
<td>223</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>68 exon</td>
<td>157,993</td>
<td>158,113</td>
<td>121</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>69 intron</td>
<td>158,114</td>
<td>160,102</td>
<td>1,969</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>70 exon</td>
<td>160,103</td>
<td>160,298</td>
<td>196</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>71 intron</td>
<td>160,299</td>
<td>160,895</td>
<td>597</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>72 exon</td>
<td>160,896</td>
<td>161,013</td>
<td>118</td>
<td>TGGAGCCTAAGAGGAGA...CTGGAGCCTAAGAGGAGA</td>
</tr>
<tr>
<td>73 intron</td>
<td>161,014</td>
<td>161,093</td>
<td>86</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>74 exon</td>
<td>161,100</td>
<td>161,315</td>
<td>216</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>75 intron</td>
<td>161,316</td>
<td>161,806</td>
<td>491</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>76 exon</td>
<td>161,807</td>
<td>162,019</td>
<td>213</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>77 intron</td>
<td>162,020</td>
<td>162,155</td>
<td>136</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
<tr>
<td>78 exon</td>
<td>162,156</td>
<td>162,283</td>
<td>128</td>
<td>GTGAGGAGACCTGCTG...CCCTCTCTCTCTCCAG</td>
</tr>
</tbody>
</table>

93
Table 3.1 DYNC1H1 genomic organisation

78 exons of DYNC1H1 were identified. Exon/intron positions given relative to position on genome contig AL118558. First and last 15 nt of exon/intron sequence is shown. (From Ahmad-Annur et al., 2003).
3.3 Mutation screening of DYNCIHI exons 8, 13 and 14, and intron 13

After determining the genomic architecture of the human DYNCIHI, the next step was to screen the gene for variants associated with disease. Detecting mutations by resequencing the entire 86.6 kb genomic locus was cost-prohibitive and therefore regions of the gene had to be prioritised.

3.3.1 Prioritising regions of DYNCIHI to screen

Screening a single gene in an association study leaves two further variables that can influence the cost and outcome of the study: the number of samples and the number of amplicons screened. With the effect size of a disease associated mutation in DYNCIHI unknown, as many samples as were available were required for screening. Therefore, the number of amplicons had to be minimised to ensure that the study did not exceed that available budget for its completion. This required careful prioritization of regions of the gene that were considered a priori to be of importance to disease pathogenesis. It was not feasible to screen all 78 exons of coding sequence for functionally relevant coding changes and conserved domains such as the motor domain were too large to screen in their entirety. The 5' and 3' UTR were considered for screening but the regions that were screened were those which carried mutations in mouse models of motor neuron disease.

Exons 8 and 13, homologous to those harbouring the mouse Loa and Cral mutations respectively, were to be screened as a priority. Aligning mouse and human protein sequences AAF91078 and Q14204 using BLAST, indicated that the regions corresponding to mouse and human exon 8 and 13 share 99% and 100% homology respectively. The conservation of amino acids between the two species suggested that mutation of either human exon may also yield a deleterious phenotype similar to that seen in the mouse.

3.3.2 Primer design and sequencing of exons 8 and 13

Primers for PCR amplification and sequencing were designed to incorporate coding sequence, splice junctions and flanking sequence to ensure that potential mutations producing aberrant amino acid sequence or splicing would be detected. Due to the small size of exon 13 (177 bp) it was possible to include exon 14 and the intervening intron 13, within a single PCR amplicon and sequencing reaction. Exons were sequenced in both forward and reverse orientation and sequences were checked for mutations by eye.

3.3.3 Patient samples screened for DYNCIHI mutations

DNA from the FALS, HSP and SMA patients screened in this study was collected with informed consent by Professors Karen Morrison and Pamela Shaw and Dr Richard Orrell. All samples collected were of UK residents although for the majority of samples, ethnicity was not known. In
order to eliminate the possibility that a potential mutation in the case samples was actually a polymorphism found at low frequency in the Caucasian population, 100 unrelated CEPH individuals were screened as controls. Prior to the screening of any samples, ethical approval for research was obtained from the Central Office for Research Committee Ethics, London, UK.

3.3.4 Mutation screening of exon 8

Exon 8 coding sequence, splice junctions and flanking sequence were sequenced in a panel of 170 FALS, 31 HSP and 26 SMA patient samples, and 100 CEPH control samples. One synonymous A/G variant was found at the third base pair of codon 836, encoding alanine (A836) (Figure 3.6). The SNP was found in a heterozygous state in two FALS samples and also in a single control sample, with minor allele frequencies (MAF) 0.6% in cases and 0.5% in controls.

![Figure 3.6 Exon 8 A/G synonymous SNP](image)

**Top.** Exon 8 genomic sequence was screened for mutations including flanking intronic sequence (grey) and coding sequence (blue and orange). Position of the A/G SNP is indicated and amino acid translation of the exonic sequence shown above each codon (red).

**Lower.** Electropherograms of the exon 8 wildtype forward sequence (left) and A/G transition (right) seen in both cases and controls. The A/G SNP was a synonymous change at codon 836, encoding an alanine residue.

Genotypes were observed to be in HWE in both cases and controls and a Fisher’s exact test for the independence of SNP frequency in cases compared to controls was not significant (Fisher’s exact; $p=0.1$) when all cases were considered together and also, when FALS samples only were tested against controls (Table 3.2).
Table 3.2 Summary of *DYNC1H1* exon 8 mutation screen, in motor neuron disease patients and controls

The mutation screen was undertaken in a cohort of familial ALS (FALS), familial hereditary spastic paraplegia (HSP) and spinal muscular atrophy (SMA) patients, and CEPH controls. Deviations of genotype proportions from Hardy-Weinberg equilibrium (HWE) and Fisher’s exact test between cases and controls were not significant (ns).

### 3.3.5 Mutation screening of exons 13, 14 and intron 13

Exons 13 and 14, intervening intron 13 and flanking sequence were sequenced in 165 FALS, 32 HSP and 26 SMA patient samples, and 93 CEPH control samples. No variants were found within exon 13 and 14 coding sequences but an A/G polymorphism was seen within intron 13 at base 122,704 of AL118558 (Figure 3.7). The SNP was identified in 8 FALS patients (MAF= 2.4%), 5 SMA samples (9.6%), 6 HSP samples (9.4%) and 9 CEPH controls (4.8%) and all samples bearing the polymorphism were heterozygous. No minor allele homozygotes were seen and all genotypes were in Hardy-Weinberg equilibrium. Fisher’s exact tests for independence of SNP frequency were not significant for the entire case cohort versus controls $p=0.83$ or for cohorts analysed by disease type (ALS $p=0.20$; SMA $p=0.19$; HSP $p=0.22$ (Table 3.3).

![Figure 3.7 Intron 13 A/G SNP](image)

**Figure 3.7 Intron 13 A/G SNP**

**Top.** Exons 13 and 14 coding sequence (blue and orange) and intron 13 sequence (grey) were screened for mutations. Position of the A/G SNP identified is indicated and the amino acid translation of the exonic sequence shown above each codon (red).

**Lower.** Electropherograms of the intron 13 wildtype forward sequence (left) and A/G transition (right) seen in both cases and controls.
Table 3.3 Summary of DYNC1H1 mutation screening of exons 13, 14 and intron 14 in motor neuron disease patients and controls

The mutation screen was undertaken in a cohort of familial ALS (FALS), familial hereditary spastic paraplegia (HSP) and spinal muscular atrophy (SMA) patients, and CEPH controls. Deviations of genotype proportions from Hardy-Weinberg equilibrium (HWE) and Fisher's exact test between cases and controls were not significant (ns).

3.4 Discussion

The results detailed in this chapter describe work to assess the significance of the candidate gene DYNC1H1 as a causal or susceptibility factor for ALS and/or other motor neuron diseases. At the time that this research began in April 2003, little information was available on the genomic organisation of DYNC1H1. The nature of research carried out in the cytoplasmic dynein community has historically focused on examining proteins and expressed sequences of non-human species and so, there existed a paucity of sequence information on the human cytoplasmic dyneins. This consequently impeded initiatives by genome databases such as Ensembl to identify the genomic architecture of predicted and known genes by automated methods. Elucidating the genomic structure of a candidate gene is a problem seldom faced by researchers today, with high quality in silico and sequence data readily available in public databases, but establishing the size and structure of a candidate gene to be screened for an association with disease is paramount.

Through comparative sequence techniques (comparing against mouse and rat sequences) DYNC1H1 was localised to chromosome 14q32.32 and to the human genome assembly on contig AL118558. The gene was found to comprise 78 exons and span 86.6kb. The use of homology to related species allowed the complete gene structure to be elucidated where only a partial structure was seen at on browsers such as Ensembl. A similar technique called "gene builds" has since been implemented by Ensembl as an automated method for identifying and describing genes from 2004 (Curwen et al., 2004). In addition, the availability of additional eukaryotic genome sequences and sequence comparison tools, such as MultiPipmaker (Schwartz et al., 2003), have facilitated the identification of functionally conserved sequences, including putative coding sequences (Morgenstern et al., 2002), and confirm the genomic structure predicted in this chapter was correct.
The final results in this chapter detail the screening of exons 8, 13 and 14 and intron 14 of \textit{DYN1HI} for mutations associated with various forms of motor neuron disease. Two variants were identified; one in exon 8 and one within intron 13. The exon 8 variant was an A/G transition at low frequency (~0.5%) in both cases and controls and constituted the third base position of an alanine codon (A837). Due to the redundancy of the human genetic code, the SNP does not alter the amino acid composition of the protein - GCA and GCG both code for an alanine - and so a deleterious phenotype in individuals possessing the SNP is unlikely. There was no significant difference in allele frequency of the SNP between cases and controls, and therefore it is likely that this SNP is simply a neutral polymorphism.

The intron 13 SNP was identified in all three disease types and controls, at varying frequencies, which is likely to reflect an ascertainment bias due to sample size. The SNP was an A/G transition which bears no obvious functional relevance as it does not occur at the splice junctions. There was no significant difference in allele frequencies between cases and controls, when cases were compared by disease type or when combined, and genotypes were seen in HWE, suggesting that this SNP is not associated with disease.

3.4.1 Updated SNP information

Since this screen was carried out, information on SNP content of \textit{DYN1HI} and SNP frequencies has become available at dbSNP. The SNPs discovered in this screen have been identified by other investigators and have been submitted to dbSNP in 2004, with SNP IDs: rs17512054 and rs4900529 for exon 8 and intron 13 respectively. Both SNPs have been validated and frequency information is available. No other SNPs have been discovered in the regions screened in this study except for rs17540908, which is 9bp downstream from rs4900529. This SNP, with a MAF of 0.6%, has not yet been validated and so may be an artefact.

Frequency data available for rs17512054 gives a MAF of 1.4% which is in contrast to the 0.6% and 0.5% we identified in cases and controls. This discrepancy is almost certainly due to differences in populations used as the dbSNP data are generated using a multi-ethnic polymorphism discovery resource panel of 90 samples (PDR90) and a Caucasian population was used in this screen (Collins et al., 1998). Population differences accounting for frequency discrepancies seen for SNP rs4900529, for which a MAF of 4.8% was found for Caucasian controls in this screen compared to a MAF of 14.7% on dbSNP using PDR90. As rs4900529 was genotyped in the HapMap project, population specific frequencies are available and it is clear that in European samples the MAF is 5.8% and in Asian and Sub-Saharan African samples the MAF varies from 27.5% to 19.3%.
Therefore the minor allele frequencies experimentally derived for both SNPs in this screen were similar to the frequencies known for these SNPs in Caucasian populations.

3.4.2 Implications and explanations of a negative screen

The lack of association between the SNPs in exon 8 and intron 13 and FALS, HSP and SMA suggests that these specific SNPs may not be relevant to disease pathogenesis but does not preclude the significance of mutations in \textit{DYNCIHI} to ALS pathogenesis and other motor neuron diseases. There are several caveats to this screen: (i) although the FALS cohort was of a reasonable size, the HSP and SMA cohorts were small, with 32 and 26 individuals respectively. It is possible therefore, that these variants may account for a rare minority of familial SMA and HSP which would not be detected with such a limited sample size. (ii) all three diseases are known to demonstrate genetic heterogeneity and although \textit{SOD1}, \textit{SMN} and \textit{Spastin} mutations had been ruled out in the FALS, SMA and HSP cases respectively, several additional loci which are known to be mutated in familial forms of these disorders, had not been screened. (iii) the study only screened a discrete region of \textit{DYNCIHI} for association with disease. Without an understanding of the extent of LD surrounding these SNPs, this screen can only be considered a direct association study and therefore, regions of the gene not investigated may still harbour disease associated mutations. To investigate the gene in its entirety, the LD pattern spanning the gene was elucidated and an association study conducted.
4 Cytoplasmic dynein 1 heavy chain 1 association study

4.1 Introduction

Despite screening DYNC1H1 candidate exons 8 and 13 in a number of familial motor neuron disease cases and controls, no variants associated with disease were found. This chapter extends those previous analyses by using a linkage disequilibrium based association study to investigate the entire DYNC1H1 genomic locus in an attempt to identify an association with sporadic ALS in a northern European-derived population. 16 SNPs were examined, of which two (rs2251644 and rs941793) were found to be sufficient to tag the majority of haplotypic variation ($r^2 \geq 0.85$). These tSNPs were tested in 261 North American sporadic ALS patients and 225 matched controls but no association with the disease was found. In addition, the genetic diversity of DYNC1H1 was examined in Japanese and Cameroonian populations to establish the evolutionary history for this gene. This chapter finishes with a brief discussion of the caveats associated with this work and how association studies have changed since this study was initiated.

4.1.1 Mutation screening of large genes can be problematic

As previously described, the genomic locus of DYNC1H1 spans over 86.6 kb and contains 78 exons. Screening genes of this size for disease associated mutations can be prohibitively expensive and the cost of such a strategy is further multiplied for complex diseases, where mutations are likely to have a small effect size and require, therefore, a large sample size to have an appreciable power to detect a disease-associated mutation. One method for overcoming the limitations associated with mutation screening of large genes is to screen a limited part of the gene instead of screening for all possible mutants, which leads to the question: which region of the gene should be screened? As previously discussed, prior hypotheses based on experimental or bioinformatic data can narrow down a candidate region for screening; for example, the screening of exons 8 and 13 of DYNC1H1 was based upon known mutations in of the mouse homolog of the gene. No mutations were discovered in this gene and therefore to investigate the entire DYNC1H1 genomic locus for an association with sporadic ALS, a linkage disequilibrium-based association study was conducted.

4.2 DYNC1H1 association study design

4.2.1 Evolving resources and methodologies in LD-based association studies influence study design

The DYNC1H1 association study was designed based on the understanding of case-control association studies and LD mapping available at the time the study was undertaken in 2002. The
paucity of genetic variation data available for \textit{DYNC1HI} in 2002 as compared to today greatly influenced the study design, with much of the preliminary work directed towards describing variation within the gene. Contemporary LD-based association studies are able to exploit large amounts of publicly available data on SNP position and frequency, LD patterns, genotypes in various worldwide populations and even tagging performance – almost none of which was available when this study was initiated. Today, the first stop for the majority of researchers interested in the genetics of complex neurodegenerative diseases are data resources such as The International HapMap Project which, although initiated in 2002, has come to fruition too late to influence the design of this study. The majority of work presented in this chapter was designed and completed before changes in this field had become common place, so many of the techniques and analyses are redundant for today’s investigators. In later sections of this chapter, the differences in study design and their effect on success are discussed.

4.2.2 Three phased study design

The \textit{DYNC1HI} study design encompassed three broad phases; discovery, tagging and testing. The goal of the SNP discovery phase was to construct both SNP and LD maps spanning \textit{DYNC1HI} at a density of approximately 1 SNP every 10kb. This SNP density was, at the time, shown to be highly effective for identifying regions of elevated LD whilst representing the true pattern of genomic variation in a region (Dawson \textit{et al.}, 2002; Gabriel \textit{et al.}, 2002; Goldstein \textit{et al.}, 2003a; Reich \textit{et al.}, 2001a). The discovery phase was necessary due to a general deficit of available marker information assigned to SNPs present in the public databases: SNPs lacked (i) validation data, essential for authenticating SNPs, (ii) data on their physical locations and (iii) information on their allele frequencies. The tagging phase was used to identify tSNPs which could efficiently represent variation across the gene. These tSNPs were then taken forward to the test phase, where they were assayed in case and control samples and analysed for association with disease status.

4.2.3 Estimating LD to assess tagging approach feasibility

The absence of \textit{a priori} information on the pattern of LD across \textit{DYNC1HI} at 14q32.32, made it initially difficult to assess how powerful a tagging SNP strategy would be and the level of economy this approach could afford. However, as there is a general relationship of inverse proportionality between LD and recombination distance (Huttley \textit{et al.}, 1999), putative LD across \textit{DYNC1HI} was estimated by considering recombination rates between markers flanking the gene.
The recombination rate was estimated by comparison of genetic and physical maps for the region. Sequence tagged sites (STSs)* flanking *DYNC1HI* were identified using the Genome Database (http://www.gdb.org/gdb; accessed November 2002) and two STS loci, *D14S1051* and *D14S293*, were found approximately 0.2Mb and 0.9Mb upstream and downstream, respectively. Both STS loci had physical map coordinates on the NCBI sequence contig Hs14_10185 and genetic map coordinates on the Généthon chromosome 14 map, which allowed a ratio (i.e. the recombination rate) to be calculated (Table 4.1).

<table>
<thead>
<tr>
<th>Marker</th>
<th>Map</th>
<th>Coordinate</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>D14S1051</td>
<td>NCBI contig Hs14_10185</td>
<td>5.2</td>
<td>Mb</td>
</tr>
<tr>
<td></td>
<td>Généthon chromosome 14</td>
<td>137</td>
<td>Kosambi cM</td>
</tr>
<tr>
<td>D14S293</td>
<td>NCBI contig Hs14_10185</td>
<td>3.9</td>
<td>Mb</td>
</tr>
<tr>
<td></td>
<td>Généthon chromosome 14</td>
<td>133</td>
<td>Kosambi cM</td>
</tr>
</tbody>
</table>

**Table 4.1 Marker information for sequence tagged sites flanking *DYNC1HI***

Physical map coordinates are given in Megabases (Mb) and genetic map coordinates given in Kosambi centiMorgans (cM).

The recombination rate was 3.08 cM/Mb, approximately twice the sex averaged recombination rate for chromosome 14 (1.36 cM/Mb) and almost 3 times greater than genome-wide sex averaged recombination rate (1.13 cM/Mb) (Kong et al., 2002). The high recombination rate did not preclude however, the existence of elevated LD at shorter distances than the 1.3Mb physical distance between the two STS markers – i.e. fine-scale LD (Reich et al., 2001a). Indeed, by considering the STS marker information available in 2006, this estimate is reduced to 2.15 cM/Mb (using *D14S1051* and *D14S272* genetic positions from the Marshfield map; *D14S272* is 0.17 Mb closer to *DYNC1HI* than *D14S293*). Without typing markers at a greater proximity to the gene and across the gene, it was difficult to accurately estimate the extent of LD across the region. Therefore, a SNP map of the region spanning the gene needed to be constructed.

### 4.2.4 Study power and estimating required sample size

Despite the absence of accurate LD estimates across the gene, power and sample size calculations were conducted using the genetic power calculator of Purcell, Cherny and Sham (http://pngu.mgh.harvard.edu/~purcell/gpc) (Purcell et al., 2003). Calculations were performed with helpful advice from Professor David Goldstein and using the conventional power threshold ≥80%.

Without prior knowledge of LD across the gene, *D'* values between 0.8 and 1.0 were used which

---

*STSs serve as landmarks on the physical map of the human genome. Each is a short DNA segment that occurs only once in the human genome and whose exact location and order of bases are known. Because each is unique, STS markers are helpful for chromosome placement of mapping and sequencing data from many different laboratories.
related to samples sizes varying from 982 to 647 respectively, to achieve 80% power of detection (Figure 4.1).

![Graph showing sample size and power for varying D' values](image.png)

**Figure 4.1** DYNC1H1 association study power and related sample size for varying D' values

Expected power was calculated by considering a causal variant under a multiplicative risk model, with genotypic relative risks of 2.0 for homozygotes and 1.41 for heterozygotes, and an allele frequency of 0.1. Alpha (type I error rate) was set to 0.05 and marker allele frequency was set at 0.1. Dashed grey lines indicate sample sizes with 80% power of detection.

### 4.3 Phase I - SNP discovery

#### 4.3.1 In silico SNP ascertainment

Several public SNP databases were used to identify the physical locations of SNPs across DYNC1H1, to construct a SNP map. The primary database interrogated was dbSNP ([www.ncbi.nlm.nih.gov/projects/SNP/](http://www.ncbi.nlm.nih.gov/projects/SNP/)), a central repository for human and mouse SNP data held at the National Cancer for Biotechnology Information (NCBI) as, dbSNP (Smigielski *et al.*, 2000). Although by September 2000 dbSNP contained SNPs at a frequency of one SNP every 2.03kb, the number of SNPs found in and around DYNC1H1 were few and with little or no information on heterozygosity or validation status. A second database, Human Genome Variation Database (HGVBASE, formerly HGBASE; [http://hgvbase.cgb.ki.se](http://hgvbase.cgb.ki.se)), independent of dbSNP was also used (Brookes *et al.*, 2000).

Compounding the paucity of SNP information was the absence of validation, which left the possibility that potential SNPs could be errors due to sequencing artefacts, the presence of paralogous sequences elsewhere in the genome or due to inaccurate laboratory assays. Initially to overcome the problem of verification, overlapping detection of SNPs by multiple independent databases was used to validate genuine polymorphisms. However this method proved impractical as only a small percentage of SNPs were seen in both databases and only 40% of final SNPs were seen in 3 databases. SNPs therefore had to be validated experimentally.
Data on SNP MAF was also unavailable. The minimum acceptable SNP MAF was selected to be 10%. Although commonly set at 5% (for later statistical chi-squared analysis to be valid), the MAF was set at a 10% threshold for the following reasons: (1) database SNPs with minor allele frequency annotation ≥10% are more likely to be genuine SNPs than artefacts, (2) to ensure that the study would have sufficient power to detect SNPs, even with a small discovery sample (3) to ensure that haplotypes later defined would be common enough to represent an ancestral state rather than more recent “singletons”, (4) imposing a frequency threshold is necessary to avoid spuriously high estimates of LD when allele frequencies are close to zero (Mateu et al., 2001). The impact of using a MAF >5% is discussed later in this chapter.

4.3.2 SNP validation and discovery

SNPs identified from the databases were validated by genotyping in a discovery sample of 32 chromosomes from 16 unrelated CEPH individuals, obtained from Dr. Howard M. Cann at the Foundation Jean Dausset (Dausset et al., 1990). Individuals from the CEPH reference families were chosen as their northern European ancestry is similar to that of the North American case and control samples which were to be tested for association (discussed later), serving as a proxy for genetic variation in the limited North American samples. Initially, the process was to be expedited by genotyping SNPs which altered endonuclease restriction sites; however these were found to be vanishingly rare, comprising less than 4% of the total number investigated. SNPs were therefore, genotyped by automated dideoxy sequencing in both DNA strand directions. Primers were designed as described in Chapter 2.3.4 to maximise amplicon size, exploiting maximum sequence length to identify novel SNPs.

4.3.3 Minor allele frequencies

In addition to validating each SNP, resequencing 32 chromosomes from unrelated individuals allowed an assessment of MAF for each SNP. Invariant loci, or where SNP MAF<10%, were disregarded and instead a neighbouring SNP was chosen from the databases and validated until at least one SNP approximately every 10 kb was identified. If no neighbouring SNPs were listed within the databases, SNPs were identified by resequencing sequential 500kb amplicons either side of the original SNP until a polymorphic site with MAF ≥10% was found.

In total, 60 SNP loci were examined through resequencing approximately 16.7kb DNA: 36 SNPs were found to be invariant when assayed in 32 chromosomes and a further 8 SNPs were seen to have MAF below the threshold value (Table 4.2). Additionally, 6 new SNPs were identified
(rs3742426*, rs941792, rs10135238, bp183293†, rs11849604 and rs1190618) not previously reported in any database. Anecdotally, fewer SNPs, and at lower frequencies, were seen in the 5' region of the gene (between SNPs 1 and 44). Approximately 13kb of resequencing (75% of total) was undertaken in the region between SNPs 1 to 44, with only one novel SNP found. In contrast, 3.7kb resequencing between SNPs 45 and 60 yielded 5 novel SNPs, suggesting possible reduced genetic diversity in the 5' region of the gene compared to the 3' region.

In total, 16 SNPs were identified with MAF ≥10%, which were all unremarkable in their locations: 8 SNPs were located within intronic sequence but not at splice junctions and 7 SNPs were outside the transcribed region. A single SNP (rs13749) was identified within coding sequence as a synonymous change (Figure 4.2), located at the fully redundant third position of the 4360 threonine amino acid codon (ACC), and therefore did not alter the amino (T4360T). The SNP density, averaged across the gene, was approximately one SNP every 6.9kb, with the largest inter-SNP gap of 21kb between rs2180510 and rs941793.

![SNP rs13749 electropherograms and the redundancy of the threonine codon](image)

**Figure 4.2 SNP rs13749 electropherograms and the redundancy of the threonine codon**

*Right.* A single synonymous T/C coding SNP rs13749 was identified. **Left.** The minor SNP allele was located at the fully redundant third position of the ACT threonine codon, so no corresponding coding change was seen.

---

* rs12895291 was merged into rs3742426 on dbSNP in May 2006
† bp183293 is not present in any database. The SNP ID is given by its position on the AL118558
<table>
<thead>
<tr>
<th>Number</th>
<th>SNP ID*</th>
<th>Position (bp)*</th>
<th>Alleles</th>
<th>Amplicon Size (bp)</th>
<th>Genotype*</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs2180511</td>
<td>27670</td>
<td>A/T</td>
<td>535</td>
<td>AA</td>
<td>6.0</td>
</tr>
<tr>
<td>2</td>
<td>rs1678034</td>
<td>30289</td>
<td>A/C</td>
<td>566</td>
<td>AB</td>
<td>6.3</td>
</tr>
<tr>
<td>3</td>
<td>rs2180513</td>
<td>61497</td>
<td>C/T</td>
<td>622</td>
<td>BB</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>rs1956297</td>
<td>74179</td>
<td>C/T</td>
<td>453</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>rs2474677</td>
<td>75613</td>
<td>C/T</td>
<td>360</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>rs2474678</td>
<td>76117</td>
<td>A/G</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>rs2448241</td>
<td>76261</td>
<td>C/T</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>rs2093025</td>
<td>76319</td>
<td>C/T</td>
<td>500</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>rs2474679</td>
<td>76323</td>
<td>C/T</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>rs1044836</td>
<td>76846</td>
<td>A/C</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>rs2281539</td>
<td>76858</td>
<td>C/T</td>
<td>510</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>rs2273434</td>
<td>76920</td>
<td>C/G</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>rs2273435</td>
<td>77049</td>
<td>A/G</td>
<td>151</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>rs2448240</td>
<td>77419</td>
<td>C/G</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>rs2448239</td>
<td>79006</td>
<td>A/G</td>
<td>520</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>rs1741151</td>
<td>80181</td>
<td>C/T</td>
<td>320</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>rs2403015</td>
<td>89054</td>
<td>A/G</td>
<td>591</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>rs1622886</td>
<td>89874</td>
<td>A/G</td>
<td>420</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>rs1741155</td>
<td>89989</td>
<td>A/G</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>rs2749911</td>
<td>90216</td>
<td>C/T</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>rs2720193</td>
<td>90225</td>
<td>A/G</td>
<td>460</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>rs2720194</td>
<td>90295</td>
<td>A/G</td>
<td>124</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>rs2448242</td>
<td>91338</td>
<td>A/G</td>
<td>440</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>rs2273438</td>
<td>107101</td>
<td>C/T</td>
<td>124</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>rs2273439</td>
<td>107195</td>
<td>C/T</td>
<td>380</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>rs2720218</td>
<td>109449</td>
<td>A/C</td>
<td>190</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>rs2720219</td>
<td>114804</td>
<td>A/G</td>
<td>200</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>rs2749894</td>
<td>119522</td>
<td>C/T</td>
<td>520</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>rs2720206</td>
<td>119787</td>
<td>C/T</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>rs2720213</td>
<td>126498</td>
<td>C/G</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>rs2251644</td>
<td>129341</td>
<td>A/G</td>
<td>280</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>32</td>
<td>rs2748986</td>
<td>132256</td>
<td>A/G</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>rs2748987</td>
<td>138855</td>
<td>A/G</td>
<td>380</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>34</td>
<td>rs2720197</td>
<td>139455</td>
<td>A/G</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>35</td>
<td>rs3742428*</td>
<td>147536</td>
<td>A/G</td>
<td>600</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>36</td>
<td>rs2180510</td>
<td>147694</td>
<td>A/G</td>
<td>124</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>37</td>
<td>rs2749899</td>
<td>150639</td>
<td>A/G</td>
<td>453</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>38</td>
<td>rs2749900</td>
<td>151540</td>
<td>C/G</td>
<td>365</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>39</td>
<td>rs2720198</td>
<td>151863</td>
<td>A/C</td>
<td>440</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>rs2403016</td>
<td>153341</td>
<td>C/T</td>
<td>359</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>41</td>
<td>rs2285445</td>
<td>157537</td>
<td>C/G</td>
<td>367</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>42</td>
<td>rs754598</td>
<td>160726</td>
<td>A/G</td>
<td>498</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>rs941632</td>
<td>163482</td>
<td>A/G</td>
<td>275</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>44</td>
<td>rs2093024</td>
<td>167571</td>
<td>A/G</td>
<td>530</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>45</td>
<td>rs941793</td>
<td>168955</td>
<td>C/T</td>
<td>260</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>46</td>
<td>rs1190605</td>
<td>171447</td>
<td>C/G</td>
<td>420</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>47</td>
<td>rs1043455</td>
<td>172220</td>
<td>C/T</td>
<td>327</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>48</td>
<td>rs13749</td>
<td>175721</td>
<td>C/T</td>
<td>430</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>49</td>
<td>rs2273656</td>
<td>175929</td>
<td>A/C</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>rs941792*</td>
<td>176207</td>
<td>C/T</td>
<td>124</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>51</td>
<td>rs1127284</td>
<td>176475</td>
<td>C/T</td>
<td>380</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>52</td>
<td>rs1604903</td>
<td>178509</td>
<td>A/G</td>
<td>160</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>53</td>
<td>rs10135238</td>
<td>180364</td>
<td>C/T</td>
<td>530</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>54</td>
<td>rs1190610</td>
<td>180468</td>
<td>A/G</td>
<td>124</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>55</td>
<td>rs1190613</td>
<td>183028</td>
<td>C/T</td>
<td>440</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>56</td>
<td>bp183293*²</td>
<td>183293</td>
<td>A/T</td>
<td>124</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>57</td>
<td>rs1190614</td>
<td>185042</td>
<td>A/T</td>
<td>570</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>58</td>
<td>rs11849604</td>
<td>185073</td>
<td>A/T</td>
<td>124</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>59</td>
<td>rs1203482</td>
<td>193407</td>
<td>C/T</td>
<td>360</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>60</td>
<td>rs1990618*</td>
<td>193460</td>
<td>C/G</td>
<td>124</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.2 SNPs identified and validated in a discovery panel of 16 CEPH individuals

SNPs were initially identified from databases and genotyped by bi-directional resequencing in 16 unrelated CEPH individuals to validate their presence and determine minor allele frequency. Relative location of DYN1CH1 represented by the yellow box; SNPs within the yellow box are intergenic. *novel SNPs identified in 2004, have since been added to dbSNP. † rs12895291 was merged into rs3742426 on dbSNP in May 2006. ‡ bp183293 not present in any SNP database. SNP ID given by position on AL118558
4.4 Genotyping validated SNPs

The 16 validated SNPs were genotyped in 32 mother, father and child CEPH trios*, to identify haplotypes and construct a detailed LD map across the region (Figure 4.3). Comprehensive genotype information was obtained by bidirectional dideoxy sequencing for all individuals using the same primers/conditions applied to the discovery panel (full genotyping results are given in Appendix 1), however, genotypes for rs3742426 and rs2180510 continually failed for the child in trio 6 (id:34603). These SNPs displayed weak PCR products and their sequencing reactions continually failed. Both of these SNPs reside in the same PCR amplicon and therefore, the genotyping failure may have been due to a polymorphism in the primer binding sequence which can reduce the affinity of a primer for the DNA template. However, as both parents were homozygous for the major allele at each locus, the child’s genotype could be easily inferred and therefore, genotyping this SNP was not pursued (although this did not rule out a potential deletion in this region).

![Figure 4.3 DYNC1H1 SNP map](image)

A schematic of the DYNC1H1 genomic locus shown positioned 5' to 3', with dbSNP accession numbers and minor allele frequencies (in parentheses). Black boxes represent first and last exons.

There was good concordance between allele frequencies of the 32 chromosome validation panel and the full CEPH panel of 64 unrelated parents (Table 4.3), especially where MAF ≥10% in the full panel. However, several SNPs were seen to decrease their MAF from 12.5% in the discovery panel to 7% in the LD panel. These SNPs were synteny, occurring in the 5' region of the gene whilst SNPs in the 3' region of the gene maintained their MAF of approximately 18%. This may have been due to a sampling anomaly whilst randomly selecting CEPH individuals for the discovery panel which may not have been representative of the full CEPH panel.

---

* Trios were used to provide information on SNP phase and aid accurate haplotype inference
<table>
<thead>
<tr>
<th>Number</th>
<th>SNP ID</th>
<th>Position (bp)</th>
<th>Alleles</th>
<th>n=32</th>
<th>n=128</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1741155</td>
<td>89892</td>
<td>A/q</td>
<td>15.6</td>
<td>13.2</td>
</tr>
<tr>
<td>2</td>
<td>2720193</td>
<td>90225</td>
<td>A/q</td>
<td>12.5</td>
<td>7.0</td>
</tr>
<tr>
<td>3</td>
<td>2720194</td>
<td>90295</td>
<td>A/q</td>
<td>12.5</td>
<td>7.0</td>
</tr>
<tr>
<td>4</td>
<td>2273438</td>
<td>107101</td>
<td>T/c</td>
<td>12.5</td>
<td>7.0</td>
</tr>
<tr>
<td>5</td>
<td>2749890</td>
<td>114824</td>
<td>A/q</td>
<td>12.5</td>
<td>7.0</td>
</tr>
<tr>
<td>6</td>
<td>2251644</td>
<td>129341</td>
<td>A/q</td>
<td>12.5</td>
<td>7.0</td>
</tr>
<tr>
<td>7</td>
<td>3742426</td>
<td>147536</td>
<td>G/a</td>
<td>12.5</td>
<td>7.0</td>
</tr>
<tr>
<td>8</td>
<td>2180510</td>
<td>147694</td>
<td>G/a</td>
<td>12.5</td>
<td>7.0</td>
</tr>
<tr>
<td>9</td>
<td>941793</td>
<td>168955</td>
<td>A/q</td>
<td>17.6</td>
<td>17.2</td>
</tr>
<tr>
<td>10</td>
<td>13749</td>
<td>175721</td>
<td>T/c</td>
<td>12.5</td>
<td>14.8</td>
</tr>
<tr>
<td>11</td>
<td>941792</td>
<td>176207</td>
<td>G/a</td>
<td>15.6</td>
<td>16.4</td>
</tr>
<tr>
<td>12</td>
<td>1004903</td>
<td>176509</td>
<td>G/a</td>
<td>15.6</td>
<td>14</td>
</tr>
<tr>
<td>13</td>
<td>10135238</td>
<td>180364</td>
<td>T/c</td>
<td>17.6</td>
<td>18</td>
</tr>
<tr>
<td>14</td>
<td>1190610</td>
<td>180468</td>
<td>A/q</td>
<td>15.6</td>
<td>18</td>
</tr>
<tr>
<td>15</td>
<td>1190613</td>
<td>183028</td>
<td>A/q</td>
<td>17.6</td>
<td>17.2</td>
</tr>
<tr>
<td>16</td>
<td>1190618</td>
<td>193460</td>
<td>C/q</td>
<td>17.6</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 4.3 SNP frequency comparison between 32 and 128 chromosome discovery panels

SNP ID refers to dbSNP identification numbers with rs- prefix. Alleles are shown with the common allele in uppercase. SNP with frequencies in red changed significantly between panels.

4.4.1 Quality control - assessing genotyping accuracy

An estimate of genotyping accuracy was empirically determined, to identify assays bias and to prevent errors affecting later association tests. This work was carried out with the help of Dr. Azlina Ahmad-Annuar. SNPs rs2251644 and rs1004903 were chosen at random and resequenced in 40 individuals, with the operator blind to all sample identifiers. A second operator coded the samples and when decoded, the genotyping error rate was ~1%. This error rate demonstrated that a high fidelity of the genotyping assay was achieved.

4.4.2 Quality control – Mendelian inheritance and Hardy-Weinberg equilibrium

Genotype data was prepared for analysis using the tagging SNP selection software TagIT (Goldstein et al., 2003b) as described in Chapter 2.3.14. As an additional genotyping quality control measure and to ensure that no errors had been introduced during manual input of data, the “check matrix” command (‘checkM(M)’) in TagIT was used to identify inconsistencies in Mendelian inheritance within trios. A single case of non-Mendelian transmission was seen in which a heterozygous parent (AC genotype) and homozygous major parent (AA genotype) resulted in a homozygous minor child (CC genotype). Upon investigation of the original sequencing data, this
was found to be due to a data inputting error and was rectified. The data were also tested for deviations from Hardy-Weinberg equilibrium using TagIT, a conventional practice in assessing the quality of genotype data. Although there are a number of circumstances including mutation and demographic changes that could lead to the perturbation of HWE, deviations from HWE proportions may also arise from genotyping errors. Non-significant $\chi^2$ values for all 16 loci ($\chi^2 < 3.841, p<0.05; 1$ d.f.), indicated consistency with HWE.

### 4.5 The haplotype structure of DYNC1H1

The utility of using CEPH trio genotype data is apparent when identifying haplotypes across DYNC1H1. Although a number of frequently occurring haplotypes were visible by manual inspection of the genotype data, TagIT was used to comprehensively identify transmitted haplotypes from parent to child within each trio using the 'knownhap' function. In this manner, TagIT resolved a total of 116 transmitted haplotypes from a possible 128, or 90% of the total haplotypes possible.

**Table 4.4 The haplotype structure of DYNC1H1 in CEPH**

<table>
<thead>
<tr>
<th>ID</th>
<th>Haplotype</th>
<th>Transmitted Count</th>
<th>Freq (%)</th>
<th>EM Count</th>
<th>Freq (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>89</td>
<td>76.72</td>
<td>95</td>
<td>74.20</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>7</td>
<td>6.30</td>
<td>8</td>
<td>6.20</td>
</tr>
<tr>
<td>3</td>
<td>G</td>
<td>3</td>
<td>2.59</td>
<td>3</td>
<td>2.30</td>
</tr>
<tr>
<td>4</td>
<td>C</td>
<td>3</td>
<td>2.59</td>
<td>3</td>
<td>2.30</td>
</tr>
<tr>
<td>5</td>
<td>E</td>
<td>2</td>
<td>1.72</td>
<td>2</td>
<td>1.56</td>
</tr>
<tr>
<td>6</td>
<td>F</td>
<td>1</td>
<td>0.86</td>
<td>2</td>
<td>1.56</td>
</tr>
<tr>
<td>7</td>
<td>D</td>
<td>1</td>
<td>0.86</td>
<td>3</td>
<td>2.30</td>
</tr>
<tr>
<td>8</td>
<td>H</td>
<td>1</td>
<td>0.86</td>
<td>2</td>
<td>1.56</td>
</tr>
<tr>
<td>9</td>
<td>I</td>
<td>1</td>
<td>0.86</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>10</td>
<td>J</td>
<td>1</td>
<td>0.86</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>11</td>
<td>K</td>
<td>1</td>
<td>0.86</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>12</td>
<td>L</td>
<td>1</td>
<td>0.86</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>1</td>
<td>0.86</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>14</td>
<td>N</td>
<td>1</td>
<td>0.86</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>15</td>
<td>O</td>
<td>1</td>
<td>0.86</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>16</td>
<td>P</td>
<td>1</td>
<td>0.86</td>
<td>1</td>
<td>0.78</td>
</tr>
<tr>
<td>17</td>
<td>Q</td>
<td>1</td>
<td>0.86</td>
<td>1</td>
<td>0.72</td>
</tr>
</tbody>
</table>

The 12 haplotypes not observed were attributable to ambiguous SNP phase in trios 9, 17 and 30; each possessing one or more SNP locus heterozygous in all three trio members (Figure 4.4). Proceeding with the readily identifiable, transmitted haplotypes to discern the pattern of LD across
DYNCIHI would have omitted information from approximately 10% of the total genotyping dataset due to ambiguous phase, therefore the phase-ambiguous haplotypes needed to be identified.

Figure 4.4 Phase-ambiguous SNPs in CEPH trios

In CEPH trio 9, both parents and child are heterozygous at all loci except SNPs 11, 12 and 16; SNP 4 is shown. In CEPH trio 17, both parents and child are heterozygous at SNP locus 16. In CEPH trio 30 both parents and child are heterozygous at SNP loci 13, 14 and 16; SNP 13 is shown. Paternal SNP alleles are in black, maternal SNP alleles in red and the two possible alleles inherited by the children are shown. CEPH identification numbers are given above/below the pedigree symbols.

The SNP/haplotype ambiguity was resolved by applying an Expectation Maximisation (EM) algorithm to infer the missing haplotypes. The EM algorithm, first proposed by Clark in 1990, determines resolved haplotypes from individuals with no haplotype ambiguity and sequentially applies these to phase-ambiguous individuals, to resolve their haplotypes (Clark, 1990). The algorithm uses maximum likelihood estimates of haplotype frequencies in the sample using an initial set of haplotype frequency estimates to calculate the conditional distribution for haplotype pairs that each individual carries (the Expectation step). The initial set of haplotype frequencies were restricted to those phased haplotypes identified from the CEPH trios. Based on these conditional distributions, haplotype frequency estimates can be updated (Maximization- step) and the algorithm iterates between the two steps until the haplotype frequency estimates converge. Haplotypes were inferred using the EM algorithm implemented within TagIT, with a minimum frequency threshold of ≥1%.

The EM analysis identified a total of 17 different haplotypes ≥1% frequency, 15 of which were previously seen transmitted amongst the trios (Table 4.4). The EM estimated frequencies agreed well with the transmitted haplotype frequencies and both data sets show that haplotype A predominates, present at approximately 75% on average. Two new haplotypes P and Q were identified through the EM procedure, although these were present at low frequency (1%) and a greater number of haplotypes were seen at 3% frequency, which most likely represent those haplotypes that were previously unresolved.
Interestingly, several pairs of haplotypes were seen with completely mismatching alleles (i.e. nucleotides differing at every SNP position in a haplotype pair) which, at the time they were identified, had not previously been reported by any other groups. The high frequency of these mismatching haplotype pairs: A+F, B+C and to a lesser extent J+M, seemed an odd occurrence and through successive discussions with colleagues it was suggested that these haplotypes may represent a genetic signature of selection or population demography. These haplotypes have since been termed by Zhang and colleagues as “yin yang haplotypes” (Zhang et al., 2003). Finally, the haplotype data provided encouraging anecdotal evidence for linkage disequilibrium across the region: the A haplotype comprised approximately 75% of the total haplotype diversity, with the remaining 25% comprised of haplotypes at much lower frequencies many of which are variants of haplotype A.

4.6 Linkage disequilibrium patterns

The empirical assessment of linkage disequilibrium across the gene was a crucial measure to determine the magnitude of economy, if any, available in using tagging SNPs to represent variation across DYNC1H1 and also in selecting the tSNPs.

The extent of LD across DYNC1H1 was calculated by assessing the frequency of 2-locus haplotypes (i.e. pairwise comparisons) within the total pool of phased chromosomes and then calculating $D'$ and $r^2$ measures from these values. In view of the haplotype frequencies discussed above and considering how the $D'$ measure is calculated ($D'$ is maximal when two of the four possible haplotypes, in a two-locus situation for example, are absent – see $D'$ derivation from Chapter 1) high $D'$ scores between the majority of pairwise SNP comparisons were predicted.

![Figure 4.5](image)

*Figure 4.5 Linkage disequilibrium ($D'$) between the 16 SNPs spanning DYNC1H1*
Assessing pairwise LD between the 16 SNPs across DYNC1H1 reveals two regions of elevated LD. SNP ID numbers arranged on the left from most 5' (top) to the 3' (bottom). Colour scale bar represents $D'$ values from 0 (white) to 1 (black).

Figure 4.5 provides a graphical illustration of pairwise $D'$ between the 16 DYNC1H1 SNPs calculated using TagIT and substantiated the high $D'$ prediction. Average $D'$ across the gene was 0.98 and although it is presented here as mainly a descriptive statistic, it supports the original power calculation made (where $D' \sim 1$). $D'$ also informs on recombination across the region, which appears to be absent except for at the 3’ most SNP.

As previously discussed, a more useful measure of LD for association studies is $r^2$. Although $r^2$ is sensitive to marker allele frequencies it does provide a simple linear relationship to sample size, as the reciprocal of $r^2$ (that is, $1/r^2$) relates to the proportional increase in sample size required to obtain the same power as testing the functional variant itself (Pritchard et al., 2001). With this linear relationship in mind, a minimum working threshold of $r^2 \geq 0.85$ was used, requiring an increase in sample size of $\sim 18\%$ to have equivalent power of typing the causal SNP itself. Figure 4.6 provides a graphical illustration of pairwise $r^2$ values between the 16 DYNC1H1 SNPs. The $r^2$ data clearly show two defined regions of elevated LD: the first region bounded by SNPs rs2720193 to rs3742426 and the second region by SNPs rs941793 to rs1190613. Although the average LD across the gene was $r^2=0.61$, within the two regions LD was $r^2=1$ and $r^2=0.91$ on average.

Figure 4.6 Linkage disequilibrium ($r^2$) between the 16 SNPs spanning DYNC1H1
Assessing pairwise LD between the 16 SNPs across DYNC1H1 reveals two regions of elevated LD. SNP ID numbers arranged on the left from most 5' (top) to the 3' (bottom). Colour scale bar represents $r^2$ values from 0 (white) to 1 (black).
Subsequent to identifying and phasing the CEPH *DYNC1HI* haplotypes and calculating pairwise LD amongst the 16 SNPs, the investigation converged to identifying tagging SNPs. Tagging SNP selection was performed using Goldstein and Weale’s TagIT software. Goldstein and Weale’s method uses a direct approach to identify a set of tSNPs that maintain high $r^2$ values with the other SNPs (thus tagging them), or that define a set of haplotypes that do so. In general, there are two approaches in which a reduced set of tSNPs can be selected to represent a larger set of known SNPs. The first approach is based on capturing as much of the original haplotype diversity present in a set of known SNPs when condensed to a smaller set of tSNPs, as exemplified by Johnson and colleagues 2001 paper (Johnson *et al.*, 2001). The second approach involves establishing a maximal association between the reduced tSNP set and the original set of known SNPs from which they are derived, as described in Chapman’s 2003 paper (Chapman *et al.*, 2003). Although TagIT incorporates both of these approaches to identify tSNP sets, the association based approach is a more directly relevant approach to SNP prediction in the context of an association study (Weale *et al.*, 2003), and therefore all tSNPs in this study were selected on association based performance.

The size of a tSNP set may be determined empirically or predefined. Identifying tSNP sets of a predefined size can be useful if tSNP set size is constrained by cost (for example, if a large number of genes are being tested in a single study); however, this method can underpower studies, especially in regions of low or variable LD. Instead, minimal tagging SNP sets can be chosen based on their ability to exceed a minimum performance threshold. In this circumstance the tSNP set size varies to provide a minimum level of association with the larger set of known SNPs from which the tSNP are derived. In selecting tSNPs to represent variation across *DYNC1HI* the latter approach was used. The measure used to evaluate tSNP performance was based on an “average locus” association criterion, termed by Goldstein and Weale as ‘*haplotype $r^2$’*, which has been suggested to be an optimal performance measure for use in association studies (Weale *et al.*, 2003) and in estimating associations between SNPs with MAF of 6% or more (Goldstein pers. comm.). This measure reduces the performance of a tSNP set against all known SNPs, to a single value by taking the weighted average (weighted relative to the minor allele frequency at each locus) of each individual performance measure. The minimal performance threshold used in selecting a tSNP set was set at $r^2 \geq 0.85$.

All tSNP sets of varying size $H$ were identified for which the average locus performance criterion for each tSNP set was $r^2 \geq 0.85$ against the 16 dynein SNPs (Figure 4.7).
Figure 4.7 Average locus haplotype $r^2$ performance of all tSNP sets sizes from $H=1$ to $H=5$

$H=1$ there are 16 possible tSNP sets, $H=2$ there are 120 possible tSNP sets, $H=3$ there are 560 possible tSNP sets, $H=4$ there are 1820 possible tSNP sets and $H=5$ there are 4368 possible tSNP sets.

The highest performing tSNP sets were identified for different tSNP sizes. As $H$ increased, the performance of all tSNP sets approached 1, as a greater number of variants were tagged. But which value of $H$ is optimal? As Figure 4.8 illustrates, the minimum tSNP set size to achieve an average locus association performance threshold $r^2\geq 0.85$ is $H=2$. The data shown in Figure 4.8 are the “best”, or highest scoring, tSNP sets taken from the average locus analysis, although more than one tSNP set of a given size can be “best”). There were diminishing returns in increasing $H>4$, with no appreciable increase in performance, and increasing tSNP set size beyond $H=7$ yields no additional benefit and therefore, in an average locus analysis, all haplotypes can be tagged by a minimum of 7 tSNPs. A two tSNP set exceeded the 0.85 threshold and also performed well in representing the diversity of haplotypes available – capturing 94% of the total haplotype diversity.
Figure 4.8 Performance of "best" tSNP sets of increasing size $H$ chosen using two criteria

Performance measure is given as average locus haplotype diversity (blue) or average locus haplotype $r^2$ (red).

The tSNPs that related to these optimal sets are shown in Figure 4.9. For tSNP sets $H>2$ there were multiple 'best' sets of equal performance. Using the TagIT 'performL' function, it was possible to decompose the performance of each best tSNP set, to show performance against individual loci (Figure 4.9). The optimal values were derived by combining the values for each tSNP against each locus and further combining the single locus scores to provide a single value for a tSNP set. The 'performL' function displayed the combined haplotype $r^2$ values for each locus and this manner it is possible to see which loci are influencing the result.

The results indicated that a solo tSNP set (SNP 5 or 9) was insufficient to capture the variation of the entire region - each single tSNP tags the region of elevated linkage disequilibrium it resides in. In concordance with the LD data and the haplotype data, the SNP tagging data confirmed that two "sub-regions" with different properties exist across the gene and that these regions could not be summarised by a single tSNP. A two tSNP set overcame this problem and although a weighted average of the performance values was greater than 0.85, it was clear that there were four SNPs that could not be tagged well (SNPs 1, 10, 12 and 16). These SNPs could only be reliably tagged by increasing the tSNP set size to $H=3$ and $H=4$ and only if they themselves were chosen as tSNPs (i.e., they are tagging nothing but themselves).
Figure 4.9 Optimal tSNP sets of size $H$ and their performance against all loci

Left. Table gives "best" tSNP sets for varying set size $H$. SNPs 1 to 16 are rs1741155, rs2720193, rs2720194, rs2273438, rs2749890, rs2251644, rs3742426, rs2180510, rs941793, rs13749, rs941792, rs1004903, rs10135238, rs1190610, rs1190613 and rs1190618 respectively. Right. Performance of "best" tSNP sets for each locus. All performance tested using association criterion 5. Broken grey line represents 0.85 performance threshold.

4.7.1 Assessing tSNPs – "SNP-dropping"

Of paramount importance in selecting tagging SNPs is the performance of a chosen tag set to capture unknown variation. This was tested using the TagIT SNP-dropping routine "excludes", to sequentially drop each SNP from the full SNP set, to simulate an "unknown" polymorphism, and
assess the performance of tSNPs selected from the remaining reduced SNP set to tag the dropped SNP. The SNP-dropping routine was carried out for varying tSNP set sizes \( H \) and single-locus performance criteria values were returned for each dropped SNP (Figure 4.10). As shown previously, more than one tSNP set was defined as equally “best” for increasing values of \( H \) and in these circumstances the single-locus performance criteria values returned were averaged over all “best” tSNP sets.

![Figure 4.10](image)

**Figure 4.10** SNP dropping performance for “best” tSNP sets of varying size

Performance relates to average single locus performance criterion 5.

Very little difference was seen in performance with increasing \( H \) indicating diminishing performance returns (Figure 4.10). The procedure confirmed the difficulty in tagging SNP 16, and to a lesser extent, SNP 10 and 12. When SNP 16 was dropped, no tag set chosen from the remaining 15 SNPs was able act as a proxy, indicating that the only method to represent SNP 16 and its associated variation would be to include SNP 16 as a tagging SNP itself. But was it imperative that this SNP and its associated variation be tagged? The position of SNP 16, approximately 15kb from the last exon of \( DYNClHI \), and the absence of appreciable LD with SNP 15 (which is only 4.5kb from exon 78) suggested that it probably did not tag any functional variants in the 3'UTR and so, tagging this SNP was not a priority.

Taken together, these analyses suggested that a two tSNP set could provide the optimal performance. It was unlikely that any appreciable gain would have been seen in adding extra tSNPs to this set, apart from tagging low frequency clades of the haplotype genealogy, however the occurrence of a functional variant in a low frequency clade cannot be ruled out. Of the 14 possible two tSNP combinations shown in Figure 4.9, tSNPs 6 (rs2251644) and 9 (rs941793) were chosen to genotype in cases and controls. As all 14 tSNP sets performed equally well, the only discriminating criteria was the ease of genotyping - both rs2251644 and rs941793 modify restriction endonuclease recognition sequences and therefore these SNPs were pursued.
4.7.2 Restriction endonuclease assays for rs2251644 and rs941793

The major allele of tSNP rs2251644 was found to destroy an NsiI endonuclease recognition site within the 270bp PCR amplicon originally used for genotyping in the CEPH trios, and the minor allele of tSNP rs941793 created a SacI recognition sequence within its the 300bp amplicon. To control for partial and complete digestion failures, additional invariant restriction endonuclease recognition sites were investigated as internal controls. These internal control endonucleases were chosen on the basis that double digestions could be performed under identical conditions (at identical temperatures and buffers) and that the cleavage products can be easily resolved on 2.5% agarose gels. No suitable restriction endonuclease was identified as an internal control for rs2251644 cleavage, however, NheI was found to cleave at a site 100bp from SacI as an internal control for rs941793 (Figure 4.11).

All tSNP genotyping and restriction endonuclease cleavage optimisations were performed at the Cecil B. Day Laboratory for Neuromuscular Research, Massachusetts General Hospital, Boston (USA) with the help of Dr. Azlina Ahmad-Annuar. Non-specific amplification of the rs2251644 SNP PCR product was sporadically seen despite identical primer sequences being used to those used for genotyping of CEPH trios, which had previously worked well in London. Inconsistency of the rs2251644 PCR product may have been due to differences in the reagents and thermal cyclers used between the two laboratories and dictated the design of new primers for genotyping the tSNP. The new primers (rs2251644CBDay, shown in Appendix 2 Primer Table) provided a repeatable and robust PCR product, approximately 392bp in length.

Due to the absence of an internal control for tSNP rs2251644 and a change of genotyping reagents, protocol and location from those previously used, the restriction endonuclease assays were checked for genotyping fidelity. Genotypes of 96 randomly selected control individuals (from a panel later used in the sporadic ALS association study) for both tSNPs, were determined by double digest and sequencing on a Beckman Coulter Genetic Analysis system (Figure 4.11). The operator was blinded to all sample identifiers to ensure impartiality. Genotypes from both assays were compared for consistency and acceptable error rates of approximately 1% were seen (<1% for rs2251644 and 1.09% for rs941793).

* A double digest is the cleavage of a single DNA moiety by two restriction endonucleases
## Table 4.11 Restriction digests for tSNPs rs2251644 and rs941793

<table>
<thead>
<tr>
<th>ID</th>
<th>Major allele</th>
<th>Minor allele</th>
<th>Enzyme</th>
<th>Primary</th>
<th>Internal control</th>
<th>PCR conditions</th>
<th>Fragment size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2251644</td>
<td>A</td>
<td>G</td>
<td>Nsil</td>
<td>atgca</td>
<td>n/a</td>
<td>37°C</td>
<td>230 392 392</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SacI</td>
<td>gact</td>
<td>Nhel</td>
<td>37°C</td>
<td>212 212 112</td>
</tr>
<tr>
<td>rs941793</td>
<td>A</td>
<td>G</td>
<td>Nsil</td>
<td>atgca</td>
<td>Nhel</td>
<td>37°C</td>
<td>88 112 100</td>
</tr>
</tbody>
</table>

### rs2251644

**Figure 4.11** Restriction digests for tSNPs rs2251644 and rs941793

**Top.** Table summarising tSNP restriction endonuclease properties and expected product sizes. No internal control was identified for rs2251644. Fragment sizes are given in base pairs after complete digestion of AA (major allele homozygotes), Aa (heterozygotes) and aa (minor allele homozygotes) by the respective enzymes. **Middle.** rs2251644 PCR amplicon schematic, gel electrophoresis results and sequence verification. **Bottom.** rs941793 PCR amplicon schematic, gel electrophoresis results and sequence verification.
4.8 Phase III - Testing in sporadic ALS cases and controls

4.8.1 Association study samples

Sporadic ALS case samples were generously provided by R.H. Brown at Massachusetts General Hospital or collaborators from various academic and clinical institutions across North America. All sporadic ALS patient samples were obtained with accompanying signed consent. Patients were diagnosed by El Escorial criteria and had not been screened for SOD1 mutations. The patient cohort comprised of 134 females and 147 males with an average age at diagnosis of 46 years (range 24 to 79 years). 34% of patients were diagnosed with early onset in the lower extremities, 35% with upper onset, 30% with bulbar onset and site of onset in the remaining 2% was unknown. As all samples were obtained retrospectively, limited phenotypic information was available.

4.8.2 Tagging SNP rs2251644 and rs941793 genotyping in cases and controls

With the endonuclease assay functioning well, the DYNCIHI tSNPs were genotyped in 281 North American sporadic ALS cases and 225 age matched controls (Table 4.5).

<table>
<thead>
<tr>
<th>tSNP ID</th>
<th>Sample</th>
<th>Genotype count observed</th>
<th>Genotype count expected</th>
<th>Allele frequencies</th>
<th>HWE Analysis</th>
<th>Association test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AA</td>
<td>171</td>
<td>173.4</td>
<td>A = 0.88</td>
<td>χ² = 2.15;</td>
</tr>
<tr>
<td>rs2251644</td>
<td></td>
<td>GA</td>
<td>53</td>
<td>48.3</td>
<td>G = 0.12</td>
<td>p = 0.357</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG</td>
<td>1</td>
<td>3.4</td>
<td></td>
<td>χ² = 1.40;</td>
</tr>
<tr>
<td></td>
<td>control</td>
<td>AA</td>
<td>202</td>
<td>202.4</td>
<td>A = 0.87</td>
<td>χ² = 0.04;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GA</td>
<td>60</td>
<td>59.3</td>
<td>G = 0.13</td>
<td>p = 0.850</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG</td>
<td>4</td>
<td>4.4</td>
<td></td>
<td>p = 0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA</td>
<td>141</td>
<td>140.8</td>
<td>A = 0.75</td>
<td>χ² = 0.004;</td>
</tr>
<tr>
<td>rs941793</td>
<td>SALS</td>
<td>GA</td>
<td>94</td>
<td>94.4</td>
<td>G = 0.25</td>
<td>p = 0.100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG</td>
<td>16</td>
<td>15.8</td>
<td></td>
<td>χ² = 3.20;</td>
</tr>
<tr>
<td></td>
<td>control</td>
<td>AA</td>
<td>120</td>
<td>126.9</td>
<td>A = 0.75</td>
<td>χ² = 6.12;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GA</td>
<td>98</td>
<td>84.1</td>
<td>G = 0.25</td>
<td>p = 0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG</td>
<td>7</td>
<td>13.9</td>
<td></td>
<td>p = 0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT</td>
<td>69</td>
<td>74.3</td>
<td>T = 0.77</td>
<td>χ² = 7.48;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TC</td>
<td>54</td>
<td>43.4</td>
<td>C = 0.23</td>
<td>p = 0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CC</td>
<td>1</td>
<td>6.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA</td>
<td>70</td>
<td>75.3</td>
<td>A = 0.78</td>
<td>χ² = 7.36;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GA</td>
<td>54</td>
<td>43.5</td>
<td>G = 0.22</td>
<td>p = 0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG</td>
<td>1</td>
<td>6.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.5 Genotype frequencies of tSNPs rs2251644 and rs941793 in sporadic ALS cases and matched controls

Observed and expected tSNP genotype frequencies are shown for sporadic ALS (dark grey panels) and controls (light grey panels). No significant differences were seen in genotype or allele frequencies between cases and controls. Deviation from Hardy-Weinberg equilibrium was seen based on χ² distribution at tSNP rs941793 in the controls and at two additional SNP loci rs10135238 and rs1190610 (cream panels).
4.8.2.1 Association study genotyping accuracy

The genotype data was tested for deviations from Hardy-Weinberg proportions using the software HWSIM (http://krunch.med.yale.edu/hwsim) to assure the quality of the data and prevent errors or cryptic genetic factors leading to a spurious association. HWSIM calculated expected genotype frequencies and performed a Pearson $\chi^2$-test for accordance between the observed and expected genotypes. As $\chi^2$-tests are a poor approximation when one or more genotype class has a value less than 5, HWSIM was used to implement a Monte-Carlo permutation test with 10,000 iterations for rs2251644 case and control data. A two-tailed test was used (obtained by counting how many times, across iterations, the original data and all other distributions that are more deviant from Hardy-Weinberg were observed) which, compared to a one-tailed test, did not require a judgement to be made on whether greater or fewer heterozygotes are to be expected. Two-tailed values for rs2251644 and rs941793, in cases were $p=0.3574 \ p=1.000$. All other genotypes were tested using HWSIM based on the $\chi^2$ distribution and were seen to be in HWE (i.e. $\chi^2<3.84$, with 1 degree of freedom; d.f.) except for rs941793 in the control panel ($p=0.01$), which showed an excess of GA heterozygotes and deficit of GG homozygotes. The relevance of which is discussed further below.

4.8.2.2 Hardy-Weinberg disequilibrium in control samples

There are several possible explanations to account for the deviation of rs941793 from HWE in control samples. The first could have been due to a poorly functioning genotyping assay. As the SacI restriction endonuclease assay for rs941793 cleaves the minor (G) allele, it was initially thought that the deficit of GG homozygotes could be due to suboptimal or incomplete digestion of PCR products. However, on closer inspection of the agarose gel images of the cleavage products, the internal control enzyme had cleaved correctly. In addition, due to the nature of this assay, suboptimal/incomplete digestion of the PCR products would have resulted in a reduction of heterozygotes rather than an excess, which is contrary to the observed data. Alternatively a polymorphism within the rs941793 primer binding region, could have caused the biased amplification of the major allele, although this is unlikely as both the case panel and the discovery CEPH panel would be similarly affected. In addition, no polymorphisms in the rs941793 primer binding regions have been identified in any SNP database to date.

As a number of 2 tSNP sets were identified as “best”, a number of additional SNPs were available to replace rs941793. Two such SNPs were rs10135238 and rs1190610, present on the same amplicon. These were both genotyped in the control cohort by sequencing using the Beckman Coulter Sequencer – this deliberate move to use a different genotyping platform was to avoid any assay-related errors that may have lead to the deviation from HWE (and it allows both SNPs to be genotyped simultaneously). Both SNPs were found to deviate significantly ($p=0.01$) from the chi-
squared distribution, however, simulated \( p \)-values showed that the results were marginally significant Monte-Carlo \( p \)-values \( rs10135238 \ p=0.0495 \) and \( rs1190610 \ p=0.0466 \). Analysis of SNP \( rs941793 \), using Monte-Carlo permutation tests reduces the significance of deviation to \( p=0.07 \), suggesting that based on the genotypes seen, the probability of obtaining equally deviant data sets with a paucity of heterozygotes is highly likely. Taken together, these SNPs show a trend to significance which is not entirely unexpected, given the high LD between \( rs941793 \), \( rs10135238 \) and \( rs1190610 \) and suggests that the result may be a genuine property of the control cohort and not an assay-related error.

### 4.8.2.3 A sampling bias?

To identify if the HWE deviation was due to an intrinsic property of the control cohort generated through a sampling bias, the details of all control samples were reviewed with Diane McKenna-Yasek, coordinator of the Day Lab DNA collection. Samples from the control panel were checked to identify related individuals, which could result in a skewing of allele frequencies by over-representing one genotype compared to another. The review confirmed that the majority of the cohort was comprised of spousal controls, precluding relatedness between the individuals.

### 4.8.2.4 Hardy-Weinberg equilibrium at genes independent of \( DYNCIHI \)

Several different factors can lead to perturbations of HWE, such as demographic processes, selection and genotyping error. Although demographic processes, such as population admixture, can lead to deviations in HWE, these effects are observed genome-wide. Could the deviation from HWE seen in the control samples be due to a demographic or sampling error? This question was answered by assaying for deviations from HWE at sites independent of \( DYNCIHI \). Data were available from two additional studies which used the same control panel to that used in the heavy chain study: 8 SNPs from a study of the cytoplasmic dynein intermediate chain 1 (\( DYNCIII \)) gene on chromosome 7 (Azlina Ahmad-Annuar, pers. comm.) and 3 SNPs in the vascular endothelial growth factor (\( VEGF \)) gene on chromosome 6 (Carsten Russ, pers. comm.) were all in HWE (Table 4.6). This suggests that the HWE perturbations seen at \( DYNCIHI \) may be a locus-specific artefact in the controls rather than genome-wide effect due to demographic processes. Genotyping errors may also result in significant deviations from HWE and although the majority of HWE deviations are due to problematic and non-specific assays, there remain a number of instances where no root cause for genotype error is detectable (Hosking et al., 2004).

### 4.8.3 A comparison of tSNP genotype frequencies between cases and controls

Despite the deviation from HWE, genotypes of the two tSNPs were analysed for significant differences between cases and controls, using the software. The null hypothesis was that there was...
no significant difference between the two samples and this was tested using a Fisher's exact test of the tSNP genotypes and $\chi^2$-test of the tSNP allele frequencies. The tSNPs showed no significant differences in genotype proportions between sporadic ALS cases and age matched controls (Table 4.5).
<table>
<thead>
<tr>
<th>tSNP ID</th>
<th>Gene</th>
<th>Genotype count</th>
<th>Allele frequency</th>
<th>HWE analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Observed</td>
<td>Expected</td>
<td></td>
</tr>
<tr>
<td>rs940424</td>
<td>DYNC1I1</td>
<td>AA 10</td>
<td>8.08</td>
<td>χ² = 0.77; p = 0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AT 51</td>
<td>54.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT 95</td>
<td>93.08</td>
<td></td>
</tr>
<tr>
<td>rs2690289</td>
<td>DYNC1I1</td>
<td>AA 46</td>
<td>50.94</td>
<td>χ² = 2.55; p = 0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AG 88</td>
<td>78.11</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG 25</td>
<td>29.94</td>
<td></td>
</tr>
<tr>
<td>rs3129314</td>
<td>DYNC1I1</td>
<td>AA 63</td>
<td>58.63</td>
<td>χ² = 1.85; p = 0.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AG 76</td>
<td>84.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG 35</td>
<td>30.63</td>
<td></td>
</tr>
<tr>
<td>rs1488513</td>
<td>DYNC1I1</td>
<td>AA 32</td>
<td>32.60</td>
<td>χ² = 0.04; p = 0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AG 80</td>
<td>78.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG 47</td>
<td>47.60</td>
<td></td>
</tr>
<tr>
<td>rs1685818</td>
<td>DYNC1I1</td>
<td>GG 28</td>
<td>29.10</td>
<td>χ² = 0.13; p = 0.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GT 83</td>
<td>80.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT 55</td>
<td>56.10</td>
<td></td>
</tr>
<tr>
<td>rs81018</td>
<td>DYNC1I1</td>
<td>AA 40</td>
<td>45.38</td>
<td>χ² = 3.14; p = 0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AG 85</td>
<td>74.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG 25</td>
<td>30.37</td>
<td></td>
</tr>
<tr>
<td>rs720780</td>
<td>DYNC1I1</td>
<td>AA 62</td>
<td>59.01</td>
<td>χ² = 1.16; p = 0.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AG 61</td>
<td>66.98</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GG 22</td>
<td>19.01</td>
<td></td>
</tr>
<tr>
<td>rs2299282</td>
<td>DYNC1I1</td>
<td>GG 9</td>
<td>7.21</td>
<td>χ² = 0.71; p = 0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GT 51</td>
<td>54.57</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT 105</td>
<td>103.21</td>
<td></td>
</tr>
<tr>
<td>rs6966540</td>
<td>DYNC1I1</td>
<td>CC 13</td>
<td>15.02</td>
<td>χ² = 0.63; p = 0.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CT 61</td>
<td>56.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TT 52</td>
<td>54.02</td>
<td></td>
</tr>
<tr>
<td>2578</td>
<td>VEGF</td>
<td>CC 45</td>
<td>44.9</td>
<td>χ² = 0.0292; p = 0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CA 104</td>
<td>105.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA 63</td>
<td>61.8</td>
<td></td>
</tr>
<tr>
<td>1154</td>
<td>VEGF</td>
<td>GG 81</td>
<td>78.9</td>
<td>χ² = 0.1380; p = 0.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GA 98</td>
<td>100.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA 33</td>
<td>32.2</td>
<td></td>
</tr>
<tr>
<td>634</td>
<td>VEGF</td>
<td>GG 103</td>
<td>100.9</td>
<td>χ² = 0.0109; p = 0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GC 90</td>
<td>90.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CC 19</td>
<td>20.4</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.6 *DYNC1I1* and *VEGF* control genotypes in Hardy-Weinberg equilibrium

SNPs 2578, 1154 and 634 are *VEGF* promoter polymorphisms

### 4.8.4 Comparison of tSNP haplotypes between cases and controls

As the tSNPs were originally selected based on their individual tagging performance and as haplotypes, two marker haplotype frequencies were next compared between cases and controls. The software FASTEHLPLUS (Zhao et al., 2002) was used to perform the haplotype association analysis based on the case control genotype data. FASTEHLPLUS was used to estimate two locus haplotype frequencies from the entirety of the genotype data and haplotypes were then compared between cases and controls. From the total dataset of 472 individuals, 64 records were discarded due to missing data.
<table>
<thead>
<tr>
<th>tSNP haplotype</th>
<th>Frequency (%)</th>
<th>( \chi^2 ) test</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2251644</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A A</td>
<td>0.75</td>
<td>0.74</td>
</tr>
<tr>
<td>A G</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>G A</td>
<td>0.09</td>
<td>0.00</td>
</tr>
<tr>
<td>G G</td>
<td>0.03</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Table 4.7 DYNC1H1 tSNP haplotype frequencies in sporadic ALS cases and matched controls

Two-marker haplotype frequencies calculated by FASTEHPLUS based on 408 unrelated individuals. No significant association was seen based on a \( \chi^2 \) test.

Based on comparing the two-marker haplotype frequencies in sporadic ALS cases and matched controls a heterogeneity statistic is outputted based on a \( \chi^2 \) distribution (\( \chi^2 = 0.32; p = 0.956 \)). Taken together, the individual tSNP loci, rs2251644 and rs941793 and their two-marker haplotypes show no significant association with sporadic ALS (Table 4.7).

4.9 Evolutionary analysis of the DYNCIH1 locus

Despite the absence of association between DYNC1H1 variants and SALS, the pattern of variation, more specifically the apparent reduction in diversity at the sequence level, was interesting and it was thought that analysis of this characteristic might hold potential clues as to the genes' role in ALS. To identify differences in genetic diversity and test for selection at DYNC1H1, the 16 SNP loci were genotyped in additional Japanese and Cameroonian populations. However, since this study began it has been shown that the hierarchal sampling method commonly used to ascertain SNPs, and the exportation of these SNPs for investigation in other populations, can introduce bias to the data which can distort tests of genetic diversity and selection (Akey et al., 2003; Weiss et al., 2002). Due to this SNP ascertainment bias, we were unable to exploit many of the indices commonly used to infer selection or deviations from neutrality. Many of the properties of these three populations presented here are affected by ascertainment bias and therefore they are provided as descriptive.

4.9.1 SNP ascertainment bias and impact on measurements of selection

To date, the most common strategy for discovering SNPs has been by full resequencing of a small number of samples, followed by targeted genotyping of these discovered SNPs in larger clinical samples. The initial appeal of this strategy is obvious – it makes sound economic sense - but it has since been shown to introduce bias which can confound subsequently population genetic analyses. The problem of this ascertainment bias has affected almost every large-scale polymorphism discovery project to some extent, including The SNP Consortium and HapMap projects (Clark et
(al., 2005) and affects the DYN1H1 study too. Ascertainment bias results from the fact that SNP discovery panels are often small, so that the probability that a SNP is identified in this sample, and later genotyped in a larger sample, is a function of the allele frequency (the probability of witnessing a rare allele is the same as its frequency). This conditional sampling of genetic variation imposes a bias, in that rare variants are likely to be missed in the larger sample. This frequency-specific distortion in SNP discovery consequently means that the SNP frequency spectrum obtained from a two-tier sampling will be different from that obtained under complete sampling (for example, by resequencing the entire study sample). As a result, statistical attributes that rely on the site frequency spectrum will be affected.

4.9.2 DYN1H1 chimpanzee genotyping

Before further analysis was undertaken, the ancestral allelic states of the 16 DYN1H1 SNP loci were assessed. Knowing the ancestral status of human polymorphisms can provide valuable evidence of evolutionary forces acting on the region. The 16 loci were genotyped, in 8 chimpanzees (Pan troglodytes) permitting alleles with MAF>6.25% be identified. DNA was extracted from chimpanzee blood (provided by the Institute of Zoology, London Zoo, Regents Park, UK) and genotyping conducted by bidirectional resequencing using identical primers and cycling conditions as those previously used on human samples.

<table>
<thead>
<tr>
<th>SNP number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chimpanzee</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>Human</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.8 Ancestral chimpanzee and common human DYN1H1 haplotypes

Ancestral alleles are shown in red and derived alleles in white. SNPs are indicated by position 5' to 3' along the gene.

In the chimpanzee, all loci corresponding to the human SNPs were either monomorphic or polymorphic at an undetectable frequency (MAF<6.25%) in this small sample. There was concordance between the ancestral chimpanzee and the common human SNP allele at only 9 loci (Table 4.8). There is a direct relationship of a derived* allele with age (Kimura et al., 1973) and it has been shown that (generally) for major SNP alleles with frequencies greater-than 80%, the major allele usually represents the ancestral allelic state (Hacia et al., 1999). This implies there should be a preponderance of DYN1H1 major SNP alleles identical to the chimpanzee alleles, as they are all present in the CEPH at frequencies >>80% - however there is concordance between the two species

* A derived allele is the term given to a new allele produced by mutation
at only 9 of the 16 loci. Moreover, the ancestral chimpanzee haplotype is completely absent from
the CEPH haplotype pool.

These relationships between allele frequencies and their age are generally true under conditions of
neutrality. So, does this evidence suggest that neutrality may have been perturbed at this locus? An
excess of high-frequency-derived mutations can be a signature of positive selection, or of genetic
hitchhiking* with linked alleles driven rapidly to fixation (Smith et al., 1974). However it may be a
pronounced feature of a population bottleneck, so to discriminate between the two, additional
populations could be useful.

4.9.3 Genetic variation of DYN1HI in additional populations

To determine if the pattern of variation seen at the DYN1HI locus in the European-derived CEPH
samples was confined to northern Europe or common across populations, the 16 DYN1HI SNPs
were genotyped in 48 unrelated individuals from east Asia (Japan) and 49 unrelated individuals
from West Africa, namely Bantu speaking Cameroonians. These samples were obtained from the
laboratory of Professor David Goldstein, with full and appropriate consent for research. With the
exception of SNP 16 in the Cameroonian samples ($\chi^2=10.8717; p<0.001$), all SNPs were in HWE.
There was no obvious reason why SNP 16 was not in HWE.

4.9.3.1 Allele frequency comparisons

Comparison of SNP frequencies, standardised by the derived CEPH allele, in each population,
showed a clear reduction in allele frequency in the northern European-derived samples as compared
to the other two populations (Figure 4.12).

Average gene diversity or expected allelic heterozygosity ($H$), which assesses the genetic diversity
in a population and may be characteristically altered by natural selection, was calculated under the
assumptions of HWE for each population, using Nei's measure of heterozygosity. Generally the
average gene diversity was expected to be higher in the West African population than the Japanese
or CEPH, as African populations are more diverse. Average gene diversity ($H \pm S.E.M$) across the
DYN1HI locus was: northern Europeans 0.211$\pm0.019$, Japanese 0.317$\pm0.016$ and Cameroon
0.440$\pm0.011$. A clear gradient in allelic heterozygosity was observed from the northern Europeans,
to the Cameroonians. This result is not unexpected as African populations generally display greater
diversity and populations which have later migrated out of Africa possess a part of that diversity.

* Genetic hitchhiking occurs when alleles linked to a selected mutation, which is driven rapidly to fixation, are
also dragged along to high frequency
Figure 4.12 A comparison of DYNC1H1 SNP allele frequencies between 3 populations

Minor allele frequencies are shown, standardized by allelic state in northern European, Japanese and Cameroonian populations.

4.9.3.2 Genetic differentiation

Interpopulation differences in gene diversity can be attributed to genetic differentiation due to causes including natural selection, population sub-division and admixture. To assess the possible role of selection in shaping population differentiation, Wright’s $F_{ST}$ is often calculated. This statistic allows allelic diversity of sub-populations to be measured against a global average to quantify genetic differentiation (varies between 0 and 1 with 1 being complete genetic differentiation). A modification of Wright’s measure is Weir’s weighted $F_{ST}$, which is used here to calculate total (global) and population pairwise comparisons of $F_{ST}$ at each locus (Figure 4.13). Global $F_{ST}$ ($\pm$ SEM) averaged across the entire locus is $0.25\pm0.04$ and pairwise comparisons of northern European against Japanese is $0.04\pm0.05$; against Cameroonian is $0.37\pm0.056$ and Japanese against Cameroonian is $0.29\pm0.065$. 
Figure 4.13 Global and pairwise $F_{ST}$ comparisons

Weir's $F_{ST}$ is shown for global and population comparisons at each SNP locus. The region bounded by SNPs rs2720193 and rs2180510 demonstrated less differentiation than other SNPs tested.

4.9.4 Haplotype and LD comparisons between three worldwide populations

The software PLEM was used to infer haplotypes for the unrelated Cameroonian and Japanese samples. Inferring haplotypes using unrelated samples can occasionally lead to spurious results and therefore haplotype frequencies were verified by reseeding the EM algorithm applied within PLEM with varying integers and also by using a second EM based program, SNPHAP, for conformation. No appreciable differences were seen in the results from varied seeds and either program. EM inference of Japanese haplotypes yields 9 haplotypes >1% frequency, suggesting a simple haplotypic structure of $DYNCHI$. In contrast with both the northern European and Japanese haplotypes, there are 28 haplotypes exceeding an estimated population frequency of 1% in the Cameroonian samples. Generally, greater allelic and haplotypic diversity was seen in the Cameroonian population compared to the Japanese and European-derived, which is consistent with
other reports (Stephens et al., 2001). Haplotype A, the most frequent haplotype in the CEPH at 74.2%, was found to be the highest frequency haplotype in all 3 populations: with frequencies of 67.9% and 21.4% in the Japanese and Cameroonian respectively (Table 4.9). Haplotype C is also common to all three populations and conversely to haplotype A which increases in frequency from the Cameroonian to northern European population, haplotype C is decreases in frequency.

A Ewens-Watterson test ($F$) was used to evaluate the observed haplotype frequency-distribution for goodness of fit with that expected under neutrality using the infinite alleles model of mutation (i.e. a model with no selection, where all mutations are selectively neutral) (Ewens, 1972; Watterson, 1978; Watterson, 1986). For the northern European population only, $F$ is positive and significant ($p=0.976$) at the 0.05 level in this two-tailed test, implying that the haplotype distribution is inconsistent with neutral conditions devoid of selection.
Table 4.9 Northern European, Japanese, Cameroonian and chimpanzee DYN1H1 haplotypes

Haplotypes >2% frequency shown with corresponding single letter identifiers. SNPs coded as major allele (-) and minor allele (+). Major SNP allele shown as uppercase.
The general pattern of pairwise LD in the Cameroonians was similar to that seen in the northern Europeans, comprising two regions of elevated LD between SNPs 2 and 8, and between 10 and 15. However, the pairwise LD values were the lowest of all three populations, reflected by a greater rate of LD decay with physical distance compared to the northern Europeans and Japanese. LD in the Japanese was a single elevated region delimited by SNPs 1 and 16 and LD was seen to decay at the lowest rate over distance.

![Linkage disequilibrium decay across DYNC1H1 in 3 populations](image)

**Figure 4.14 Linkage disequilibrium decay across DYNC1H1 in 3 populations**

Plotting pairwise $r^2$ values amongst the 16 SNPs spanning DYNC1H1 in N. European, Japanese and Cameroonian populations against distance between each comparison illustrates three very different patterns of LD decay.

Pairwise haplotypic $F_{ST}$ between populations was also calculated and $F_{ST}$ comparisons were highly significant for both the northern Europeans and the Japanese against the Cameroonians ($F_{ST}=0.228$, $p<<0.000001$ and $F_{ST}=0.1375$, $p<<0.000001$) and just significant for the CEPH versus Japanese ($F_{ST}=0.0348$, $p<0.02$).
These results showed a general reducing cline in average gene diversity at *DYNC1H1* from the northern European to the Cameroonian. This reduction in variation was further illustrated by an elevated global $F_{ST}$ value for all three populations combined, compared to the empirical genome-wide average of 0.123 (Akey *et al*., 2002), largely due to the extent of differentiation between the northern Europeans and Japanese against the Cameroonian.

### 4.9.5 Natural selection at *DYNC1H1?*

The haplotype results show an approximate cline in frequency of haplotype A from West Africa, to East Asia and to northern Europe (21%, 68% and 74%). Although this may be due to population demography which can confound the genetic signals of selection, the concomitant decrease in frequency of haplotype C in these populations (11%, 15% and 2%) suggests mechanisms other than demography may operate. We asked if the increase in frequency of the common haplotypes, reduction in diversity and to some extent genetic differentiation, could be signals of selection at this locus rather than demographic or neutral processes. Positive selection for an allele can reduce genetic diversity, elevate LD and lower the rate of LD decay with physical distance (Bamshad *et al*., 2003 see for review). The elevated LD levels in the northern Europeans compared to the Cameroonian may be a feature of selection rather than demography, although due to confounding by ascertainment bias, our results are unable to conclusively differentiate between the two.

### 4.10 Discussion

#### 4.10.1 *DYNC1H1* association study caveats

The involvement of *DYNC1H1* in the pathology of sporadic ALS remains unclear. Although this study has not detected an association of *DYNC1H1* variants with SALS, it does not preclude a role of the gene as a susceptibility factor as there are several caveats to the approach that must be acknowledged.

##### 4.10.1.1 An under-powered study?

The numbers of case and control samples used are likely to have not been amenable to detecting associated variants. As described previously, under optimistic allelic association conditions where $D'=1$, a sample size of 647 case individuals are required to achieve 80% power. In this study, although $D'$ was approximately equal to 1, a maximum of only 281 cases were available for screening, giving 51% power to detect an association (at a marker/disease allele frequency of 12%). This relative power of detection could be improved by supplementing this analysis with additional cases and controls, collected since this study was undertaken.
4.10.1.2 Changes in study design?

Since this work began in 2002, there have been a number advances in our understanding of the problems and pitfalls associated with association studies and their design. There are countless variables that can be potentially altered in any LD-based association study, including: SNP ascertainment strategies, densities, minor allele frequencies, haplotype inference, tagging SNP selection which have all come under recent scrutiny. Many of the changes proposed were introduced too late to influence this study but mark essential considerations for future association studies:

4.10.1.2.1 Two-stage SNP ascertainment

SNP discovery and validation are two issues that many contemporary association studies will not have to face, having access to *a priori* information on LD patterns and tSNPs in many different populations from sources such as the HapMap project. SNP discovery and validation was essential in this study to identify the underlying pattern of LD across DYNCIHI and to select tSNPs, however, this may have introduced a source of potential bias that reduced the overall power of the association study. Identifying or validating SNPs in a smaller sub-sample of individuals, although a common sense approach for any investigator with limited resources, biases the distribution of variants detected towards high frequency common variation. The deficit of SNPs of low frequency in this two-stage approach to SNP ascertainment means that the power to detect associations is reduced when the variants that actually cause disease are rare. Therefore, ascertainment bias may have eroded the power of this study to detect and tag rare variants associated with disease, when applied to a larger case/control sample. Conversely, the power to detect associations when the causal SNP is common is actually thought to be increased (Clark *et al.*, 2005).

* Two-stage SNP ascertainment describes the identification of SNPs in a limited sub-sample of individuals from which tagging SNPs are derived (stage one) and applied to test a larger sample of cases and controls (stage two). N.B. two-stage is not referred to here in terms of replication of a tSNP association in a larger sample set, as described in (Hirschhorn *et al.*, 2005).
Ascertainment bias is a common issue, known to also have affected large polymorphism discovery projects including the Human Genome Project and the HapMap and to date few robust methods exist to avoid introducing this type of bias or for correcting biased data (Clark et al., 2005).

4.10.1.2.2 What is the "right" SNP density and allele frequency?

Over the last five years, the most appropriate SNP density for association studies has been under considerable debate. Early studies using marker densities varying from SNP per kb to one SNP per 15kb (Dawson et al., 2002; Jeffreys et al., 2001; Patil et al., 2001) demonstrated that the extent of linkage disequilibrium is inversely related to marker density. Although this holds true for fine-scale LD, broader patterns of LD remain relatively insensitive to SNP (Ke et al., 2004). Gabriel and colleagues suggested that as a trade-off, using a SNP density of one common SNP every 7.8kb would be sufficient to capture the majority of common haplotype variation throughout the genome (Gabriel et al., 2002). However, Wang and Todd showed in 2003 using simulated data, that incomplete SNP maps of 1 SNP per 2.5kb, 5kb and 10kb ascertained approximately 75%, 50% and 40% of the total underlying common SNP variation, respectively, implying that only dense SNP maps would be useful for indirect association mapping (Wang et al., 2003). Choosing an appropriate marker density for an association study SNP map relies on a number of factors including SNP ascertainment criteria, SNP frequency, LD and tagging methods used (Wang et al., 2003). The HapMap has settled on map densities of 5kb (Phase I) and 1kb (Phase II) to provide researchers with data on fine-scale variation in LD and it is this information that will most likely
influence future studies: modifying SNP density for an association study according to the extent of LD (Dunning et al., 2000; Kruglyak, 1999; Pardi et al., 2005; Reich et al., 2001a)

Almost all recent association studies utilising LD patterns and haplotype structures in the human genome to identify disease susceptibility loci, have focused on common SNPs with a minor allele frequency of 5% or greater. The main rationale for this is the "common allele-common disease hypothesis" suggesting that the majority of common genetic disorders are caused by genetic variants common in the general population (Bueler et al., 1993; Collins et al., 1997; Lander, 1996). However, despite the limitations stipulated by the CDCV hypothesis, several authors have recently demonstrated that tag SNPs chosen from common (MAF ≥5%) variants can capture a proportion of rare alleles (1%<MAF<5%), dependent on the method of tag selection (using an exhaustive allelic transmission disequilibrium test, or EATDT, approach) (de Bakker et al., 2005; Lin et al., 2004). An interesting development of the influence of allele frequencies to the outcome of association studies comes from the frequency matching of SNPs. It was recently empirically shown that using the \( r^2 \) LD measure, if SNPs of equal or near equal frequency are compared, the correlation between them more accurately represents the true correlation than of SNPs of unmatched frequencies (Eberle et al., 2006) (an intuitive result considering the derivation of the \( r^2 \) equation). Although the resulting impact of this finding on the power of tag SNPs selected from these frequency matched variants has not yet been investigated, this information taken together with the EATDT data, may mean that future studies will not be worried so much about minimum MAF of each SNP, but more so about ensuring that MAFs are approximately equal.

4.10.1.2.3 Tag SNP selection

There are myriad algorithms implemented in software currently available to select tagging SNPs from genotype data (see Table 1 from (Halldorsson et al., 2004) for example and see (de Bakker et al., 2005; De La Vega et al., 2006; Howie et al., 2006; Liu et al., 2006)). It is beyond the scope of this thesis to review the relative merits of these algorithms against the TagIT method implemented in this study however, the TagIT routine (i.e. not the algorithm, but the way it is implemented) has been updated since its use in this study by Ahmadi et al. (Ahmadi et al., 2005) and can be assessed.

If an identical data analysis was undertaken today, the criterion used in the TagIT routine would be based on a ‘worst locus’ (TagIT criterion 11) rather than ‘average locus’ (TagIT criterion 5) which describes how the tSNP set of any size is described in terms of its performance against all of the original SNPs it was derived from (Ahmadi et al., 2005). The ‘worst locus’ approach returns a single value based on the minimum performance of any tSNP set, where as the ‘average locus’
approach returns a weighted average value. In this respect, the 'worst locus' approach is more conservative and provides a minimum performance for a tSNP set. Based on the original data and analysed with criterion 11, a minimum of 3 tSNPs would be needed to provide a minimum haplotype $r^2$ performance >0.85 (data not shown), implying that some regions of the gene may not have been sufficiently represented in the original study, thus reducing power to detect an association.

4.10.2 Case sample heterogeneity

As previously described in this chapter, the 281 sporadic ALS samples used in this study were obtained retrospectively with little accompanying phenotypic information. There was some variation in the pattern of onset amongst patients, with approximately a third of all patients presenting with either (i) early onset in the lower extremities, (ii) upper onset and (iii) with bulbar onset. Although these patients were screened by the El Escorial criteria, it is not known what proportion of patients were possible, probable or definite ALS and what proportion were diagnosed later with primary lateral sclerosis or primary muscular atrophy. These variations of ALS may have slightly different pathogenesis and therefore it is conceivable that different genes may be involved. Such phenotypic (and potentially genetic) heterogeneity can reduce the power of a candidate gene association study to identify a true association. In addition, genetic heterogeneity may have been introduced by the unknown $SOD1$ status of the case samples, which may have again reduced study power.

4.10.3 The availability of HapMap data

With SNP and HapMap data now available, the extent of power lost in using the study design adopted here could be cursorily examined. At the beginning of 2007, information on 42 coding SNPs within $DYNCIHI$ was available from dbSNP (which contrasts starkly to the single coding SNP discovered in this study) and 95 SNPs were identified from HapMap in the ~121kb sequence surrounding the $DYNCIHI$ locus. After trimming of SNPs with MAF<1%, 56 SNPs remained, giving an average SNP density of 1 SNP per ~2.1kb – over three-times the density empirically determined in this study. The HapMap data was analysed using Tagger, with MAF set to 0.01, 0.05 and 0.07, HWE p-value cut-off at 0.001, and minimum percentage genotype at 75%. At all MAF thresholds, strong LD was identified.
Figure 4.16 Linkage disequilibrium ($r^2$) across DYNC1H1 identified using HapMap data

SNPs from 121kb region centring on DYNC1H1 were analysed using Haplovie. Only SNPs with MAF $\geq$1% were included in the analysis.

It is difficult to directly compare the experimentally derived data with that from the HapMap as they are composed of different SNP sets, obtained in different ways and analysed by different methods. Using the aggressive tagging approach implemented in Haplovie (pairwise tagging and by 2 and 3 SNP haplotypes), only 63% of SNPs with MAF $\geq$1% could be tagged by two tSNPs $r^2$>0.8 (mean $r^2$=0.97). Both of these tSNPs (rs10135238 and rs2251644) were identified in the DYNC1H1 study and one, tSNPs (rs2251644), was used to interrogate case and controls. It could be extrapolated therefore, that the original tSNPs from the association study may have only tagged $\sim$63% of SNPs $\geq$ 1% frequency.
5 Genetic analysis of the cytoplasmic dynein subunit families

5.1 Introduction

The absence of association between \textit{DYNC1H1} and sporadic ALS and the propensity for ALS and other neurodegenerative disorders to demonstrate genetic heterogeneity provides a compelling case to investigate additional components of the cytoplasmic dynein complex for association with disease. The work presented in this chapter was born through a necessity for clarification of the mouse and human cytoplasmic dynein subunit data presented in the literature and in databases including nomenclature, genetic mapping positions and family relationships. The chapter begins with a justification for why this work was considered necessary and the results of an effort to collate the varied aliases historically applied to the subunits. Clarifying the cytoplasmic dynein subunit mapping positions is presented next, describing an \textit{in silico} methodology which also provided an excellent opportunity to survey both mouse and human genomes for novel cytoplasmic dynein paralogs. This chapter concludes with an \textit{in silico} experiment to identify cytoplasmic dynein subunit orthologs, a description of how these data have aided the implementation of a new system of nomenclature and the direction in which these results may lead a future dynein and ALS association study.

5.1.1 Beyond the cytoplasmic dynein heavy chain – the need to consider pathways

Despite the evidence implicating \textit{DYNC1H1} as both a causal locus and susceptibility factor for ALS and other motor neuron disorders, no association with disease has yet been found. As discussed previously, one potential explanation may be the confounding effect of non-allelic genetic heterogeneity (mutations in different genes resulting in identical phenotypes) reducing the power to detect a disease-associated variant. The propensity for familial ALS cases to demonstrate genetic heterogeneity and the precedence of genetic heterogeneity within single pathways in other neurodegenerative diseases (such as the APP pathway in Alzheimer's), it is plausible that genetic heterogeneity within the dynein pathway may elicit an ALS phenotype. In addition, tantalising evidence for dynein associated non-allelic genetic heterogeneity has come from the investigation of the dynein-associated complex dynactin, a subunit of which is mutated in familial forms of motor neuron disease (Munch \textit{et al.}, 2005; Puls \textit{et al.}, 2003). The next line of enquiry of the role of the cytoplasmic dyneins in ALS and other motor neuron disorders should consider subunits additional to the heavy chains of the dynein complex.
5.1.2 The cytoplasmic dynein subunits and a need for clarity

Cytoplasmic dynein is a large multi-subunit complex which, at its core, consists of a heavy chain homodimer bound by various intermediate, light intermediate and light chain polypeptides (Figure 5.1A, right panel).

![Figure 5.1 The mammalian cytoplasmic dynein complexes](image)

A. (left panel) Polypeptides of immunoaffinity purified rat brain cytoplasmic dynein. Polypeptide mass (kDa) and consensus family names are indicated. (right panel) Structural model of the cytoplasmic dynein complex: comprising a DYN1H1 homodimer core and intermediate, light-intermediate and light chain subunits. B. Structural model of the cytoplasmic dynein 2 complex. DYN2H1 is predicted to be similar to that of the cytoplasmic dyneins. DYN2L11 is the only known subunit of this complex. (Pfister et al., 2005b)

Investigation of these subunits as potential candidate loci for ALS and other motor neuron diseases exposed confusion associated with their nomenclature and the mapping positions of their genes in the literature and within databases. Before further study could be undertaken to investigate the dynein complex as a candidate in ALS, it was necessary to clarify the number of dynein subunits, their names, mapping positions and other information relevant for further study.

5.1.3 Standardising nomenclature – why agreeing a name for each gene is important.

The need to standardise human gene nomenclature was recognised as early as the 1970s when the first guidelines for human gene nomenclature were presented at the Edinburgh Human Genome Meeting (Shows et al., 1979) and is based largely on the fact that many genes share a common origin. The accepted dogma is that all present-day genes are likely to be derived from a ‘core’ of between 7,000 and 12,000 ancestral genes that existed more than 500 million years ago (Nebert et al., 2003). Therefore present-day genes, which exhibit homology within and between species, can be clustered together as families, and as such, should be identified as being related by being designated with a similar symbol.
However, in practice scientific communities working on diverse organisms have regularly named genes and proteins in an uncoordinated manner (see for example names such as ‘daughterless’, ‘groucho’ and ‘saxophone’ used by fruit-fly geneticists). These genes often have homologs in other species that may be given names based on a different nomenclature system specific to that community’s research. The situation becomes more complicated when you consider bias in letter usage which can result in different genes and proteins of different organisms, with quite different functions, having identical names and symbols (Nebert et al., 2004) - or vice versa. With the ever expanding number of genome sequencing initiatives (409 eukaryote genomes have been sequenced, of which 107 are animals; www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj, as of November 2006), providing vast amounts of gene and sequence data for myriad species, the only feasible method to handle and utilize data is to use computers. The ability of current computer software to accurately and efficiently cross-reference, cross-index and analyse all of these data, requires a systematic and co-ordinated approach to gene nomenclature to aid in data retrieval and to minimise errors (Nebert et al., 2003). Each gene therefore should have a unique core symbol which can be applied and amended to other members of a gene family (Wain et al., 2002).

One well utilised method for identifying members of a gene family prior to standardising their nomenclature is to use evolutionary trees – a phylogenetic approach – which has already been used for over 124 families/superfamilies (Nebert et al., 2004).

5.2 Clarifying mouse and human cytoplasmic dynein subunit nomenclature, genomic locations and accession numbers

Research on the cytoplasmic dynein subunits has historically been undertaken in myriad organisms from yeast (Saccharomyces cerevisiae) to humans (Homo sapiens). The nomenclature of mammalian genes encoding these proteins has drawn on homologs in other organisms, which in turn have often been defined and named by criteria such as their molecular mass and mobility on sodium dodecyl sulphate–polyacrylamide electrophoresis (SDS-PAGE) gels (Figure 5.1A, left panel) (Pfister et al., 2005b). Consequently a number of synonyms can be found for any given human or mouse cytoplasmic dynein subunit. Much of the early research into the cytoplasmic dyneins was conducted in the biflagellate green alga Chlamydomonas, on the dyneins found in the flagellar axoneme, and therefore some mammalian cytoplasmic dynein nomenclature derives from these studies. For example, mammalian members of the cytoplasmic light chain families DYNLIRB and DYNLL have commonly been referred to as LC7 and LC8 respectively, which are the names of homologous Chlamydomonas axonemal dynein subunits.
The lack of consistency and coordination of dynein nomenclature has also led to the erroneous cross-indexing of gene names and other data on databases. Consequently, this has introduced errors in nomenclature, mapping position and in the sequence accession numbers into the databases, which are propagated onto additional databases and even into the literature (the opposite is also true; errors in the literature are propagated onto database entries). For example, searching NCBI LocusLink for human and mouse gene entries with the root symbol ‘dhc’, often prefixed to the cytoplasmic dynein gene names, gives 17 results which include cytoplasmic and axonemal dyneins, several different enzymes and transmembrane proteins (Table 5.1).

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Organism</th>
<th>Gene symbol</th>
<th>LocusLink*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytoplasmic dynein heavy chain 1</td>
<td>Homo sapiens</td>
<td>DHC1</td>
<td>1778</td>
</tr>
<tr>
<td>Dynein heavy chain domain 1</td>
<td>Homo sapiens</td>
<td>DHCD1</td>
<td>144132</td>
</tr>
<tr>
<td>Axonemal dynein heavy chain 12</td>
<td>Homo sapiens</td>
<td>DHC3</td>
<td>8679</td>
</tr>
<tr>
<td>Transmembrane 7 superfamily member 2</td>
<td>Homo sapiens</td>
<td>DHC14</td>
<td>1717</td>
</tr>
<tr>
<td>Lamin B receptor</td>
<td>Homo sapiens</td>
<td>DHC14B</td>
<td>3930</td>
</tr>
<tr>
<td>7-dehydrocholesterol reductase</td>
<td>Mus musculus</td>
<td>Dhc7</td>
<td>13360</td>
</tr>
<tr>
<td>3-beta-hydroxysterol delta-24 reductase</td>
<td>Mus musculus</td>
<td>Dhc24</td>
<td>116932</td>
</tr>
</tbody>
</table>

Table 5.1 NCBI LocusLink human and mouse genes containing the root symbol ‘dhc’
Search limited to records created between 1995 and 2004. *LocusLink accession number

5.2.1 Collating human and mouse cytoplasmic dynein subunit synonyms

To identify the extent and variety of synonyms in use, literature searches were conducted using PubMed and the online databases MGI, NCBI (including LocusLink/Entrez Gene) and Ensembl (all searches were carried out between January and July 2004). For fast interrogation of online bioinformatics resources, the bioinformatics meta-search engine Bioinformatic-Harvester was also used. Table 5.2 lists aliases, map position and protein/DNA sequence accession data identified for each known mouse and human cytoplasmic dynein gene. More synonyms were observed in online databases than cited in the literature, which most likely reflects both the automated collection/annotation approach of database records and the lack of curation/review of these entries. The origins of each cytoplasmic dynein subunit synonym were traced to verify their validity ensuring for example, that the synonym did not simply represent a typographical error.

Many of the synonyms identified were named in accordance with their method of detection – projects responsible for originally identifying a gene or gene product, or projects detecting previously identified genes or gene product. For example (from Table 5.2), the ‘MGC’ prefixes are from the National Institutes of Health Mammalian Gene Collection (MGC) as part of an initiative to identify and sequence cDNA clones containing a full-length open reading frame for human, mouse and rat (Strausberg et al., 2002). ‘CGI’ prefixes are assigned by the Comparative Gene
Identification study, which use the *Caenorhabditis elegans* (*C. elegans*) proteome as an alignment template to assist in novel human gene identification from human EST nucleotide databases (Lai et al., 2000). 'Rik' suffixes are aliases assigned by the Riken Genomic Sciences Centre (http://genome.gsc.riken.jp).

5.2.2 The many names of cytoplasmic dynein 1 heavy chain 1

Amongst the cytoplasmic dynein subunits, *DYNC1HI* had the greatest number of synonyms published in the literature and online, with 15 aliases published for the human heavy chain and 9 for the mouse. Some alternative cytoplasmic dynein gene names were from large-scale gene and transcript identification efforts such as the partial *DYNC1HI* clone KIAA0325 and its mouse homolog 'mKIA00325', generated by the Kazusa cDNA project (Ohara et al., 1997). A small number of gene names were derived from DNA markers and cDNA clones used to identify the genes, for example, *DYNC2LI1* was named DKFZp564A033 after the cDNA sequence and clone of the same name and *DYNC1HI* referred to as Hp22 after a marker generated from its human cDNA sequence, as well as the rat-derived marker Rk3-8 and a cDNA clone named HL-3.

Two synonyms for mouse and human *DYNC1HI* that were investigated and ultimately rejected were ‘cell division cycle 22’ (CDC22) and ‘cell division cycle 23’ (CDC23), respectively. The mouse CDC22 alias was annotated on to the NCBI entry for *Dyn1h1* full-length mRNA (Acc. NM_030238; sequence revision history 21 December 2003) and partial mRNA (Acc. BC004751; sequence revision history 16 April 2003), and was also found at the MGI database (Figure 5.2). The human CDC23 alias was annotated on to the NCBI entry for *DYNC1HI* full-length mRNA (Acc. NM_001376; sequence revision history 23 December 2003) and partial mRNA mRNA (Acc. BC021297; sequence revision history 9 December 2003) (Figure 5.2). The validity of these aliases were investigated by first conducting literature searches using the search terms ‘CDC22’, ‘CDC23’ and ‘cell division cycle’ and identifying any papers citing these terms in association with cytoplasmic dynein. No citations of these terms were found or with the name ‘cytoplasmic dynein chain’ with the suffixes ‘22’ and ‘23’. This suggested that the error was most probably generated and propagated solely on the database.
Figure 5.2 Mouse and human DYN1H1 synonyms from the Entrez database at NCBI

Screen-shot of the NCBI sequence database entry for human DYN1H1 mRNA sequence, accession number NM_001376 (A) and mouse Dyn1h1 mRNA sequence, accession number NM_030238 (B). Synonyms listed within the entry are indicated by the red arrow and CDC22 and “cytoplasmic dynein complex” are boxed in red. Database accessed April 2004.

To test the validity of these synonyms and to identify if they had been assigned based on sequence homology to other known proteins, homology searches were conducted. Human and mouse CDC23 and CDC22 cDNAs and protein sequences were obtained by searching GenBank using the terms ‘CDC23’ and ‘CDC22’. The GenBank entries for CDC23 were listed as “cell division cycle protein homologous to Saccharomyces cerevisiae cdc23”, however no entries were seen for CDC22. The yeast cdc23 protein, a component of anaphase-promoting complex (APC), is essential for cell cycle progression through the G2/M transition and highly conserved in eukaryotic cells, was also included for comparison. The human and mouse dynein heavy chains were compared against the human and mouse cell division cycle subunits and against two different yeast cell division cycle subunits using BLAST. The Schizosaccharomyces pombe cdc22 protein, which shows no significant homology to S. cerevisiae cdc23 was included as a possible homolog of the mouse CDC22. Both cDNA and protein alignments were conducted using ClustalW. No significant similarity was seen between the dynein heavy chains and the cell division cycle proteins at both the nucleotide of protein level (data not shown).
<table>
<thead>
<tr>
<th>Cytoplasmic dynein gene family</th>
<th>Official gene name HUGO* (human)</th>
<th>Aliases</th>
<th>Location Human (Hsa) NCBIn* Mouse (Mmu) MGI*</th>
<th>Entrez Gene ID* (human and mouse)</th>
<th>mRNA (NCBI RefSeq accession numbers)</th>
<th>Protein (NCBI and SwissProt* accession numbers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytoplasmic dynein 1 heavy chain</td>
<td>Human</td>
<td>DNHC1 (Ahmad-Annuar et al., 2003)</td>
<td>Hsa14q32</td>
<td>1778</td>
<td>NM_001376</td>
<td>NP_001367 Q14204</td>
</tr>
<tr>
<td></td>
<td>Dynein 1 heavy chain DYNC1H1</td>
<td>Hsa14q32</td>
<td>1778</td>
<td>NM_001376</td>
<td>NP_001367 Q14204</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>Dnchc1 (Vaughan et al., 1996)</td>
<td>Mmu12 (55cM)</td>
<td>13424</td>
<td>103147</td>
<td>NP_064514 Q5JHU4</td>
</tr>
<tr>
<td>Cytoplasmic dynein 2 heavy chain</td>
<td>Human</td>
<td>DHC2 (Criswell et al., 1996; Vaisberg et al., 1996a)</td>
<td>Hsa11q21-q22.1</td>
<td>79659</td>
<td>XM_370652</td>
<td>XP_370652 O00432</td>
</tr>
<tr>
<td></td>
<td>Dnchc2 (Mouse Genome Informatics, 2004)</td>
<td>Mmu9 (1cM)</td>
<td>110350</td>
<td>107736</td>
<td>XM_358380</td>
<td>XP_358380 O08822</td>
</tr>
<tr>
<td>Cytoplasmic dynein 1 intermediate chain</td>
<td>Human</td>
<td>IC1 (Vaughan et al., 1995)</td>
<td>Hsa7q21.3-q22.1</td>
<td>1780</td>
<td>NM_004411</td>
<td>NP_004402 Q14576</td>
</tr>
<tr>
<td></td>
<td>Dncl1 (Mouse Genome Informatics, 2004)</td>
<td>Mmu6 (4cM)</td>
<td>13426</td>
<td>107743</td>
<td>NM_010063</td>
<td>NP_034193 O88485</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>Dnc1 (Mouse Genome Informatics, 2004)</td>
<td>Hsa2q31.1</td>
<td>1781</td>
<td>NM_001378</td>
<td>NP_001369 Q13409</td>
</tr>
<tr>
<td>Protein Name</td>
<td>Species</td>
<td>Gene Name</td>
<td>OMIM ID</td>
<td>RefSeq ID</td>
<td>Protein ID</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
<td>-----------</td>
<td>------------</td>
<td>-------------</td>
<td>------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>Dync1I2</td>
<td>Mouse</td>
<td>Dncc2</td>
<td>41cM</td>
<td>13427</td>
<td>NM_010064</td>
<td>NP_034194</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc2</td>
<td></td>
<td>107750</td>
<td></td>
<td>O68487</td>
</tr>
<tr>
<td>Dync1I1</td>
<td>Human</td>
<td>Dnclc1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>Dncc1</td>
<td>41cM</td>
<td>235661</td>
<td>NM_146229</td>
<td>NP_666341</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc2c1</td>
<td></td>
<td>2135610</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync1I2</td>
<td>Human</td>
<td>Dncc2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>Dncc2</td>
<td>41cM</td>
<td>110801 and see</td>
<td>XM_134573</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc2c2</td>
<td></td>
<td>234663</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync1I1</td>
<td>Mouse</td>
<td>Dncc1</td>
<td>41cM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync1I2</td>
<td>Human</td>
<td>Dncc2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync2I1</td>
<td>Mouse</td>
<td>Dncc1</td>
<td>41cM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync2I1</td>
<td>Human</td>
<td>Dncc2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynlt1</td>
<td>Mouse</td>
<td>Tctex1</td>
<td>41cM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tcxl</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynlt1</td>
<td>Human</td>
<td>Tctex1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tcxl</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynlt3</td>
<td>Mouse</td>
<td>Tctex1</td>
<td>41cM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tcxl</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynlt3</td>
<td>Human</td>
<td>Tctex1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tcxl</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync1I1</td>
<td>Mouse</td>
<td>Dncc1</td>
<td>41cM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>Dncc2</td>
<td>41cM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc2c2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytoplasmic dynein 1 light intermediate chain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync1I2</td>
<td>Human</td>
<td>Dncc2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync2I1</td>
<td>Mouse</td>
<td>Dncc1</td>
<td>41cM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync2I1</td>
<td>Human</td>
<td>Dncc2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytoplasmic dynein 2 light intermediate chain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync2I1</td>
<td>Mouse</td>
<td>Dncc1</td>
<td>41cM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync2I1</td>
<td>Human</td>
<td>Dncc2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytoplasmic dynein light chain Roadblock</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dync1I1</td>
<td>Mouse</td>
<td>Dncc1</td>
<td>41cM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dnc1c1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** The table includes information on protein names, species, gene names, OMIM IDs, RefSeq IDs, and protein IDs for various proteins related to the dynein family. The data comes from various sources, including Mouse Genome Informatics (MGI), NCBI, and literature references.
### Table 5.2 Human and mouse cytoplasmic dynein genes and map positions

<table>
<thead>
<tr>
<th>Cytoplasmic dynein light chain LC8</th>
<th>Mouse</th>
<th>Mmu2 H1</th>
<th>67068</th>
<th>1914318</th>
<th>NM_025947</th>
<th>NP_080223</th>
<th>Q86567</th>
</tr>
</thead>
<tbody>
<tr>
<td>DYNLRB2</td>
<td>Human</td>
<td>Hsa16q23.3</td>
<td>83557</td>
<td>NM_130897</td>
<td>NP_570967</td>
<td>Q8TF09</td>
<td></td>
</tr>
<tr>
<td>Dynlr2</td>
<td>Mouse</td>
<td>Mmu8 E1</td>
<td>75465</td>
<td>1922715</td>
<td>NM_029297</td>
<td>NP_063573</td>
<td>Q8DAJ5</td>
</tr>
</tbody>
</table>

### Dynamic light chain LC8

<table>
<thead>
<tr>
<th>Mouse</th>
<th>Mmu5 F</th>
<th>56455</th>
<th>1861457</th>
<th>NM_019682</th>
<th>NP_062656</th>
<th>Q9DF6F6</th>
</tr>
</thead>
<tbody>
<tr>
<td>DYNLL1</td>
<td>Human</td>
<td>Hsa12q24.3</td>
<td>1</td>
<td>8655</td>
<td>NM_003746</td>
<td>NP_003737</td>
</tr>
<tr>
<td></td>
<td>Mouse</td>
<td>Dncl1  (Mouse Genome Informatics, 2004)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynl1</td>
<td>Mouse</td>
<td>Mmu11 C</td>
<td>68097</td>
<td>1915347</td>
<td>NM_026556</td>
<td>NP_080832</td>
</tr>
</tbody>
</table>

### Dynamic light chain LC8

<table>
<thead>
<tr>
<th>Mouse</th>
<th>Mmu2 H1</th>
<th>67068</th>
<th>1914318</th>
<th>NM_025947</th>
<th>NP_080223</th>
<th>Q86567</th>
</tr>
</thead>
<tbody>
<tr>
<td>DYNLRB2</td>
<td>Human</td>
<td>Hsa16q23.3</td>
<td>83557</td>
<td>NM_130897</td>
<td>NP_570967</td>
<td>Q8TF09</td>
</tr>
<tr>
<td>Dynlr2</td>
<td>Mouse</td>
<td>Mmu8 E1</td>
<td>75465</td>
<td>1922715</td>
<td>NM_029297</td>
<td>NP_063573</td>
</tr>
</tbody>
</table>

Table 5.2 Human and mouse cytoplasmic dynein genes and map positions

Mapping prefixes are Hsa (Homo sapiens) and Mmu (Mus musculus) followed by the chromosomal localisation. All databases were accessed November 2005.

- HUGO: www.gene.ucl.ac.uk/nomenclature/
- MGI: www.informatics.jax.org/; mapping position shown in cM or cytogenetic band
- SwissProt: http://ca.expasy.org/sprot/
- * name as given by SwissProt

#### 5.2.3 Clarifying subunit mapping positions and identifying paralogs

Mapping positions of mouse and human cytoplasmic dynein subunits were clarified by identifying regions of sequence homology between the dynein subunit sequences, available from the sequence databases, and the mouse or human genome assembly. The elegance of this method was that it provided an assumption-free approach to identifying cognate genetic loci as it did not rely on a priori data from the literature or databases. However, published data obtained through PubMed
literature searches and from the online databases MGI, NCBI (including LocusLink and OMIM) and Ensembl, was also used, but only when corroborated by the sequence alignment results. Genome alignments were conducted using megaBLAST, with default settings and restricted to human and mouse genomes only. To designate the alignment of a query sequence against the genome as a cognate dynein subunit genetic locus, the following specified criteria had to be met:

1. the entire query sequence must align to discrete region of the genome
2. the alignment must demonstrate a rational genomic architecture – conservation of splice junctions, promoter sequences etc.
3. *in silico* splicing and translation of the putative cognate locus results in a known protein product of that subunit
4. supporting evidence from the literature, if available

The mapping results are shown in Table 5.2. This approach also provided a valuable secondary analysis by identifying potential unknown dynein subunit paralogs.

The dynamic nature of mouse and human genome evolution can generate paralogous sequences through mechanisms of duplication at the level of entire genomes, chromosomes, chromosomal segments and retrotransposition events (see for example, (Koonin, 2005; Sankoff, 2001; Taylor et al., 2004)). Each distinct mechanism of gene duplication and subsequent functional or non-functional sequence divergence generates a specific signature of evolution; from intron containing paralogs created by duplication to intron-less pseudogenes created by retrotransposition. However, as with many bioinformatics techniques, the limitation of this approach was anticipated to be the quality and completeness of the query sequences and the genome assembly. Therefore homology searches were also made against the nucleotide and protein sequence databases (GenBank) to identify potentially expressed and translated paralogs. Although low homology sequence was seen frequently, no functional loci were identified (data not shown) except for DYNLL1. Although the DYNLL1 locus appeared functional, no protein product was found. In the interest of brevity two examples of this work are given here, which are representative of the full work undertaken.

5.2.3.1 Example 1: Cytoplasmic dynein light chain 1

The cytoplasmic dynein light chain 1 (DYNLL1) gene illustrates many of the methods used to clarify dynein subunit information. DYNLL1 possessed two different loci, the cognate locus 12q24.21 and the incorrect locus at 14q24 which was reproduced and propagated on genome browsers. The origin of the discrepancy most likely stemmed from the work of Dick and colleagues.

* Paralogs are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.
who took N and C terminal sequences of *Drosophila* and *C. elegans* and designed primers by reverse translation and used PCR to design 300bp probes and isolated 2 cDNAs (Dick *et al.*, 1996a). FISH probes constructed with this sequence mapped to 14q24. This error was also seen at the Human Genome Nomenclature Committee (HGNC) website at the Galton Laboratory, UCL, London, UK (www.gene.ucl.ac.uk/nomenclature) during a search of all loci containing the name “cytoplasmic dynein”. The Galton Laboratory gave *DYNLL1* a 14q24 locus and two additional pseudogene loci, *DNCL1P1* and *DNCL1P2* on chromosome 14. In addition Online Mendelian Inheritance in Man (OMIM) website, used to collate information about human genetic diseases reported a 14q24 location but LocusLink (now EntrezGene) gave a 12q24.23. To resolve these differences the full-length *DYNLL1* cDNA was aligned against the human genome using megaBLAST (Table 5.3).
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Hsa</th>
<th>Contig</th>
<th><strong>DYNLL1</strong></th>
<th>Contig Position</th>
<th><strong>% identity</strong></th>
<th><strong>ORF?</strong></th>
<th><strong>Protein</strong></th>
<th><strong>% identity</strong></th>
<th><strong>Ensembl Gene</strong></th>
<th><strong>NCBI Gene</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1p36.13</td>
<td>NT_077921</td>
<td>2 – 371</td>
<td>456983 – 457323</td>
<td>91</td>
<td>Yes</td>
<td>89</td>
<td>87</td>
<td>DYLL_HUMAN</td>
<td>ENSG00000124854</td>
</tr>
<tr>
<td>1</td>
<td>1p36.11</td>
<td>NT_037485</td>
<td>307 – 550</td>
<td>2790937 – 2790695</td>
<td>93</td>
<td>No</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>3q25.3</td>
<td>NT_005612</td>
<td>39 – 94</td>
<td>60870198 – 60870252</td>
<td>83</td>
<td>No</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>4</td>
<td>4q23</td>
<td>NT_016354</td>
<td>23 – 546</td>
<td>2545035 – 25457518</td>
<td>83</td>
<td>No</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>7</td>
<td>7q21.1</td>
<td>NT_007933</td>
<td>1 – 66</td>
<td>10245012 – 10244953</td>
<td>93</td>
<td>No</td>
<td>—</td>
<td>—</td>
<td>Hs_7_c1269</td>
<td>OTTHUMG00007007517</td>
</tr>
<tr>
<td></td>
<td>7q21.1</td>
<td>NT_007933</td>
<td>79 – 370</td>
<td>1024494 – 10244659</td>
<td>89</td>
<td>No</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>7q21.1</td>
<td>NT_007933</td>
<td>377 – 640</td>
<td>10245024 – 10245285</td>
<td>93</td>
<td>No</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>12</td>
<td>12q24.13</td>
<td>NT_009775</td>
<td>87 – 456</td>
<td>4746361 – 4745993</td>
<td>89</td>
<td>No</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>12q24.13</td>
<td>NT_009775</td>
<td>472 – 578</td>
<td>4745993 – 4745887</td>
<td>89</td>
<td>No</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>12q24.13</td>
<td>NT_009775</td>
<td>560 – 640</td>
<td>4745587 – 4745506</td>
<td>82</td>
<td>Yes</td>
<td>89</td>
<td>100</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>14*</td>
<td>14q31.1</td>
<td>NT_026437</td>
<td>1 – 640</td>
<td>61633039 – 61632401</td>
<td>94</td>
<td>Yes</td>
<td>89</td>
<td>91</td>
<td>K14_NN_963_1</td>
<td>ENSG00000183992 and OTTG00000002869</td>
</tr>
<tr>
<td>14</td>
<td>14q31.1</td>
<td>NT_026437</td>
<td>38 – 635</td>
<td>61799299 – 61799896</td>
<td>94</td>
<td>No</td>
<td>—</td>
<td>—</td>
<td>K14_NN_968_1</td>
<td>OTTG00000002889</td>
</tr>
</tbody>
</table>

Table 5.3 **DYNLL1** megaBLAST alignments against the human genome

An intact open reading frame (ORF) was seen on three chromosomes, with three potentially translated proteins 89 amino acids (aa) in length. Chromosome 12q24.31 was identified as the cognate genetic locus and loci on chromosome 1 and 14 as untranslated pseudogenes. Mapping prefix Hsa refers to Homo sapiens chromosomal location and all alignments are given by nucleotide (nt).
Alignments to chromosome 12q24.31 had the highest percentage identity, spanning the entire length of the cDNA sequence. To ensure this megaBLAST alignment was correct, the DYNLL1 cDNA sequence was aligned against chromosome 12 contig only (Acc: NT_009775) using BLAST2Seq and the corresponding alignments checked for the presence of conserved sequence features indicative of an expressed gene. The sequence alignment was identical to that seen using megaBLAST (data not shown). The entire cDNA length was represented against the chromosome 12 contig in three distinct exonic regions and several conserved sequence motifs, such as Kozak consensus sequence, splice junctions and polyadenylation signals were identified which supported this chromosomal locus (Figure 5.3). The exonic sequences were artificially spliced and translated using TRANSLATE and resulting protein compared to DYNLL1 (Acc: NP_003737) using BLAST2seq - the two sequences were identical (Figure 5.4).

**Figure 5.3 Schematic representation of the putative DYNLL locus at 12q24.31**

Exonic sequence shown in blue font and intronic in grey font. Translational start codon (methionine) is shown in red at +1 within Kozak conserved signal (grey box). Conserved splice junctions indicated within green lines, TAA stop codon and polyadenylation signal highlighted in yellow.

**Figure 5.4 In silico translation of artificially spliced DYNLL1 alignments**

DYNLL1 alignments from various chromosomes were spliced and translated in silico and aligned to the known DYNLL1 protein NM_003746 using ClustalW. The alignment at chromosome 12q24.31 showed complete homology to DYNLL1 and all other loci showed evidence of sequence degeneration associated with pseudogenes.

This method of analysis was carried out for all DYNLL1 alignments shown in Table 5.3. Several loci including those on chromosomes 1, 7 and 14 were characteristic of processed pseudogenes; lacking introns, encompassing a poly(T) tail and comprised of degenerate sequence (i.e. possessing
stop codons in the ORF) (Vanin, 1985). The “processed” nature of these loci implied duplication of the cognate locus by retrotransposition. This hypothesis was confirmed by analysing the sequence context of each suspected locus using RepeatMasker (www.repeatmasker.org/) – all loci except chromosome 12q24.31, were seen to possess hallmark flanking sequences of non-long terminal repeats (LTR) Class I retrotransposable elements: chromosome 1 loci possessed short interspersed nuclear elements (SINEs) and chromosome 12 possessed long interspersed nuclear elements (LINEs). Corresponding sequences were seen in the mouse in regions syntenic to those in humans suggesting that duplication of the cognate locus occurred in the common ancestor of both human and mouse species. Locus 12q24.31 was proposed identified as the cognate locus, which was supported by the mapping of the mouse Dynll1 to a region syntenic to 12q24.31 on Mmu5.

To ensure that no potential paralogs had been missed the cDNA and protein sequences of DYNLL1 were used to search the sequence databases using nucleotide and protein BLAST. A match to “Homo sapiens lung cancer oncogene 5” (HLC5) was found with 99% identity over 85% of the length of DYNLL1 cDNA. In silico translation of the antisense strand of HLC5 produced a 100% match to the full length DYNLL1. It is likely that the HLC5 data, which remains unpublished and is based on a single observation from a study of mRNA species in cancerous tissue, is incorrect and most likely represents a detection of DYNLL1.

5.2.3.2 Example 2: Cytoplasmic dynein 1 heavy chain 1

A search for sequences paralogous to DYNC1H1, using BLAST to query the cDNA sequence against the nucleotide database, did not yield significant results however a BLAST search of the protein database yielded a sequence with 27% identity to DYNC1H1. The sequence (Acc. XP_085578, XM_085578) was the predicted product of a hypothetical gene, FLJ46675, located at chromosome 17p13.1. The existence of a second heavy chain species DYNC2H1 was already known and so this result suggested a possible third heavy chain existed. By aligning the sequences for DYNC1H1, DYNC2H1, FLJ46675 and DNAH9 used as an outgroup, it was shown that the sequence was more likely to be axonemal (33% identity, 51% similarity to DYNC1H1) than cytoplasmic (Figure 5.5).

Figure 5.5 Cladogram of the relationships between human dynein heavy chains and the hypothesised heavy chain FLJ46675
5.2.4 Cytoplasmic dynein subunit accession numbers

Cytoplasmic dynein subunit nucleotide and protein accession numbers were obtained from NCBI RefSeq. The majority of sequences were found through links from the central gene-based data resource LocusLink. Human DYNC2H1 sequences XM_370652 and XP_370652 were predicted by analysis of genomic sequence (NT_033899) using the NCBI gene prediction method GNOMON, supported by mRNA and EST evidence. Mouse Dynlli2 sequences XM_134573 and XP_134573 were also predicted on genomic sequence (NT_078575) using a similar method.

5.3 Identifying cytoplasmic dynein orthologs

With the availability of sequence data from different organisms, cytoplasmic dynein orthologs were identified by searching the GenBank non-redundant protein database, with the human dynein proteins using PSI-BLAST with default parameters and the BLOSUM-62 substitution matrix. Searches of pufferfish sequence Takifugu rubripes were performed by using human protein sequence against the translated Fugu Genome nucleotide database (version release 2.0; www.ensembl.org/Fugu_rubripes). Where possible, outgroup sequences were identified from homologs in Chlamydomonas reinhardtii as these axonemal dynein proteins have been extensively studied in this species and many of the subunits in higher-order species are known to be related. The accession numbers of all protein sequences identified are given in the Appendix. Phylogenetic analyses were conducted using these sequences by Dr James Cotton at the Natural History Museum, London, UK; using the methods outlined in Chapter 2.3.18. For the Markov-chain approach, the consensus of three Markov-chain analyses was used to ensure that the algorithm had converged to a stationary state. For one of the cytoplasmic dynein heavy chain gene family, the three chains had reached different likelihood values after 1,000,000 generations, suggesting failure to converge. Running another three independent chains resulted in five out of six chains agreeing on the likelihood values, suggesting that only one chain had not converged properly. In all cases, the phylogeny presented is the majority rule consensus of the posterior sample of tree topologies from all three Markov chains. A positive control in the identification of homologous sequences to construct phylogenetic trees was the multiple independent detection of subunits: e.g. searching for human DYNC1H1 paralogs identified DYNC2H1 with an alignment score of 30%.

5.3.1 Cytoplasmic dynein heavy chain gene family (DYNC1H1, DYNC2H1)

Figure 5.6A illustrates the phylogenetic relationships amongst the sequences identified as orthologous to DYNC1H1 and DYNC2H1 from various organisms. The existence of several
axonemal dynein heavy chain proteins in the *Chlamydomonas* proteome made identifying an outgroup sequence difficult, however the outer arm heavy chain (ODA11) was arbitrarily chosen (all *C. reinhardtii* heavy chains had comparable sequence similarity to DYNC1H1). A partial fragment of a suggested third dynein heavy chain identified in the literature (Vaisberg *et al.*, 1996a) was included in the analysis (fragment was termed “DNAH12frag”). Only partial sequence (336aa) of the mouse DYNC2H1 (XP_35830) protein was available in the GenBank database. Adding this partial sequence to the analysis resulted in spurious clustering and therefore, an extended putative sequence was required. The extended sequence was obtained by aligning human and rat sequences XP_370652 and NP_075413 respectively, against the translated mouse genome (Build 32) using TBLASTN. Incomplete mouse genomic assembly at the DYNC2H1 locus yielded a truncated sequence 3455 amino acids in length, 85% the length of human DYNC2H1.

The heavy chain sequences fell into two distinct clades and the relationships within each clade were generally consistent with known evolutionary distances between the organisms shown*. The deep branching of the dynein 1 and dynein 2 clades, and the presence of myriad species on both branches, suggests that the origin of these two dynein heavy chain proteins is ancient, possibly predating the divergence of *C. reinhardtii* from a primordial species. The phylogeny complemented and extended, previous phylogenetic analyses of the heavy chain proteins (Porter *et al.*, 1999; Vaisberg *et al.*, 1996b). Importantly, the analysis indicated that the partial human sequence, suspected to be a third cytoplasmic heavy chain, was unlikely to be a cytoplasmic dynein; it possessed good homology to DNAH12 (Acc:AAB09729) and was termed “DNAH12frag”.

5.3.2 Cytoplasmic dynein intermediate chain gene family (DYNC1I1, DYNC1I2)

Figure 5.6B shows the dynein intermediate chain protein phylogeny. The protein sequence data demonstrated an evolutionary distant relationship between axonemal and cytoplasmic dynein intermediate chains; for example, mammalian rat DYNC1I1 (NP_062107) has 48% similarity to the *Chlamydomonas* axonemal outer arm dynein intermediate chain encoded by the ODA6 locus and therefore ODA6 was used as an outgroup (Mitchell *et al.*, 1991; Paschal *et al.*, 1992a). The intermediate chain sequences fell into two clades, intermediate chain 1 and 2, which were comprised of vertebrate species only.

An alternative placement of a *Takifugu* sequence, as a member of the intermediate chain 1 clade was almost as well-supported by the data as the placement shown in Figure 5.6B (49% bootstrap

* Relationships amongst dynein sequences of different species do not necessarily reflect the evolutionary relationships amongst species as a gene tree does not always reflect a species tree; see (Tajima, 1983) and (Pamilo *et al.*, 1988) for further details.
support against 51% support). In view of this and with all non-vertebrate species falling outside these clades, the data suggest a recent evolutionary origin for an intermediate chain 1 and 2 gene split, perhaps as part of a ‘2R’ event of genome duplication (see Wolfe, 2001) for review. The absence of an amphibian intermediate chain 1 protein may have been due to the paucity of Xenopus laevis sequences in the GenBank sequence database at the time of analysis.

5.3.3 Cytoplasmic dynein light intermediate chain family (DYNC1LI1, DYNC1LI2, DYNC2LI1)

Figure 5.6C illustrates the phylogenetic relationships amongst the dynein light intermediate chain protein sequences from various organisms. The light intermediate chains separated into two deep clades, mirroring the dynein 1 and 2 heavy chain tree. The dynein 1 clade was further split the two intermediate chain components of cytoplasmic dynein 1 complexes, DYNC1LI1 and DYNC1LI2 (Entrez GeneID: 51143 and 1783, respectively) which are more closely related to each other than to the cytoplasmic dynein 2 light intermediate chain, DYNC2LI1. DYNC2LI1 homologs were only identified in species possessing DYNC2H1, emphasising the distinct cellular identities and roles of these separate dynein complexes (Figure 5.1B). C elegans appeared to possess one light intermediate chain (dli-1) (Entrez GeneID: 178260) for cytoplasmic dynein (DYNC1H1-based complexes), and one (xbx-1) (Entrez GeneID: 184080) for cytoplasmic dynein 2 (DYNC2H1-based complexes) (Schafer et al., 2003).

5.3.4 Cytoplasmic dynein light chain Tctex1-family (DYNLT1, DYNLT3)

Cytoplasmic dynein light chain Tctex1-family is one of three known dynein light chain gene families that are components of cytoplasmic dynein 1: (1) the t-complex associated family (DYNLT1, DYNLT3), (2) the Roadblock family (DYNLRB1, DYNLRB2), and (3) the LC8 family (DYNLL1, DYNLL2). These gene families are named according to their original discovery, through the effect of mutations in mouse (t-complex associated, Tctex1) and Drosophila (Roadblock) or according to the size of the protein in Chlamydomonas (LC8). The dynein light chain Tctex1-family protein phylogeny is shown in Figure 5.8A, and shows distinct clades for DYNLT1-like and DYNLT3-like sequences.
Figure 5.6 Protein-based phylogenies of the cytoplasmic dynein heavy chain, intermediate chain and light intermediate chain families

Species names for heavy chain (A), intermediate chain (B) and light intermediate chain (C) phylogenies are given with NCBI/GenBank gene/protein names (sequence accession numbers are given in Appendix). Orthologous human, mouse, and rat gene names use the revised systematized consensus nomenclature (Pfister et al., 2005b). Named clades are indicated in the right margins. Bayesian and maximum-likelihood bootstrap values are shown as percentages adjacent to branch points. *bootstraps below 50%. ^bootstraps at 100%. Scale-bar represents evolutionary distance (estimated numbers of amino-acid substitutions per site).
5.3.5 Cytoplasmic dynein light chain Roadblock family (DYNLRB1, DYNLRB2)

Figure 5.8B shows the phylogenetic relationships amongst the dynein light chain Roadblock protein family in various organisms. The Roadblock sequences appeared well-conserved from *T. rubripes* to humans, with 96% of pairwise sequence comparisons amongst all sequences demonstrating sequence identity >50% (data not shown). However, the high alignment scores may have been produced through the limited number of sequences used and the paucity of more distant species. The conservation of Roadblock sequences (i.e. Roadblock domains) within the coding regions of genes of mammals and other species (Bowman *et al.*, 1999; Koonin *et al.*, 2000) highlights a necessary functional constraint. However, these genes are not thought to be cytoplasmic dyneins; e.g. *MAPBPIP* (Entrez GeneID: 28956) in human and mouse functions mainly in the endosome/lysosome pathway (Lunin *et al.*, 2004).

5.3.6 Cytoplasmic dynein light chain (LC8) 1, DYNLL1

Figure 5.8C shows the phylogenetic relationships amongst the dynein light chain LC8 family protein sequences from various organisms. The phylogeny shows that the mammalian LC8 light chain family falls into two distinct clades containing DYNLL1- and DYNLL2-like genes. DYNLL is highly conserved from algae to humans, which is illustrated by: (i) a high degree of sequence similarity between distant species - 92% similarity exists between *Drosophila* Cdlc1 (NP_525075) and the 8kDa flagellar outer arm dynein light chain from *Chlamydomonas* (Q39580), and 91% between human and *C. elegans* light chain 1 (NP_498422) and (ii) a flat tree with short branch lengths between taxa. In total, 67% of pairwise sequence alignments illustrated >50% sequence identity which, although fewer that seen for the Roadblock proteins, includes sequences from several more distant species and therefore accounts for a lower number of high identity alignments (Figure 5.7). Taken together these data suggest that *DYNLL1* and *DYNLL2* are both likely to be under strong purifying selection to conserve protein function.

Figure 5.7 Cytoplasmic dynein light chain (LC8) family protein alignments

67% of all alignments were >50% sequence identity. Protein sequences are presented in the same order as the tree in Figure 5.8C. Alignments were conducted using ClustalW.
Figure 5.8 Protein-based phylogenies of the cytoplasmic dynein light chain family

Species names for Tctex1 (A), Roadblock (B) and LC8 (C) dynein light chain phylogenies are given with NCBI/GenBank gene/protein names (sequence accession numbers are given in Appendix). Orthologous human, mouse, and rat gene names use the revised systematized consensus nomenclature (Pfister et al., 2005b). Named clades are indicated in the right margins. Bayesian and maximum-likelihood bootstrap values are shown as percentages adjacent to branch points. *bootstraps below 50%. #bootstraps at 100%. Scale-bar represents evolutionary distance (estimated numbers of amino-acid substitutions per site).

5.4 Discussion

In preparing to undertake a systematic analysis of the components of the cytoplasmic dynein pathway for a large scale association study, confusion and errors regarding the nomenclature, mapping positions and sequence accession numbers were observed in the literature and databases.
The accuracy of electronic data should be a concern for all research communities as it is an increasingly common problem; it is estimated, for example, that ~15% of annotation in GenBank contains errors, many of which have been found to be propagated in peer-reviewed literature (Pennisi, 1999). The dynein community is not exceptional in this respect and as the data presented in this chapter catalogue, several errors were identified in cytoplasmic dynein subunit annotation. Where mistakes were identified in databases and other online resources, the hosting institute was contacted to review their data and amend it where necessary.

Table 5.2 catalogues the key data for mouse and human dynein subunits including subunit synonyms, mapping positions and mRNA and protein accession numbers, which has since been published as a valuable resource for the cytoplasmic dynein community (Pfister et al., 2005b). The investigation of novel mouse and human paralogs was a valuable secondary analysis in confirming the cognate genetic locus of each dynein subunit. In both species, for many of the subunits tested, alignments of expressed sequences were seen against various regions of the genome however, these regions were identified as either: (i) genes and proteins containing homologous domains to the dynein subunits tested, (ii) pseudogenes of whole or partially duplicated loci or (iii) incorrect genome assembly. DYNLL1 possessed a large number of homologous sites which were determined to be pseudogenes created by duplication events. All loci identified except chromosome 1, have since been included on the human pseudogene database (www.pseudogenes.org). The "intronless" nature of the pseudogenes and the SINE and LINE flanking repeat sequences were indicative of duplication by a non-LTR Class 1 transposable element, which moves by reverse transcription of an RNA intermediate (Silva et al., 2004). The presence of identical transposed sequences in the mouse provides an indication of the age of the DYNLL1 duplication – predating the mouse and humans species divergence ~75 million years ago (Waterston et al., 2002).

5.4.1 Cytoplasmic dynein nomenclature

Collation of the myriad names used to describe the cytoplasmic dynein subunits illustrated the extent of confusion in the cytoplasmic dynein field. All subunit synonyms were investigated to establish their accuracy and many, such as the "CDC23" synonym were proven to be incorrect (Chapter 5.2.2). The data presented in this chapter, including the phylogenetic analyses, were influential in facilitating a revision of the human and mouse dynein nomenclature to a standardised format. Standardising nomenclature of large gene families has been undertaken for many different proteins; see for example (Marenholz et al., 2004; Miki et al., 2005; Rodgers et al., 2007; Wilson et al., 2006) and have in common the use of phylogenetic analyses to establish inclusion/exclusion criteria for family members. The investigation of paralogous and orthologous cytoplasmic dynein
sequences allowed a complete investigation of known and unknown dynein subunit family members, which could then be named accordingly.

5.4.1.1 A new nomenclature system for the cytoplasmic dynein subunits

Revision of the mammalian cytoplasmic dynein nomenclature system was undertaken by Pfister and colleagues (Pfister et al., 2005a) and was approved by the Human Genome Organisation Nomenclature Committee (HGNC, 2004) and the International Committee on Standardised Nomenclature for Mice (Maltais et al., 1997). Phylogenetic analyses of the dynein proteins substantiated the existence of two distinct dynein complexes, cytoplasmic dynein 1 and cytoplasmic dynein 2 and the exclusive occurrence of DYNC2H1 and its binding protein DYNC2LI1. In accordance with HGNC policy, Pfister and colleagues designated each unique cytoplasmic dynein subunit with the root “DYNC” (for dynein, cytoplasmic), followed by the specific dynein complex subtype 1 or 2 (e.g. cytoplasmic dynein 2 was designated DYNC2). The shared light chains were rooted with “DYN” and other subunits were designated with a letter(s) for the size of the polypeptides: H for the Heavy Chain, I for the Intermediate chain, LI for the Light Intermediate Chain, and L for Light Chain. Additional letters (T, RB and L) were used to distinguish the three distinct light chain families and individual members of the gene families were assigned numbers. The revised nomenclature has been used throughout this thesis.

5.4.2 Changing databases and future studies

As with many in silico approaches, the results of an investigation are often correlated to the quality and completeness of the raw data used and as such, the data presented in this chapter reflect the state of the sequence and genome databases at the time they were accessed. Identifying paralogous dynein sequences in the mouse and human benefited from the availability of complete genome, nucleotide and protein data, however, the identification of orthologous sequences in newly or partially sequenced genomes may have suffered from incomplete sequence ascertainment. Therefore, the cytoplasmic dynein subunit phylogeny is currently accurate but with the publication of genome sequences from additional species, the phylogeny may be refined in the future and weak nodes provided with greater support.

5.4.2.1 Priority candidate genes: DYNLL family and DYNLRB family

The cytoplasmic dynein light chain LC8 phylogenetic tree suggests that human DYNLL1 and DYNLL2 may both be promising candidates for a future ALS candidate gene association study. The flat branch network illustrated the extent to which proteins of this family are conserved between distantly related species that diverged ~1717 million years ago (Nei et al., 2001). More closely related species such as human, mouse, rat, pig, and cow show extraordinary sequence conservation;
between these species the amino acid sequences of both DYNLL1 and DYNLL2 are identical (Wilson et al., 2001). The observed DYNLL amino acid sequence homogeneity between species can be explained by the action of purifying selection which serves to eliminate those mutations that have a deleterious effect on protein function. Purifying selection is generally a more pervasive mechanism of selection than other forms of directional selection (Hughes et al., 2003; Kimura et al., 1974) and has been shown to maintains stringent interspecies sequence homogeneity in proteins with essential functions, such as H4 histones (Piontkivska et al., 2002). The conservation of the DYNLL protein family raises questions as to what essential function these proteins play and whether perturbation of this function may yield a neurodegenerative phenotype. As yet, no neuronal phenotypes are known.

The significance of DYNLL to neuronal function (and conversely, degeneration) is not known. However, DYNLL is known to be an integral part of brain cytoplasmic dynein, although it is ubiquitously expressed in all cells of the body and large amounts of brain DYNLL are not associated with the dynein complex (King et al., 1996a). The functional role of these LC8 light chains is varied: DYNLL is a substrate of a p21-activating kinase (Vadlamudi et al., 2004), a component of the actin-based motor myosin V (Naisbitt et al., 2000b) and a component of the nNOS complex (Grissom et al., 2002). DYNLL mutants display various developmental and fertility phenotypes (reviewed in (Pfister et al., 2005b) and total loss-of-function Drosophila mutants are embryonic lethal (Dick et al., 1996a).

*Drosophila* Roadblock mutants also result in a number of phenotypes including the accumulation of axonal cargoes, mitotic defects, female sterility, and larval/pupal lethality (Bowman et al., 1999). The Roadblock proteins also demonstrated high sequence similarity of interspecies pairwise alignments which suggests that these proteins have also been conserved. Although, more distantly related species were not available for comparison during the original analyses (which may have inflated the apparent level of protein conservation), several new protein sequences have recently become available and continue to show high sequence identity: *Dictyostelium discoideum* hypothetical protein (XP_637964) displays 62% sequence similarity to human DYLRB1 (NP_054902). The Roadblock proteins are promising candidates for neurodegeneration with mutants affecting neuroblast proliferation, reducing dendritic complexity and causing defects in axonal transport (Reuter et al., 2003).

The human DYNLL and DYNLRB genes should therefore make valuable candidates for future ALS and other motor neuron disease association studies.
6 Identifying kuru susceptibility loci

6.1 Introduction

Mapping disease susceptibility loci by identifying genes affected by natural selection imposed by the disease, is a novel approach to mapping complex traits. In this chapter such evolutionary analyses of the prion gene (PRNP) is undertaken as a paradigm for future studies of prion disease candidate loci and whole-genome analyses. This chapter begins with an updated analysis of the role of codon 129 as both a susceptibility locus and as a protective factor for kuru, in the kuru-exposed linguistic groups of the Eastern Highlands of Papua New Guinea. The resulting genetic impact of the kuru epidemic is investigated in the surviving populations, with codon 129 genotypes analysed stratified by age and sex. A large sample of elderly South Fore females and slightly younger South Fore males show significant Hardy-Weinberg disequilibrium at codon 129, driven by an excess of heterozygosity, an observation consistent with their participation in mortuary feasts. Examination of codon 129 valine allele frequency worldwide illustrates a significant West-East reducing cline, against which the PNG valine frequency contrasts starkly. An update and refinement of individual village exposure to kuru by establishment of an Exposure Index, allowed a more subtle analysis of valine allele frequency. A significant increasing cline was seen from non-Highland to Highland populations however no local cline within the Eastern Highlands was seen. Similarly, no increase in LD between microsatellites flanking PRNP is seen. This chapter concludes with a feasibility study on the whole-genome analysis of a small number of elderly Fore women and whole genome amplified kuru samples using 250,000 SNP arrays. Although the amplified kuru samples genotyped poorly, the remaining 7 samples from elderly Fore women are used to determine analysis criteria and examine cursory LD data.

6.1.1 Kuru and the evolution of human PRNP

The evolution of human PRNP is a contentious topic which has been investigated by several authors. In 2003, the work of Mead and colleagues investigated the PRNP locus in a number of worldwide populations including the people of the Eastern Highlands of Papua New Guinea who experienced the human prion disease epidemic of kuru (Mead et al., 2003). Mead identified that the prion locus had undergone kuru imposed balancing selection in the Fore people of the Eastern Highlands by identifying a significant excess of heterozygotes in the surviving Fore population; a significant excess of human coding polymorphisms as tested by a McDonald-Kreitman test against both Old and New World monkeys; a significant positive Tajima’s D for polymorphisms within 4.7kb sequence; a bifurcating haplotype genealogy characterised by two long deep branches with highly divergent clades (which were associated with each codon 129 allele) and extended linkage
disequilibrium with microsatellites flanking ~30 kb either side of \textit{PRNP} with reduced diversity of alleles. In addition, several of these features were shared by other worldwide populations prompting the suggestion that the prion gene had undergone balancing selection throughout human history and speculation that (as one hypothesis) the selective pressure was provided by widespread cannibalistic practices in prehistoric humans. Several authors have challenged Mead's proposition with regards to balancing selection at \textit{PRNP}, imposed by cannibalism-associated prion-like diseases during the evolution of modern humans, citing polymorphism ascertainment as a bias introducing factor (Brookfield, 2003; Soldevila \textit{et al.}, 2003; Soldevila \textit{et al.}, 2006).

Despite this, there is general agreement that the data presented for the Fore does illustrate the action of natural selection in this population and that the basis for this selection pressure is the marked survival advantage of \textit{PRNP} codon 129 heterozygotes (Cervenakova \textit{et al.}, 1998; Mead \textit{et al.}, 2003). The strength of balancing selection in the Fore, within a single generation, has been identified as the strongest documented to-date in any human population (Hedrick, 2003) and surpasses the commonly quoted example of malaria on the \textit{C} and \textit{S} alleles of \textit{\beta}-haemoglobin (Hedrick, 2004). Such great selection pressures could rapidly establish/eliminate mutations within the population and therefore by examining the evolutionary signatures of selection, kuru may provide a unique resource for the investigation of disease modifying loci in prion disease and possibly other neurodegenerative diseases. Therefore, the main aim of this chapter was not to resolve the debate regarding worldwide balancing selection but to investigate further the genetic signature of balancing selection on the Fore, how this can be utilised to identify other mediators of human prion disease and the applicability to this method of analysis to other neurodegenerative diseases. Further, this chapter aims to:

to update previous assessments of the role of codon 129 in kuru susceptibility and protection, by analysing larger data sets.

(i) to identify a correlation between the extent of kuru experienced by Eastern Highland linguistic groups and the pattern of genetic variation at \textit{PRNP} (i.e. identify signatures of selection) as a precedent for performing larger candidate gene or whole-genome study

(ii) to establish a resource to identify novel susceptibility loci. By identifying a correlation of genetic variation at \textit{PRNP} proportional to exposure to kuru, a panel of samples from different linguistic groups with differing exposure to kuru could be used to identify/support additional susceptibility loci

(iii) to pilot a whole genome hyper-case-control association study between kuru and multiple exposure samples.
6.2 Kuru susceptibility is mediated by PRNP codon 129 genotype

Cervenáková and colleagues originally illustrated the influence of codon 129 genotype on kuru susceptibility by examining genotype data from 92 kuru patients (Cervenakova et al., 1998). They correlated homozygosity at codon 129 with an earlier age of kuru onset and shorter duration of illness, compared to heterozygosity. The MRC Prion Unit’s collection of 161 kuru samples provided an opportunity to re-examine the effect of codon 129 on kuru susceptibility with a larger sample size and for a sex-specific genotype distribution.

Kuru samples were obtained as either DNA or blood/sera from which DNA was extracted by the MRC Prion Unit Human Genetics Group, using QIAmp mini DNA extraction kits and following safety procedures relevant to handling infectious material. PRNP codon 129 genotyping was performed by allelic discrimination on all samples (described in Materials and Methods Chapter 2.3.13) using a FAM-tagged PRNP codon 129M probe, VIC-tagged PRNP codon 129V probe and flanking forward and reverse primers taqmvf2 and taqmvr respectively (see Primer Table in Appendix). Data on age at sample collection and sex were obtained for all samples however, 18 samples lacked this information and were removed from further analysis. Deviations from HWE were calculated using PLINK.

![Figure 6.1 Distribution of PRNP codon 129 genotypes by age in 147 kuru cases](image)

The proportion of kuru MV heterozygotes increases with age of onset.

Genotype proportions in early onset kuru cases were significantly distorted away from those estimated by HWE (Figure 6.1) with excess homozygosity ($P_{exact} = 0.0003$, in patients ≤10 years old). Concomitant increase in heterozygosity was seen with age, with a significant excess of heterozygotes in the age groups 31 to 40 years ($P_{exact} = 0.003$) and 41 to 50 years ($P_{exact} = 0.006$). The 11 to 20 and 21 to 30 years age groups were both in HWE ($P_{exact} = 0.70$ and 0.087 respectively), which represented the intersection of decreasing homozygosity and increasing heterozygosity with age and therefore the genotype proportions appeared normal. Comparison by Pearson $\chi^2$ test of the
most (<20 years old) and least susceptible (>30 years old) kuru groups was highly significantly different ($P<0.001$, 2 d.f.). These data illustrate the protective effect of heterozygosity at codon 129 which results in a later age of kuru onset.

When partitioned and analysed by sex (Figure 6.2) the data illustrated and was consistent with several epidemiological aspects of kuru including: (i) the extent to which kuru largely affected females as compared with males and (ii) the peak kuru incidence in males occurred between 11 to 20 years old, which supports the participation of only boys less than 8 years old at mortuary feasts (given a mean kuru incubation period of ~12 years).

![Figure 6.2](image.png)

**Figure 6.2** Sex specific distribution of PRNP codon 129 genotypes by age at collection in 146 kuru cases

*Top.* Female kuru cases. *Bottom.* Male kuru cases.

Detailed inspection of the earliest onset males and females cases ($\leq 10$ years old), identified an excess of female valine homozygotes and methionine homozygotes in males (two-tailed Fisher’s exact, $P=0.015$). Lee and colleagues also noted this result in young Fore males $\leq 20$ years and suggested that it reflected an increased susceptibility of methionine homozygotes to kuru (Lee *et al.*, 2001a) although their analysis may have been affected by small sample size. Taking this into account, the conflicting absence of MM and VV genotypes in males and females most likely reflects a sampling artefact.

### 6.3 PRNP codon 129 heterozygosity mediates protection against kuru

Expanding on the work of Mead and colleagues who examined codon 129 status in 30 elderly Fore women, blood was obtained from an additional 104 women (Mead, 2002). All women were aged 50
years or over in the year 2000, had a history of multiple exposures to kuru at mortuary feasts and resided in a village documented to have a moderate to high level of exposure. The 134 samples were obtained from the South Fore (n=68), North Fore (n=36), Gimi (n=27) and Keiagana (n=3) and ranging in age from 50 to 82 years old, with a mean age of 60 years. On closer inspection of village-specific kuru exposure (explained later in this chapter), 9 individuals were found to originate from an unexposed village and were removed from further study.

The remaining 125 samples were genotyped at codon 129 (as described in 2.3.13). The elderly multiple-kuru exposure women demonstrated a highly significant deviation from Hardy-Weinberg equilibrium ($P_{\text{exact}}=3.1 \times 10^{-5}$) with a significant excess of heterozygotes ($P<<0.0000001$) as tested using the Score test implemented in the software GENEPOP (Figure 6.3).

The added data support the original analyses and illustrate the remarkable survival advantage proffered by heterozygosity at codon 129 to a cohort with acute and repeated exposure to kuru. An interesting question raised by these data relates to the “protection” afforded to homozygotes – in a cohort of women acutely exposed to kuru on multiple occasions, how have women with a “susceptible” codon 129 genotype managed to survive? These data suggest the possibility of codon 129 independent protection, which is examined later in this chapter.

---

* The Score test (or U test) as implemented in GENEPOP, is an exact HWE test which specifically tests for excess heterozygosity as an alternate to the null hypothesis, which is that the population examined is in HWE.
6.4 Investigating *PRNP* for a kuru-mediated signatures of selection

6.4.1 Codon 129 genotypes in the surviving Eastern Highlands population

Codon 129 genotype proportions were used to examine the scale and the extent of the kuru epidemic in the surviving population of the Eastern Highlands. The reduced fitness of codon 129 homozygotes and their subsequent removal from the population should have produced a characteristic signature in the genotype proportions of the remaining population, detectable as deviation from HWE.

![Figure 6.4 Schematic map of the Eastern Highlands of Papua New Guinea](image)

Figure 6.4 Schematic map of the Eastern Highlands of Papua New Guinea

Approximate localisation of neighbouring linguistic groups are shown. For simplicity, geographical and topographical features have been omitted. Linguistic groups with experience of kuru are shaded brown and groups unaffected by kuru are shaded green.

The extent of Hardy-Weinberg disequilibrium will be proportional to the extent of kuru experienced by each population examined, with the greatest deviations seen in the North and South Fore. To examine HWE by kuru exposure, an estimate of exposure level for each linguistic group was obtained based on an assessment of number of kuru cases recorded since records began in 1957 made by the MRC-PNGIMR field team. The Eastern Highlands were divided into kuru exposed and unexposed areas (Figure 6.4).

A total of 3200 DNA samples representing 12 linguistic groups of the Eastern Highlands were obtained, either as previously extracted DNA samples held at the MRC Prion Unit or extracted from whole blood. All samples were genotyped at codon 129 as described previously. For each sample the following data were collated and samples lacking these data were not retained for analysis: the individual’s sex, age at collection, year of collection and village. Approximately 2350 samples from the South Fore (n=1279), North Fore (n=276), Keiagana (n=220), Yagaria (n=114), Agarabi (n=90),
Tairora (n=77), Siane (n=74), Awa (n=46), BenaBena (n=46), Jate (n=43), Morae (n=43) and Gadsup (n=42) were carried forward for further analysis.

6.4.1.1 The special case of heterozygote advantage

The availability of age at collection and date of collection data was imperative for accurate HWE analysis to be undertaken. These data allowed the analysis of individuals born during the period of the kuru epidemic, as opposed to individuals born after the cessation of cannibalism. To understand the importance of obtaining the effect of balancing selection on allele frequencies in generations subsequent to the removal of selective pressure should be considered:

Balancing selection establishes a unique condition in which although Hardy-Weinberg's law is disobeyed (i.e. the assumption of an absence of selection is violated), the allele and genotype frequencies in successive generations, after the selective pressure is removed, remain stable and in equilibrium. Consider the condition where complete heterozygote advantage is seen (i.e. homozygosity is lethal) in a single locus system with alleles A and B. Within a single generation all AA and BB homozygotes perish and only AB heterozygotes remain. Following mating of the surviving heterozygotes (frequencies of both alleles A and B are 50% each) all genotypes are represented at equilibrium frequency (AA and BB at 25% frequency and AB at 50%). With the removal of selective pressure, only a single generation is required to return genotypes to equilibrium conditions and any signal of selection that is based on allele frequency and genotype proportions is lost.

Therefore, the PNG samples were analysed stratified by age and exposure to kuru, calculated based on epidemiological information available. The enforced cessation of cannibalism in the mid-1950s was used as a watershed for age calculations. The majority of PNG samples were collected between 1997 and 2002 and individuals recorded as being >50 years old at sample collection were assumed to have lived during the kuru epidemic.

6.4.2 Hardy-Weinberg equilibrium in the surviving population

The surviving linguistic groups of the Eastern Highlands were tested for deviations from HWE, stratified by age. Linguistic groups were analysed (i) as a whole, (ii) stratified by age and (iii) for the North and South Fore only, stratified by age and sex, as these cohorts were sufficiently large enough to avoid small sample sizes. Significance was tested using the exact $\chi^2$ test implemented in PLINK (Table 6.1). Significant Hardy-Weinberg disequilibrium was seen in the South Fore for the whole group ($P_{exact}=0.005$) and all South Fore individuals >50 years alive during the kuru epidemic ($P_{exact}=0.003$). By analysing the >50 years group by sex, it was seen that females >50 years were
driving this result \( (P_{exact} = 0.0004) \) and no significance was seen for the corresponding male group. Significant deviation from HWE was also seen in South Fore males <50 years old. This group was further analysed by decade of birth (data not shown) and only those men born between 1950 and 1960, during the peak of the kuru epidemic, showed a significant deviation \( (P_{exact} = 0.0287) \). These data support the epidemiological and cultural data suggesting that boys up to 8 years of age would have participated at feasts with their mothers.

HWE in the North Fore population and additional kuru exposed groups including Keiagana, Jate, Agarabi, Yagaria and Awa, was puzzling. Kuru cases have been recorded in all of these linguistic groups and the North Fore, in particular females >50 years old, are recorded to have experienced frequent kuru cases, second only to the South Fore. The absence of a significant deviation from HWE in these linguistic groups may relate to a sample ascertainment problem, with all linguistic groups sharing a paucity of individuals >50 years old.
<table>
<thead>
<tr>
<th>Linguistic Group (years old)</th>
<th>N=</th>
<th>Codon 129</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MM</td>
<td>MV</td>
</tr>
<tr>
<td>S.Fore</td>
<td>1279</td>
<td>248</td>
<td>687</td>
</tr>
<tr>
<td>&gt;50</td>
<td>122</td>
<td>16</td>
<td>77</td>
</tr>
<tr>
<td>&lt;50</td>
<td>1157</td>
<td>232</td>
<td>610</td>
</tr>
<tr>
<td>S.Fore male</td>
<td>426</td>
<td>72</td>
<td>240</td>
</tr>
<tr>
<td>&gt;50</td>
<td>74</td>
<td>11</td>
<td>40</td>
</tr>
<tr>
<td>&lt;50</td>
<td>352</td>
<td>61</td>
<td>200</td>
</tr>
<tr>
<td>S.Fore female</td>
<td>853</td>
<td>176</td>
<td>447</td>
</tr>
<tr>
<td>&gt;50</td>
<td>48</td>
<td>5</td>
<td>37</td>
</tr>
<tr>
<td>&lt;50</td>
<td>805</td>
<td>171</td>
<td>410</td>
</tr>
<tr>
<td>N.Fore</td>
<td>276</td>
<td>52</td>
<td>146</td>
</tr>
<tr>
<td>&gt;50</td>
<td>81</td>
<td>10</td>
<td>43</td>
</tr>
<tr>
<td>&lt;50</td>
<td>195</td>
<td>42</td>
<td>103</td>
</tr>
<tr>
<td>N.Fore male</td>
<td>109</td>
<td>19</td>
<td>53</td>
</tr>
<tr>
<td>&gt;50</td>
<td>42</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>&lt;50</td>
<td>67</td>
<td>14</td>
<td>32</td>
</tr>
<tr>
<td>N.Fore female</td>
<td>167</td>
<td>33</td>
<td>93</td>
</tr>
<tr>
<td>&gt;50</td>
<td>39</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>&lt;50</td>
<td>128</td>
<td>28</td>
<td>71</td>
</tr>
<tr>
<td>Keiagana</td>
<td>220</td>
<td>42</td>
<td>106</td>
</tr>
<tr>
<td>&gt;50</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>&lt;50</td>
<td>216</td>
<td>40</td>
<td>105</td>
</tr>
<tr>
<td>Jate</td>
<td>43</td>
<td>7</td>
<td>26</td>
</tr>
<tr>
<td>&gt;50</td>
<td>14</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>&lt;50</td>
<td>29</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>Agarabi</td>
<td>90</td>
<td>17</td>
<td>37</td>
</tr>
<tr>
<td>&gt;50</td>
<td>13</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>&lt;50</td>
<td>77</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>Yagaria</td>
<td>114</td>
<td>28</td>
<td>56</td>
</tr>
<tr>
<td>&gt;50</td>
<td>36</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>&lt;50</td>
<td>78</td>
<td>20</td>
<td>37</td>
</tr>
<tr>
<td>Awa</td>
<td>46</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td>&gt;50</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>&lt;50</td>
<td>42</td>
<td>4</td>
<td>25</td>
</tr>
</tbody>
</table>

**Table 6.1** Hardy-Weinberg analysis of the Eastern Highland linguistic groups

HWE P-value calculated using exact test – significant deviations from HWE are in bold
6.4.3 Heterozygosity at codon 129

The Hardy-Weinberg data in Table 6.1 clearly indicates that deviations from HWE are due to excess heterozygosity at codon 129. It has been proposed that reduced diversity at various loci can be used as a tool to identify genes undergoing directional selection (Hughes et al., 2003) and so it was hypothesised that conversely, loci with greater-than-expected diversity may indicate genes undergoing balancing selection. To further refine the HWE data and develop a test to apply to other loci, heterozygosity (H) was examined at codon 129. Linguistic groups were grouped based on kuru exposure levels: 1668 high kuru exposure samples (North and South Fore), 333 low exposure samples (Awa, Gimi, Jate, Keiagana and Yagaria) and 487 unexposed samples (Agarabi, Asaro, BenaBena, Gadsup, Gahuku, Labogai, Morae, Siane, Tairora and Yabiyufa) were analysed.

![Figure 6.5 Heterozygosity at PRNP codon 129 stratified by sex and age](image)

Mean heterozygosity (H) shown ± 1.96 SEM

The difference between expected and the observed heterozygosity (ΔH) was calculated and highly exposed women >50 years and men <50 years both differed significantly (P<0.001) from the global mean ΔH ± 1.96 SEM (Figure 6.5). In addition, all cohorts examined for HWE were tested using the Score test implemented in GENEPOP. All results were negative except highly exposed women >50 years, P=0.0002 and South Fore men between 40 and 50 years P=0.0049.

6.5 Variation of PRNP codon 129 valine allele frequency

The geographical co-variation of allele frequencies with intensity of selective pressure has been routinely cited as a signature of natural selection in human populations, see for example lactase persistence alleles and the establishment of dairy farming (Harvey et al., 1998; Swallow, 2003) and blood pressure regulating alleles co-varying with latitude (Young et al., 2005). Several authors have
identified a geographical cline of PRNP codon 129 allele frequency worldwide. Mead and colleagues first reported an approximate West-East reducing cline of the codon 129 valine allele, towards South East Asia and Oceania, identified through analysis of several worldwide populations (Mead et al., 2003). The valine allele frequency of the Fore population of PNG was found to be strikingly increased compared to neighbouring populations. Mead postulated that, taken with other data, one hypothesis describing the selective pressure responsible for creating and maintaining such a cline was the widespread occurrence of acquired prion disease, transmitted through cannibalism in prehistoric human populations and that in stark contrast the high valine frequency in the Fore represented a more recent episode of selection. This result was supported by the work of Soldevila and colleagues and Hardy and colleagues who investigated worldwide codon 129 allele frequencies in populations comprising the CEPH diversity project (Hardy et al., 2006b; Soldevila et al., 2003). Both Soldevila’s and Hardy’s data supported a longitudinal West-East cline in valine frequency and identified the highest known valine allele frequencies in several Central American populations with suspected histories of cannibalism. In addition, although Lucotte and Mercier’s work investigating valine frequency in 12 French and Western European populations failed to identify a significant longitudinal cline, they did however identify a significant correlation of allele frequency with latitude (North-South) (Lucotte et al., 2005). Population data from these and other international studies provided the opportunity to perform a combined analysis to assess worldwide clines in codon 129 allele frequency.

6.5.1 Variation of codon 129 valine allele frequency worldwide

Published codon 129 genotype data for normal worldwide populations were collected from the literature representing 41 countries. Data from 281 unexposed Fore individuals was obtained. Where data existed for distinct ethnic groups within a country, these were analysed separately. For simplicity, the geographical positions used for distance calculations were either (i) the country capital or (ii) the regional capital or largest regional town (for studied populations comprising of multiple ethnic groups). For example, several authors have studied PRNP variation in multiple Chinese ethnic groups including the Hui and the Uygur peoples, whose provincial capitals are separated by almost 2000 km and whose valine allele frequencies differ markedly by almost 15% (Figure 6.6).

The online map tool MultiMap (www.multimap.com, accessed 2006) was used to identify latitudinal and longitudinal coordinates in decimal degrees and distances from the Prime Meridian (at Greenwich, UK; 51.28N, 0E respectively), calculated using the Haversine formula employed in a java script (www.movable-type.co.uk/scripts/LatLong.html). Figure 6.7 illustrates the remarkable reducing cline seen in valine allele frequency worldwide. The correlation of allele frequency with
distance from the Prime Meridian was highly significant (Correlation coefficient \( R=0.80 \), \( R^2=0.65 \) and \( P=4\times10^{-12} \)). The highest published valine frequency was found in Central America (Colombia 81% and Brazil 79%) (Hardy et al., 2006b) and the lowest frequencies were found in East and South East Asia (Taiwan 1.5% and Japan 2%) (Mead et al., 2003; Ohkubo et al., 2003). In stark contrast to other populations at comparable distances east of the Prime Meridian, the Eastern Highlands valine allele frequency (based on the Fore population) significantly deviated from its predicted value.
Figure 6.6: Worldwide PRNP codon 129 valine allele frequency
See Figure 6.7 legend.
These data consolidate previous studies of worldwide variation of codon 129 and illustrate the Fore population of PNG as an outlier to this trend. As other authors have noted the two highest occurrences of valine allele frequency occur in the Fore population of PNG and in Central American populations, which all have documented cases of cannibalism.

![Figure 6.7 Variation of PRNP codon 129 valine allele frequency worldwide](image)

Red data points: Austria (n=300) (Zimmermann et al., 1999), China (Uygurs n=223; Han n=205) (Yu et al., 2004), France (n=161) (Deslys et al., 1994; Laplanche et al., 1994), Germany (n=722) (Vollmert et al., 2006), Greece (Crete n=205) (Plaitakis et al., 2001), Greece (n=348) (Saetta et al., 2006), Holland (n=117) (Bratosiewicz-Wasik et al., 2007), Iceland (n=208) (Georgsson et al., 2006), Italy (n=318) (Del Bo et al., 2005), Japan (n=466) (Ohkubo et al., 2003), Korea (n=529) (Jeong et al., 2004), Poland (n=194) (Gacia et al., 2006), Slovakia (n=613) (Mitrova et al., 2005), Slovenia (n=97) (Galvani et al., 2005), Spain (n=268) (Combarros et al., 2000), Turkey (n=100) (Erginel-Unaltuna et al., 2001), United Kingdom (n=406) (Collinge et al., 1991), USA (n=86) (Brown et al., 1994), Scotland (n=150), Northern Ireland (n=150), Republic of Ireland (n=203), Finland (n=1957) (Nurmi et al., 2003), Mexico (Maya n=22; Pima n=25), Brazil (Surui n=21; Karitiana n=24), Colombia (n=11), Senegal (Mandenka n=24), Scotland (Oradian n=16), Algeria (Mozambique n=30), Russia (n=25), Russia (Adygei n=17), Israel (Bedouin n=24), Central African Republic (Biaka Pygmy n=36), Cameroon (n=39), Yoruba (n=25), Pakistan (Balochi, n=26), Siberia (Yakut n=26), Democratic Republic of Congo (Mbuti Pygmy n=15), Kenya (Bantu n=20) (Hardy et al., 2006b), Sri Lanka (n=35), Taiwan (n=70), PNG/Madang (n=83), Bougainville (n=22), Fiji (Taveuni n=10, other Fijian n=18), Tonga (n=22), Vanuatu (Port Olry n=33; Maewo n=32) (Mead, 2002). Blue data point: Papua New Guinea Fore population (n=281). Linear regression to the mean (R^2) indicated.

### 6.5.2 Variation of codon 129 valine allele frequency in the Eastern Highlands

As the results above show, the equilibrium valine allele frequency in the Eastern Highlands contrasts distinctly with the reducing cline seen worldwide particularly with neighbouring populations, including the costal PNG population of Madang (valine frequency 30%). Within the Eastern Highlands of PNG, kuru incidence is known to have varied geographically amongst the various linguistic groups that inhabit the region. The Fore linguistic groups comprised the epicentre of the kuru epidemic and the burden of kuru decreased, or was completely absent, in neighbouring linguistic groups. This geographical variation of kuru incidence afforded the opportunity to investigate the variation of PRNP codon 129 allele frequencies with kuru exposure amongst the linguistic groups of the Eastern Highlands, providing further evidence of the role of PRNP as a kuru susceptibility locus and a robust method for assessing other candidate susceptibility loci. Valine
allele frequency was calculated for each linguistic group and compared by exposure levels which were approximately proportional to distance from the focus of the epidemic in the South Fore region.

**Figure 6.8 PRNP codon 129 valine frequencies across the Eastern Highlands**

A. Linguistic groups with high exposure to kuru are shaded dark brown (Keiagana n=220; North Fore n=276; South Fore n=1279 and Gimi n=25). Low exposure linguistic groups are shaded light brown (Siame n=73; Tairora n=77; Yagaria n=114; Awa n=46 and Jate n=43). Unexposed linguistic groups are shaded green (Agarabi n=90; Yabiyufa n=30; Asaro n=26; BenaBena n=46; Gadsup n=42; Gahuku n=39; Labogai n=23 and Morae n=43). B. Mean valine allele frequency by kuru exposure level - no significant differences seen. Error bars are 1.96xSEM.

Unexpectedly, an equilibrium valine allele frequency was identified across the linguistic groups of the Eastern Highlands, which was a surprising result (Figure 6.8A). There were no significant differences in allele frequency between those linguistic groups known to have been acutely exposed to kuru and those with low exposure ($T$-test, $P=0.16$) and unexposed groups ($T$-test, $P=0.46$) (Figure 6.8B). In addition there was no significant difference between the low exposed and unexposed groups ($T$-test, $P=0.28$). This intriguing result may have been due to: (i) demography – the uniform high valine frequency may reflect a historical bottleneck due migration and settling of the region rather than disease (ii) an older and more widespread incidence of kuru in the Eastern Highlands than previously reported (Jerome Whitfield, pers. comm.) and (iii) the absence of balancing selection – highly unlikely with evidence of HWE deviation presented earlier. In addition, the insensitivity of these analyses to the extent of kuru exposure of individual villages, as opposed to entire linguistic groups, may explain have confounded the results. It is known that within linguistic groups which experienced high occurrences of kuru, there were some villages that remained unexposed (Jerome Whitfield pers. comm.).
6.6 Linkage disequilibrium measures in PNG

Linkage disequilibrium was also investigated as a differentiating signature of kuru imposed selection. Under the balancing selection model, the frequency of the valine allele increased in the Fore populations from an initial value (possibly that similar to the coastal Madang population ~30%) to its present equilibrium frequency. The oral evidence collected from kuru survivors suggests that the kuru epidemic was swift, affecting at most 3 generations of Eastern Highlanders, implying an equally dramatic and swift increase in valine frequency. Such a large increase in allele frequency at a locus can cause a "hitchhiking" effect of the surrounding genetic variation, known as a selective sweep. One consequence of such a selective sweep is increased LD around the functional allele.

6.6.1 LD between microsatellites flanking PRNP

Microsatellites 108 and 53*, previously identified by Dr Simon Mead 30 kb upstream and 24 kb downstream respectively of codon 129 (Mead, 2002), were genotyped in high and low kuru exposure populations and unexposed populations as described in Chapter 2.3.5. Haplotype phase was resolved by genotyping codon 129 homozygotes only. Mead previously demonstrated reduced microsatellite diversity and extended LD at both 108 and 53 in the Fore compared to European, Japanese and African populations. The additional samples and data on exposure index allowed LD comparisons across the Eastern Highlands to be made, controlling for differences due to demography of the international samples.

Microsatellites 108 and 53 and codon 129 were in significant LD ($P_{\text{exact}} < 0.000001$, tested using the log likelihood method implemented in Arlequin) in all three groups examined on both haplotypic backgrounds. Similar patterns of allelic diversity and LD were seen in both directions for high, low and unexposed groups on both codon 129 methionine (Figure 6.9) and valine (Figure 6.10) allelic backgrounds. This result implied that either kuru had no effect on LD in the Eastern Highlands, or that the extent of LD surrounding PRNP in the Eastern Highland populations extends further than the two microsatellites examined and therefore, extended LD could not be detected. If the latter were true and kuru had affected LD surrounding PRNP, a reduction in allelic diversity would also be expected on the valine haplotype, which is not seen either.

* also referred to as a108990 and a54000 in (Mead, 2002)
Figure 6.9 Microsatellite diversity upstream and downstream of the PRNP codon 129M allele
Allele frequencies at microsatellites 108 (upstream) and 53 (downstream) on the PRNP codon 129M haplotype. Maximum number of haplotypes: high exposure n=406, low exposure n=176 and unexposed n=136. No significant differences were seen in allele diversity or LD when compared between groups.

Figure 6.10 Microsatellite diversity upstream and downstream of the PRNP codon 129V allele
Allele frequencies at microsatellites 108 (upstream) and 53 (downstream) on the PRNP codon 129V haplotype. Maximum number of haplotypes: high exposure n=562, low exposure n=246 and unexposed n=162. No significant differences were seen in allele diversity or LD when compared between groups.
Microsatellites D20S97, D20S482 and D20S889 (~720kb, ~161kb and ~143kb upstream of codon 129, respectively) flanking \textit{PRNP} at greater distances were identified using the STS database UniSTS. These microsatellites were genotyped as previously discussed with help from James Uphill, in samples homozygous at codon 129, to help phase haplotypes: high exposure samples from the kuru region, unexposed samples bordering the kuru region and UK ECACC samples for comparison. Although microsatellite diversity in the PNG samples was markedly different form the UK samples, there was no significant difference between the kuru exposed and unexposed samples on both haplotypic backgrounds (Figure 6.11 and Figure 6.12).

\textbf{Figure 6.11 Microsatellite diversity upstream of the \textit{PRNP} the codon 129M allele}

Allele frequencies at microsatellites 108, D20S97, D20S482 and D20S889 upstream of codon 129M allele. Maximum number of codon 129M haplotypes PNG high exposure \(n=44\), unexposed \(n=228\) and UK \(n=288\). Allele diversity is reduced in the PNG samples compared to the UK and also in the high kuru exposed group compared to the unexposed. Although linkage disequilibrium is extensive in the PNG samples compared to the UK samples, little difference is seen between the PNG cohorts acutely exposed to kuru and those unexposed.
Allele frequencies at microsatellites 108, D20S97, D20S482 and D20S889 upstream of codon 129V allele. Maximum number of codon 129V haplotypes PNG high exposure n=40, unexposed n=344 and UK n=60. Allele diversity is reduced in the PNG samples compared to the UK and also in the high kuru exposed group compared to the unexposed. Although linkage disequilibrium is extensive in the PNG samples compared to the UK samples, little difference is seen between the PNG cohorts acutely exposed to kuru and those unexposed.

### 6.6.2 Microsatellite $F_{ST}$

$F_{ST}$ analysis of the above populations was performed using Arlequin. On both methionine and valine haplotypic backgrounds significant differentiation was seen between the two PNG cohorts against the UK cohort ($P<0.000001$, data not shown). Marginal significance was seen for valine associated alleles between kuru high exposure and unexposed ($P=0.02 \pm 0.02$) cohorts, indicating that some genetic differentiation exists between these two groups. The methionine comparison was not significant ($P=0.06 \pm 0.01$).

### 6.7 Refining the exposure of linguistic groups – exposure index

The continuing efforts of the MRC-PNGIMR field team collecting biological samples, clinical and historical data relating to the kuru epidemic permitted a refining of kuru exposure estimates to the level of individual villages within linguistic groups. A detailed explanation of how the new exposure index (EI) for each village was calculated is provided in the Materials and Methods. Briefly the EI for each village community was defined as the number of recorded kuru deaths in the
database for a village, normalise by its estimated population in 1958 and scaled by 1000. The EI was calculated for each individual village in the Fore, Gimi and Keiagana linguistic groups, however for the smaller number of villages, and deaths, in other linguistic groups the EI was calculated for each linguistic group, based on the estimated kuru-affected population. Broadly three levels of kuru exposure were adopted: high exposure (EI>200), medium exposure (30<EI<200) consisting of villages bordering the kuru region with at least some documented cases of kuru and low exposure (EI<30) consisting of villages with at least one recorded case of kuru. The new index permitted a refining of the level of exposure and the area of exposure, which may have confounded previous analyses.

6.7.1 **Hardy-Weinberg equilibrium analysis indexed by exposure**

A proportion of the Eastern Highland samples were reanalysed using the new EI (Table 6.2). As previously observed, increased homozygosity at codon 129 was associated with early onset kuru and increased heterozygosity associated with protection from kuru. 39/56 kuru cases with age at onset <20 years were homozygous at codon 129 compared to 39/125 elderly women EI>30 ($P_{exact}=1.8 \times 10^{-6}$) and 14/50 elderly women EI>200 ($P=3.4 \times 10^{-5}$). Hardy-Weinberg disequilibrium was found in elderly women with EI>30 ($P=6.6 \times 10^{-5}$) and with EI>200 ($P=0.004$), increasing the significance which was a more significant deviation than that seen before in the linguistic group analysis. Loss of HWE was not found in a stratum of slightly younger women aged 40-50, born towards the end of the practice of mortuary feasts, and was also absent from elderly men and unexposed elderly women. As previously identified, a stratum of men aged 40-50, who would have participated in mortuary feasts as children, displayed marginal HWE deviations ($P=0.016$).

6.7.2 **PRNP codon 129 valine allele frequency analysis indexed by exposure**

Using the new EI classifications, codon 129 genotypes from 282 elderly women (>50 years old in 2000) from villages with EI>30 in the North Fore, South Fore, Keiagana and Gimi were obtained (referred hereafter as the kuru region; Figure 6.14). Codon 129 data were also obtained for low exposure populations, comprising individuals from Gimi (n=87), Jate (n=157), Keiagana (n=221), Kanite (n=35) and Awa (n=46) linguistic groups. Mean valine allele frequency was not significantly different between the kuru region and low exposure populations (57.4% and 54.3% respectively) by a two-tailed $\chi^2$ test (Figure 6.13).

Additional Highland populations with no recoded cases of kuru comprising of individuals from villages speaking Asaro (n=19), BenaBena (n=20), Gadsup (n=22), Gahuku (n=42), Labogai
(n=47), Siane (n=29), Tairora (n=79), and Yabiyufa (n=27) were also tested against the mean allele counts of the kuru region data and found to be significant (two tailed $\chi^2$ test $P=0.004$; Figure 6.13).

Figure 6.13 An increasing cline in codon 129 valine frequency within Papua New Guinea

Codon 129 allele counts were compared between the kuru region (n=282) and: low exposure populations (n=546), Highland populations with no recorded kuru (n=281) and non-highland populations (n=313). Subpopulations are shown in grey. (From Mead et al., 2007).

Codon 129 genotypes were also obtained from more distant, non-Highland populations including: Vanimo/wewak (n=5), islands neighbouring PNG (n=44), Port Morseby (n=11), Western Highlands (n=4), Madang and its neighbouring inland area (n=239), and Lae (n=10) and comparison against the kuru region for allele count was highly significant (two tailed $\chi^2$ test at $P<0.0001$; Figure 6.13).

The mean distances between the villages examined and Waisa in the South Fore were calculated as previously described for each analysis group. Valine allele frequency significantly co-varied with distance from Waisa and exposure to kuru (correlation coefficient $R=-0.99$, $P=0.007$; Figure 6.13).

6.8 Additional PRNP susceptibility loci

During genetic screening studies of PRNP at the MRC Prion Unit by the Molecular Genetics Group, a novel coding change was detected in a highly conserved and structured region of PrP. The change at codon 127 of PRNP from glycine to valine was identified in elderly women >50 years from the geographically restricted region of the Purosa Valley and neighbouring villages (Figure 6.14). In this area, with the highest kuru exposure, G127V is a common variant, occurring at ~7% frequency and exclusively on the 129 methionine background (based on a retrospective analysis of Fore pedigrees collected prior to the discovery of the allele).
The G127V change was genotyped by dideoxy resequencing in 161 WGA kuru samples and was found to be completely absent. This absence was significant when 127V-129M haplotype counts were compared by a two-tailed exact Pearson $\chi^2$ test, to 125 elderly women (>50 years old in 2000) from the exposed region ($P=0.005$; Table 6.2). Based on the codon 129 heterozygous/homozygous resistance/susceptibility model, homozygosity at both loci was investigated. Homozygosity at both codon 127 and 129 was seen in 37/49 of the kuru cases <20 years old, compared to 39/125 of elderly women (<20 years old during the peak of the epidemic) which was highly significant ($P=6.6 \times 10^{-9}$, exact $\chi^2$ test, see Table 6.2). In addition, division of the kuru cohort by long (>30 years, onset after 1990) and short (aged <20 at sampling) average incubation time, illustrated that only 12/49 of the short incubation cohort were heterozygous at either codons 127, 129 or both, compared with 86/117 of elderly women ($P=6.5 \times 10^{-7}$; exact $\chi^2$ test, see Table 6.2).
Table 6.2 Genotypes of kuru patients and age-stratified healthy population controls

<table>
<thead>
<tr>
<th></th>
<th>G127V</th>
<th>M129V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>GG</td>
</tr>
<tr>
<td>All Kuru*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incubation &lt;20* years</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>Incubation &gt;30 years</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Women &gt;50°</td>
<td></td>
<td></td>
</tr>
<tr>
<td>El &gt;30</td>
<td>125</td>
<td>119</td>
</tr>
<tr>
<td>Women &gt;50°</td>
<td></td>
<td></td>
</tr>
<tr>
<td>El &gt;200</td>
<td>50</td>
<td>44</td>
</tr>
<tr>
<td>Women 40-50° years</td>
<td>150</td>
<td>144</td>
</tr>
<tr>
<td>Men &gt;50° years</td>
<td>122</td>
<td>121</td>
</tr>
<tr>
<td>Men 40-50° years</td>
<td>83</td>
<td>81</td>
</tr>
<tr>
<td>Unexposed w&gt;50 years</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

The relative fitness of each codon 127 and 129 haploptypic combination was calculated for the entire exposed population of both genders aged over 40 (n=472), which was chosen to be a large sample representative of the genetic effect of kuru on the descendent population. The relative fitness of each heterozygous combination (127GV-129MM, 127GG-129MV and 127GV-129MV) was assumed to be equal as there was insufficient data to calculate these separately. The relative fitness for 127GG-129MM was 0.64 and 127GG-129VV was 0.74, relative to a combination of the three heterozygous genotypic combinations. More highly kuru-exposed subgroups had an even lower relative fitness of homozygous genotypes: elderly exposed women (El>30), 127GG-129MM relative fitness was 0.38, and 127GG-129VV was 0.48; elderly exposed women (El>200), 127GG-129MM relative fitness was 0.23 and 127GG-129VV was 0.42.

6.8.1 Genealogy and codon 127

To confirm or refute the function of G127V mutation as a resistance factor, the incidence of kuru in the parents of the current living 127V probands, who would have lived through the kuru epidemic, was ascertained. This hypothesis was based on the assumption that as one of the parents would be expected to be a 127V carrier, an increased or reduced history of kuru in this generation would indicate whether 127V was acting as a susceptibility or resistance factor.

51 individuals carrying the 127V mutation, aged 16-78 years (mean age 35 years, 11 individuals aged 50 or older), were interviewed by an MRC-PNGIMR research nurse and none were found to
be symptomatic of dementia, ataxia or other evidence of neurodegenerative disease. Genealogies had been obtained, prior to the detection of the variant, from 18 probands with 127V. 127V genealogies were matched to all 127G pedigrees obtained from villages in the Purosa valley in which more than one 127V individual had been detected to obtain a control cohort. With the exception of Agakamata and Ilesa with moderate exposure, the village were all found in the region of highest kuru exposure. The village Ei values were as follows: Purosa-Takai (172), Ai (221), Takai (362), Kamira (337), Ketabi (265), Ivaki (295), Mugaimutu (346), Kalu (200) and Waisa (217). Only 1/36 parents from 127V genealogies was recorded as dying from kuru, whereas 33/218 parents were recorded as dying from kuru in the matched 127G pedigrees ($P=0.04$; two tailed exact $\chi^2$ test).

Given the apparent geographically restriction of the 127V mutation (Figure 6.14), it was suspected that there was a recent common ancestor of all alleles. 13 PRNP linked microsatellite markers were genotyped over 3 MB ($D20S181, D20S193, D20S473, D20S867, D20S889, D20S116, D20S482, D20S97, PRNP Codon 129, D20S895, D20S849, D20S873, D20S95 and D20S194) in all 127V individuals to test this hypothesis and haplotypes were identified using PHASE. 25/52 127V chromosomes shared at least one identical microsatellite allele across the region (Figure 6.15) consistent with a 127V-linked haplotype. The same haplotype was found in only 1/70 127G chromosomes.

The age of the 127V mutation was estimated by using the microsatellite data to find the time to most recent common ancestor (MCRA). Of the 13 microsatellites 4 were uninformative and one was excluded because of doubt over the genetic distance between this marker and PRNP. The time to MCRA was calculated by modelling recombination-mediated LD decay over time using the formula of Risch et al. (Risch et al., 1990), as corrected by Colombo (Colombo, 2000). Importantly, this calculation did not consider the likelihood of selection acting on the variant and was therefore likely to be a conservative calculation. The median result of the 8 remaining markers gave a point estimate of the most recent common ancestor of 127V occurring within 13 generations (95% confidence intervals: 0 to 30 generations). Confidence intervals based on 10,000 bootstraps of the data (See Efron, 1993).
Figure 6.15 Size of 127V-linked haplotype compared with the same haplotype on 127G alleles

13 microsatellites linked to PRNP were genotyped and 8 were then used to date the most recent common ancestor which are illustrated here (D20S181, D20S889, D20S116, Codon 129, D20S895, D20S849, D20S873, D20S95 and D20S194; map locations relative to codon 129: -1494556, -720087, -613627, 419565, 527025, 929772, 1049310, 1475648). Codon 127-linked haplotypes were determined using PHASE software. The central dash refers to an individual haplotype, 127V-linked haplotypes are red and 127G are blue. Individuals are ordered by increasing size of linked haplotype. (From Mead et al., 2007).

6.8.2 Additional PRNP independent susceptibility loci

The identification of the codon 127 mutation in the South Fore women supported the hypothesis that susceptibility factors other than codon 129 of PRNP may exist. The utility of examining neighbouring linguistic populations as a means of supporting the role of codons 127 and 129 in kuru susceptibility demonstrated the potential for this sample resource to identify PRNP-independent susceptibility loci using a genome-wide approach. Genome-wide SNP data for example could be used to obtain both an empirical distribution of “neutral” variation data and also be used to conduct genome-wide association studies.
6.9 Genome-wide analyses in the Fore linguistic group of Papua New Guinea

The analyses thus far have addressed PRNP mediated kuru susceptibility and the genomic signatures of variation that could be used to identify the participation of a locus. Developing a resource from the PNG Eastern Highland population, has allowed genetic variation at potential susceptibility loci to be examined for co-variance with exposure. Although this has been shown above to work effectively at PRNP, possibly because of its large effect size, other loci may be susceptible to confounding due to factors such as population demography. Therefore, attention was turned to obtaining an empirical distribution of SNP genotypes, against which signatures of selection may be assessed and which could be used to conduct whole-genome association studies and whole-genome selection scans.

6.9.1 Study design

Although several whole-genome SNP genotyping platforms were available, the opportunity arose to utilise the Affymetrix GeneChip Human Mapping 500K Array Sets*. The Affymetrix 500K array provided the greatest SNP density as compared to other commercial platforms with comparable genomic coverage. In total, 14 surplus NspI Affymetrix arrays were provided by Dr Simon Mead, each capable of genotyping approximately 250,000 SNPs each. The 14 arrays were used to pilot a kuru versus elderly kuru-exposed Fore females hyper-case-control study and so the study design was considerate of this and several other factors including time, cost and sample/array availability. 7 samples in each group were chosen for screening, permitting the analysis of common SNPs at MAF> 7.1% in the population. In addition the feasibility of using whole-genome amplified (WGA) kuru samples (required to maintain sufficient quantities of DNA for future research) could be tested.

6.9.2 Whole-genome amplification of kuru samples

The majority of kuru samples held at the MRC Prion Unit were collected in the years immediately following the peak of the epidemic, between 1957 and 1960 by investigators from the National Institutes of Health (Maryland, USA). Many of these samples had either degraded over time, which affected the quality of the DNA subsequently isolated, or were provided as blood sera, which typically yields low concentrations of DNA (compared to erythrocytes). To fully exploit these limited resources Multiple Displacement WGA was performed by Geneservice Ltd (Eng Ang et al., 2007) on a subset of samples to investigate the efficacy of using amplified sample for genetic investigations. 8 kuru samples (PDGs 7450, 7456, 7468, 7470, 7504, 7520, 7595 and 7744), comprising 6 females and 2 males aged from 8 years to 40 years old, which were provided as genomic DNA extracted from sera, underwent WGA. The DNA yield was increased by 1000 fold

* Each Affymetrix 500k Array set comprises two 250K arrays which genotype SNPs proximal to either NspI or StyI restriction sites
and all samples passed Geneservice’s quality control criteria except for PDG7520, which was dropped from further analysis due to allele drop-out.

6.9.3 Pre-hybridisation preparation of the PNG samples

The samples were prepared for hybridisation onto the Affymetrix chips as described in Materials and Methods (Chapter 2.3.12). During pre-hybridisation, PCRs of the adaptor-ligated fragments were visualised on 2% TBE agarose gel as shown in Figure 6.16. Prominent banding was seen at approximately 2.25kb and 2.75kb in all WGA samples only, possibly an artefact of the WGA process such as biased amplification.

![Figure 6.16 PCR of 7 South Fore multiple kuru exposure samples and 7 kuru samples](image)

The multiple exposure and kuru samples were amplified in triplicate to provide a sufficient quantity of PCR product. The whole-genome amplified kuru samples amplified poorly compared to the multiple exposure samples with prominent bands at approximately 2.75kb and 2.25kb. Sample PDG7595 showed an atypical band at 4kb. To exclude poor sample handling as a cause of the atypical band, the sample preparation was repeated with a second WGA sample with an untreated control. PDG7595 again showed an atypical band in contrast to the control samples (data not shown) however, it was retained for downstream hybridisation to identify what effect (if any) it might have on the genotyping call-rate. The PCR banding of PDG7595 had no effect on the fragmentation pattern (Figure 6.17).
Fragmentation of the ligated PCR amplicons resulted in a range of fragment sizes. The relative intensity of the multiple exposure samples compared to the WGA kuru samples reflects the differing amounts of starting material. *The fragmentation of sample PDG7595 was similar to the other WGA kuru samples. Run on 4% TBE agarose gel at 120V for 30 minutes. Ladder sizes: 10kb, 8kb, 6kb, 5kb, 4kb, 3kb, 2.5kb, 2kb, 1.5kb, 1kb and 0.5kb

**Figure 6.17** Fragmentation of ligated PCR products

6.9.3.1 **Assigning calls and data analysis**

As much of the data presented in the following sections rely on different SNP call assignments, a brief overview of how SNP genotypes are assigned is provided. Each 250K Affymetrix array contains over 6.5 million features*, each consisting of more than one million copies of a 25nt oligonucleotide probe of a defined sequence which is complimentary to the sequence surrounding a SNP. Probes are arranged as quartets comprising perfect match and mismatches to each allele and 6 or 10 probe quartets are used for each SNP at various locations on the array. SNP genotypes are called using an algorithm which utilises the fluorescent intensities of the 6 or 10 quartets distributed across the chip and the intensities of matches/mismatches within each quartet.

The hybridised arrays were scanned using a GeneChip Scanner 3000 and data analysis was conducted using the Affymetrix GeneChip DNA Analysis Software (GDAS). Genotypes were called using the Dynamic Model (DM) Mapping Algorithm employed within GDAS which also provided quality information for each call, based on the Wilcoxon's signed rank sum test. Data were obtained at three rank-confidence score thresholds (P=0.33, 0.16 and 0.1). Increasing the rank score threshold (moving the dashed line in Figure 6.18 towards the centre of the triangle) can increase the number of genotypes called, thus reducing the number of no calls, but may also reduce the confidence of these additional calls. The converse is also true.

---

* squares etched onto the glass array
Figure 6.18 Dynamic Model scatter plot for calling genotypes

Graphical representation of the Dynamic Model (DM) algorithm. Genotypes are called based on measured fluorescence intensities, converted using the DM algorithm and tested against one of four genotype models (AA, AB, BB and null), with the rank score of the most likely model used to give a confidence measure in each called genotype. High confidence calls (~0) are plotted at the vertices and scores closer to 1 are near the centre of the triangle. Decreasing the rank score threshold (dashed line) increases the number of no calls and increasing the threshold decreases the number of no calls. Adapted from (Affymetrix, 2005b).

6.9.4 Hybridisation efficiency and call rates

Subsequent to scanning, each NspI array image was visually inspected for an assessment of quality. All samples were successfully scanned except for PDG7470 for which the array image was exceptionally dark (Figure 6.19). GDAS analysis was undertaken at a default threshold (P=0.19) and call rates obtained. WGA array call rates ranged from 49.6% to 72.9% (mean ~61%) and mean multiple exposure call rate was ~95% (good quality DNA with proper handling should yield call rates >93% (Affymetrix, 2005b)). With an atypical array image and a low call rate of 60%, PDG7470 was repeated for fear that post-hybridisation (staining and washing) reagents used for this chip only were contaminated. In addition, repeating the sample provided an opportunity to obtain a technical replicate, allowing the concordance rate* to be calculated between replicates and an assessment of assay repeatability on WGA samples.

Figure 6.19 GeneChip NspI hybridisation array images for sample PDG7470

Labelled DNA hybridised poorly onto the NspI array (right) which gave a low genotype call rate. Repeating the sample gave a better call rate and a brighter array image.

* Concordance is calculated by identifying the fraction of SNPs which are assigned identical genotypes on a set of replicate arrays (two arrays hybridised with the same sample)
The repeated PDG7470 call rate improved to 69.8%, representing a net gain of 10,719 genotypes (37,081 SNP calls gained and 26,362 lost). Importantly, the majority of new calls were heterozygotes, which are often under-represented by fluorescence based detection due to bias in detecting a 50% decrease in allelic signal compared to homozygotes. This result implied that the cause of the low call rate was staining of the hybridised DNA molecules which resulted in lower fluorescence compared to background.

With 10% of the total concordance between the two arrays comprising of “no calls”, the possibility of increasing the call rate by varying the stringency of the rank score threshold was examined. This would require a trade off between call rate and genotyping accuracy and therefore the relationship between these two variables was examined. Additional replicated NspI array genotyping data were obtained for 6 vCJD samples from Dr Mead and analysed with PDG7470. The 6 vCJD samples were not true technical replicates as one sample in each replicate was WGA treated however, they could provide data on WGA DNA performance on arrays. Analyses were conducted at thresholds 0.5, 0.3 and 0.19 and compared to a base-line analysis at 0.1 (Figure 6.20).

![Figure 6.20 SNP gains and data reliability at various rank score thresholds](image)

**Figure 6.20 SNP gains and data reliability at various rank score thresholds**

Increased call rates, relative those achieved at P=0.10, were seen for all threshold values. Mean values were calculated for the 6 vCJD samples. The maximum gain in SNP calls was achieved with P=0.19 with high SNP concordance and low discordance, after which increasing the threshold yielded diminishing returns. Error bars 1.96xSEM

The 6 WGA vCJD samples had similar call rates to their genomic replicates at various rank threshold values and comparing the genomic and WGA replicates, concordance varied directly with both threshold and call rate. At the most stringent threshold of P=0.1, a low mean call rate of 82.5% was seen with a low mean concordance rate of 72%. As the threshold stringency was decreased to
0.19, 0.33 and 0.5, both the mean call rate and concordance increased proportionally suggesting that genotypes were stable between replicates and therefore the majority of the data gained were accurate. Some increase in discordance was seen however as stringency was reduced (not including genotypes which were converted into “no calls” or vice versa). The mean discordance amongst the 6 arrays increased from 0.2% to 3.6% as stringency decreased. Taken together, these data suggest that for the vCJD samples, the WGA treatment had little effect on the quality of genotype data and that for chips with low call rates, the rank score threshold could be relaxed to obtain extra reliable data.

In comparison, the kuru chips performed poorly with, at P=0.5, a maximum mean call rate of 80%, concordance of 70% and an unacceptable high genotype discordance of 10%. At P=0.1, only 57% of the SNP genotypes were called and with only 40% concordance between chips. Whether this poor performance was due to the WGA of poor starting material or if it reflected the poor post-hybridisation treatment of the first chip, was not known.

To identify the optimal threshold value for WGA treated arrays – one that would accurately yield the maximum number of SNP genotypes with the minimum amount of discordance between replicates – the mean change in SNP call rate relative to that at P=0.1 baseline was examined, against mean percentage change in concordance and discordance (Figure 6.20). For both vCJD and kuru chips the greatest increase in call rate was seen at P=0.19 with a large increase in concordance and a marginal increase in discordance. Relaxing the threshold further yielded diminishing returns in terms of call rate and although concordance increased, so too did discordance, reducing confidence in the accuracy of the additional data. As previously noted, the kuru chips performed poorly and although there was a greater increase in SNP calls, the discordance rate was much greater, suggesting that the additional data gained was not accurate.

6.9.5 Genome-wide linkage disequilibrium

To determine LD patterns genome-wide, genotype data was exported from Affymetrix GeneChip Genotyping Analysis Software (GTYPE) into Haploview to calculate pairwise LD comparisons between SNPs. Unless stated, Affymetrix data was exported at a rank score threshold of P=0.33. Pairwise LD ($D'$) was determined for loci with 85% completeness of genotype data, MAF>15% and for comparisons <100kb. The MAF cut-off of 15% was chosen based upon genome-wide data published for isolated population of Kosrae in which the allele frequency distribution of common alleles >15% was similar to other worldwide populations (Bonnen et al., 2006) and was therefore not affected by SNP ascertainment bias. Mean pairwise LD was calculated for comparisons of increasing 2.5kb distances.
6.9.5.1 Genome-wide LD is inflated in small samples

Initially, pairwise LD was assessed for the 7 samples from elderly Fore women with multiple exposure to kuru with genotype data acquired at three different rank score thresholds (Figure 6.21). Varying the threshold had little effect on the genome-wide LD profile.

**Figure 6.21 Pairwise LD in 7 elderly multiple kuru-exposed samples at different thresholds**

Mean $r^2$ was calculated for pairwise comparisons between markers with minor allele frequencies $\geq 15\%$ that fell within the same 2.5kb intermarker distance bins for 7 multiple kuru-exposed Fore females. The LD profile is similar at all three thresholds. Error bars are $1.96\times$SEM.

With only 7 samples available for analysis, it was expected that the mean LD would be inflated due to the small sample size (Jorde, 2000). To test the extent of sample size on genome-wide LD NspI array genotype data on from 90 UK control samples was obtained from Dr Simon Mead. Genome-wide LD was calculated for all 90 samples (genotypes called at threshold $P=0.33$) and for random sub-samples of 7 individuals which were sampled without replacement 4 times (Figure 6.22).

**Figure 6.22 The effect of sample size on genome-wide pairwise LD**

Mean $r^2$ was calculated for pairwise comparisons between markers with minor allele frequencies $\geq 15\%$ that fell within the same 2.5kb intermarker distance bins for 90 UK samples and four sub-sampled sets of 7 UK individuals. Pairwise LD was found to be consistently inflated for the 7 sub-sampled cohorts compared to the complete data set at all distances. Error bars are $1.96\times$SEM.
LD was consistently inflated for the smaller 7 UK sub-sampled cohort than the full 90 UK sample data set. Furthermore, long range LD was more susceptible to inflation (230% inflation for 100kb comparisons) compared to short range LD (113% inflation for 5kb comparisons).

### 6.9.5.2 Genome-wide LD compared between PNG and UK samples

To overcome discrepancies due to sample size, LD from 7 individuals from the UK were compared to the 7 kuru and 7 elderly multiple kuru-exposure Fore females (Figure 6.23). The Fore females demonstrated elevated LD compared to the UK samples and the half-life of LD decay with genomic distance was substantially longer in the Fore women than in the UK samples (approximately 1.25-fold higher than the UK samples).

![Figure 6.23](image)

**Figure 6.23 Linkage disequilibrium decay over distance**

Mean $r^2$ for pairwise comparisons between markers with minor allele frequencies ≥15% that fell within the same 2.5kb intermarker distance bins, for 7 UK samples (sub-sampled four times from 90 samples), 7 WGA kuru and 7 elderly multiple kuru exposure Fore women. LD was greater in the 7 elderly Fore women compared to the 7 UK samples and LD amongst kuru samples was starkly reduced. Error bars are 1.96xSEM.

Pairwise LD in the 7 WGA kuru samples was considerably lower than the other samples and decay over distance occurred more rapidly. This reduced LD reflected the poor quality of the data obtained from the NspI arrays for the WGA kuru samples. The reduced LD is most likely an artefact of either (i) spuriously increased heterozygosity (which serves to reduce LD) due to genotyping inaccuracies across the 7 chips or (ii) a paucity of data to conduct a sufficient number of pairwise comparisons to obtain robust LD data.

### 6.9.5.3 Genome-wide LD between highly correlated SNPs

An important consideration for any future whole-genome PNG association study, using the Affymetrix 500k platform, is coverage. As the SNPs included on the Affymetrix array were ascertained in reference populations, it is important to assess how well these SNPs represent
variation within the PNG population. The importance of coverage and several methods of directly assessing this statistic are examined in detail in the discussion (the current study was too preliminary with too few samples to be able to utilise these methods). However, an approximate indicator of coverage was ascertained by examining the proportion of minor SNP alleles that were highly correlated \((r^2 \geq 0.8)\) and therefore might serve as proxies for other unknown variants not directly genotyped on the array. The proportion of highly correlated pairwise comparisons (of the total comparisons made), within increasing 2.5kb genomic distance bins, was inflated in small samples of 7 individuals repeatedly sub-sampled from a total of 90 UK samples (Figure 6.24).

Figure 6.24  Inflation of highly correlated \((r^2>0.8)\) LD comparisons in small sized samples
Mean proportion of pairwise comparisons \(r^2 \geq 0.8\), between markers with minor allele frequencies \(\geq 15\%\) that fell within the same 2.5kb intermarker distance bins, for 90 UK samples and four sub-sampled sets of 7 UK individuals. Error bars are 1.96xSEM.

Comparing across the three sample sets, Figure 6.25 illustrates the increased LD in the elderly Fore women as compared to the UK samples and the relative paucity of highly correlated SNPs in the WGA kuru samples. The lack of highly correlated SNPs in the kuru samples, taken together with the lower overall LD, as compared to the 7 multiple-kuru exposure samples is likely to be attributable to poor genotyping of these samples (i.e. low call rates and a greater discordance) compared to the other PNG samples.
Figure 6.25 Decay of linkage disequilibrium of highly correlated \((r^2 \geq 0.8)\) alleles with distance

Mean proportion of pairwise comparisons \(r^2 \geq 0.8\), between markers with minor allele frequencies \(\geq 15\%\) that fell within the same 2.5kb intermarker distance bins, for 7 UK samples (sub-sampled four times from 90 samples), 7 WGA kuru and 7 elderly multiple kuru exposure Fore women. The proportion of comparisons \(r^2 \geq 0.8\) was greater in the 7 elderly Fore women compared to the 7 UK samples and amongst kuru samples, was starkly reduced. Kosrae data from (Bonnen et al., 2006). Error bars are 1.96xSEM.

For comparison, linkage disequilibrium data from a study of the isolated Micronesian population of Kosrae is also illustrated in Figure 6.25. The Kosrae data represents 60 unrelated individuals and was generated using the Affymetrix GeneChip 100K platform. The smaller sample size of the 7 elderly Fore women compared to the 60 Kosrae samples implies that LD on Kosrae will likely be greater than that on PNG, however, the lower SNP density 100K array may inflate LD estimates slightly as it is known that SNP density can affect the pattern and extent of LD. As the only whole-genome genotyping study of an isolated population, the Kosrae data will make an interesting comparison for future PNG data.

Several additional analyses that would provide information on the genetic diversity and the utility of the PNG samples could not be run due to small sample size

6.10 Discussion

The results presented in this chapter have expanded on previous investigations of the effect of kuru on the PNG Eastern Highland population and the effect of \(PRNP\) codon 129 on kuru survival. Moreover, the work presented in this chapter comprises the largest investigation of the genetic effect of kuru to date.

6.11 Codon 129 and kuru susceptibility

Analysis of 161 kuru samples and 125 elderly kuru-exposed females expanded on several published studies which observed significant correlation of \(PRNP\) codon 129 status with kuru survival but
were limited by small sample sizes (Cervenakova et al., 1998; Lee et al., 2001a; Mead et al., 2003). The work presented in this chapter corroborates the correlation of homozygosity at codon 129 with early kuru onset (and conversely, heterozygosity at codon 129 with late onset). In addition, these genetic analyses have recapitulated several known epidemiological aspects of the kuru epidemic:

- the extent to which the epidemic was largely restricted to females, reflecting the majority participation of women at mortuary feasts
- females were exposed to kuru at multiple mortuary feasts throughout childhood and adulthood. This is reflected in both the peak age of kuru onset in females, which occurred a full decade in life later than in males and the gradual decline in kuru incidence in subsequent decades which, in contrast to the males’ data, implies a sustained exposure to kuru through childhood and adulthood.
- male exposure to kuru largely occurred up to the age of 6 to 8 years. This is demonstrated by a peak age of male kuru onset at 11 to 20 years, given that the mean kuru incubation time is 12 years, and a rapid decline thereafter. This observation is consistent with reports that young males lived with and attended mortuary feasts with their mothers. Boys older than 6–8 years were taken from their mothers and brought up in the men’s house and from this point on, they were exposed only to the same risk as adult men, who participated little in feasts and did not eat brain (the most infectious organ in kuru) (Collinge et al., 2006). This practice explains why adult men in 1957–58 contributed only 2% to the total number of kuru cases. Consistent with a single peak period of male kuru exposure, the number of male kuru cases with onset in later decades declines rapidly

Within the surviving population from the kuru-exposed regions, highly significant deviations from HWE were seen in the surviving elderly population. The analysis of 125 women with multiple kuru exposure, corroborated and expanded the work of Lee et al. and Mead et al. demonstrating the protective advantage of heterozygosity at codon 129 and increasing the significance of this finding from the original observations (Lee et al., 2001a; Mead et al., 2003). Use of the EI classifications for elderly women from kuru exposed villages further increased the level of significance.

### 6.11.1 Codon 129 independent susceptibility?

The analyses of both the kuru-affected and surviving populations pose further questions of additional unidentified protective/susceptibility loci. The presence of MV heterozygotes in the kuru-affected cohort raises the possibility that these individuals may possess some other codon 129-independent susceptibility locus. Similarly, the presence of women homozygous at codon 129 in the surviving multiple-kuru-exposure population may be evidence of a codon-129 independent
protective factor, although long incubation times in these women cannot be discounted. Alternatively, the presence of these genotypes in either cohort may reflect non-uniform exposure within villages and linguistic groups i.e. a "hotspot" or "coldspot" of consumption within a village of linguistic group than is represented by an average exposure statistic.

6.11.2  PRNP and kuru-mediated signatures of selection

Several signatures of selection were tested on codon 129 data from 12 different linguistic groups of the Eastern Highlands, with limited success. Investigation of co-variation of Hardy-Weinberg disequilibrium with levels of kuru-exposure identified significant deviations from HWE in only the South Fore samples. South Fore women >50 years and men aged between 40-50 years showed significant departure from HWE disequilibrium. HWE was re-established in South Fore males born after 1960, demonstrating both the abrupt end to the kuru epidemic and the swift return of genotype proportions of return to HWE subsequent to removal of the selective pressure.

HWE was seen in other exposed linguistic groups including the North Fore, which is known to have experienced acute exposure to kuru second only to the South Fore. This may have been due to the early cessation of mortuary feasting in the communities of the North Fore in the 1950's or earlier, who were the first of the Fore people to lose their traditional practices in the wake of Australian administrative control (Collinge et al., 2006).

HWE in non-Fore kuru-exposed linguistic groups is probably due to the absence of gender stratification which was precluded by small sample sizes. In addition, these analyses were insensitive to the differential exposure of villages within each linguistic group. Reclassification of the PNG samples using the new EI permitted a more robust analysis: women >50 years and men aged 40-50 years both remained highly significant results and no significant HWE deviations were seen for women between 40-50 years, men >50 years or unexposed women >50 years.

The analysis of codon 129 valine allele frequency worldwide and in PNG as a signature of selection provided an interesting result. The valine allele frequency demonstrated a significant cline worldwide to which the Fore clearly contrasted. Based on the early linguistic group analysis, an equilibrium valine allele frequency was seen across the Eastern Highlands, regardless of kuru exposure status. However, under the EI classification an increasing local PNG cline was seen from the non-Highland populations, to Eastern Highland populations with no record of kuru and the kuru exposed region itself. No significant difference was seen between the low and high kuru exposure populations from the Eastern Highlands. This may be due to a historically higher incidence of kuru in the low exposure populations than was recorded since the 1950s.
Investigating the extent of LD between microsatellites flanking \textit{PRNP} did not yield conclusive results associated with the extent of kuru exposure. Kuru exposed and unexposed cohorts. Only microsatellite \textit{F}_{ST} demonstrated marginal significance.

6.11.3 \textbf{Additional susceptibility loci - PRNP G127V}

The data presented in this chapter support the conclusion that G127V is a highly geographically restricted, but locally common, polymorphism that confers resistance to kuru in the heterozygous state. How glycine to valine change protects against kuru remains unclear. The polymorphism may result in a more bulky amino-acid side-chain which impedes beta-sheet formation and protects against prion formation, although this remains speculation and additional work would be needed, to clarify the effect on \textit{PrP} structure and the mechanism of protection.

The possibility that, rather than protecting against kuru, the 127V polymorphism might have instead triggered the epidemic cannot be completely excluded. Given that life expectancy in the kuru region is 42 years, it is possible that 127V is associated with a late-onset and low penetrant inherited prion disease which might have triggered the kuru epidemic. There are few examples of mutations causative of autosomal dominant neurodegenerative disease that achieve the polymorphic frequency observed in the Purosa Valley (Wexler \textit{et al.}, 2004). A wealth of data indicates that prion transmission is generally more efficient where there is complementarity between the primary \textit{PrP} sequence of donor and host, implying that if 127V was pathogenic it would be expected to confer susceptibility to the corresponding acquired prion disease (Collinge, 2001). Additionally, the glycoform type of \textit{PrP}\textsubscript{Sc} in kuru, although restricted to a small number of autopsy samples, closely resembles sporadic CJD rather than point mutation inherited prion disease (Parchi, 2000). Finally, if 127V had triggered kuru, the localization to the southeastern part of the Fore linguistic group would be inconsistent with oral history that the first kuru patient was observed in the Keiagana linguistic group located to the northwest of the Fore.

6.11.4 \textbf{Future study of PNG and kuru samples}

An important issue that should be investigated further to conclusively resolve the use of genetic variation to identify kuru susceptibility loci is the description of “neutral” genetic variation against which variation at \textit{PRNP} or other loci may be compared. The availability of 14 Affymetrix NspI arrays permitted the limited analysis of genomic variation in 7 Highland samples and the feasibility of utilising degraded kuru DNA through WGA.
6.11.4.1 The effect of whole-genome amplification of low quality material on whole-genome genotyping platforms

Recent studies at the MRC Prion Unit have proven the reliable use of WGA technologies for use on various genotyping arrays on high quality starting DNA. The results presented in this chapter have also illustrated the benefits of using WGA on good quality starting material but have shown that WGA on low concentration, degraded kuru DNA is not an effective method of increasing DNA quantity for Affymetrix NspI array analyses.

6.11.4.2 Ascertainment bias

All SNPs on the Affymetrix GeneChip arrays were pre-selected based on technical quality in the three HapMap reference populations (CEPH, Japanese/Chinese and Yoruban), essentially representing a quasi-random SNP set (Affymetrix, 2005a), and therefore it is unknown how representative the alleles ascertained in these reference populations are of common variation in PNG. Although the Affymetrix SNPs are likely to be representative of the majority of human SNP alleles, there is a possibility that a major class of common allele on PNG may not be represented on these arrays. Future work should concentrate on deep resequencing of several genomic regions in 16 PNG samples from unexposed PNG linguistic groups. Resequencing a subset of the ten 500kb regions investigated in the HapMap ENCODE resequencing project (www.hapmap.org) would permit a direct comparison of PNG genetic diversity against freely available data from African, European and Asian populations. The frequency distribution of all alleles could be compared to ensure that the Affymetrix 500k arrays also represent common variation in PNG. In addition, resequencing a subset of the 17 genomic regions studied in the Kosrae population would permit direct comparisons to be made with a second isolated population with available SNP data.
7 PRNP copy number polymorphisms and prion disease susceptibility

7.1 Introduction

This chapter investigates PRNP copy number as a mediator of sporadic CJD and kuru susceptibility. The chapter begins with a brief consideration of the relevance of copy number changes to neurodegenerative diseases and presents the hypotheses that PRNP duplications may mediate susceptibility to sporadic CJD, whilst deletion may afford protection against kuru. PRNP copy number was investigated using quantitative real-time PCR in 28 sporadic CJD samples and 104 elderly women from the Fore. A suspected triplication was identified in a sCJD sample and although no copy number changes were identified in the Fore women, a known polymorphism at codon 127 of PRNP was observed. This chapter ends with a brief discussion of the relevance of these results and the relevance of copy number polymorphisms to human prion disease.

7.1.1 Copy number polymorphisms in neurodegenerative disease

In recent years the insertion, deletion and duplication of DNA segments greater than 1 kb have been shown to occur frequently in the human genome (see for example (Conrad et al., 2006; Freeman et al., 2006; Khaja et al., 2006)). With approximately 12% of the human genome undergoing dynamic rearrangements, these copy number polymorphisms (CNPs) can have dramatic phenotypic consequences, altering gene dosage, disrupting coding sequences, or perturbing long-range gene regulation (Redon et al., 2006).

The accumulation and deposition of proteins is a common feature of many neurodegenerative diseases, and variability in the expression of an associated protein-coding gene is often related to the severity of neurodegeneration (i.e. age of onset, duration of disease etc.). This dosage hypothesis is supported by several examples of pathogenic copy number changes which have been observed to give rise to neurodegenerative diseases: (i) tandem duplication of a 1.5Mb region of the gene PMP22 results in the dominantly inherited demyelinating neuropathy, Charcot–Marie–Tooth type 1A (CMT1A) (Lupski et al., 1991), (ii) duplication of the amyloid precursor protein gene (APP) has been reported in five different families with autosomal dominant early onset Alzheimer’s disease (Rovelet-Lecrux et al., 2006), (iii) duplication of the proteolipid protein gene (PLP) in the rare X-linked dysmyelination disorder, Pelizaeus-Merzbacher disease (PMD) (Woodward et al., 1998) and partial triplication of exons 2 to 4 of Parkin (PARK2). Perhaps most convincing support for a dosage hypothesis to neurodegenerative disease comes from the pathogenic multiplication the α-synuclein gene (SCNA) in autosomal dominant Parkinson’s disease. Triplications of SCNA have been independently identified in two unrelated kindreds with onset in the fourth decade (Farrer et
al., 2004; Singleton et al., 2003). In addition duplication of the SCNA locus has also been shown to be pathogenic in several families, with a less severe clinical phenotype than triplication families and with onset a decade later (Chartier-Harlin et al., 2004; Ibanez et al., 2004; Nishioka et al., 2006).

7.1.2 Variability in PRNP expression and disease susceptibility

There are several lines of evidence that suggest that PRNP expression is correlated to disease outcome. In mice, the level of expression of Prnp, regardless of coding sequence, is a major determinant of incubation time. For example, transgenic mouse models with additional copies of the hamster PrP gene demonstrate an inverse correlation between PrP expression level and scrapie incubation time (Prusiner et al., 1990). Bueler and colleagues importantly showed that Prnp-ablated mice are resistant to developing disease when inoculated with mouse scrapie (Bueler et al., 1993) and that hemizygous Prnp-ablated mice have a greatly prolonged incubation time (Bueler et al., 1993). Conversely, increasing Prnp expression by increasing Prnp transgene copy number in Prnp<sup>Δα</sup> mice, increases susceptibility to Rocky Mountain Laboratory (RML) prion infection (Fischer et al., 1996). PrP levels following disease onset have also been able to modify disease outcome. Mice displaying disease following RML prion infection are seen to reverse spongiform pathology and cognitive and behavioural defects following depletion of neuronal PrP<sup>C</sup> (Mallucci et al., 2003; Mallucci et al., 2007).

In addition, a QTL mapping study of inbred mouse strains has identified a region of chromosome 2, encompassing Prnp, as one of three major QTLs that determine prion-disease incubation time after intracerebral inoculation with mouse-adapted scrapie (Lloyd et al., 2001). Intriguingly, this result has been replicated in humans - a 10kb prion haplotype confers risk to sporadic CJD (Mead et al., 2001). Although it is not known if the associated haplotype results in altered expression of PRNP, it suggests that genetic variability in prion expression may contribute to sporadic disease risk.

Taken together, these data suggest that PRNP mediated disease susceptibility extends beyond polymorphisms of the amino acid sequence and that PRNP expression may influence human prion disease outcome. The genetic elements that determine PRNP expression have been studied. Mahal and colleagues were the first to isolate and characterise the human PRNP promoter, identifying by progressive 5' deletions of the promoter region, several candidate sites where nucleotide polymorphisms may influence expression (Mahal et al., 2001). To date three SNPs in the regulatory region of the PRNP have been identified, a C to G transversion at position -101, a G to C transversion at position +310 and T to C transition at position +385. It has been shown that among sCJD patients homozygous for codon 129 methionine alleles, the numbers of who carried the rare promoter SNP alleles was higher than in controls, suggesting that the regulatory region
polymorphisms may be a risk factor for CJD (McCormack et al., 2002). It has yet to be demonstrated if PRNP expression is mediated by copy number changes.

7.1.3 Copy number polymorphisms in sporadic CJD?

The majority of examples involving pathogenic copy number alterations in neurodegenerative diseases demonstrate familial or pseudo-familial inheritance. Conversely, the majority of neurodegenerative diseases including human prion disease occur sporadically, so can copy number changes explain sporadic disease? Essentially, the first appearance of a pathogenic copy number change is a sporadic de novo event which is subsequently inherited, for example the various SCNA families possess duplications and triplications of different allele sizes reflecting their independent origins. Reduced penetrance has been observed for several diseases that result from CNPs, including the 22q11 deletion DiGeorge syndrome (Gong et al., 2001) as well as speech problems in patients with duplication of the Williams syndrome region (Gong et al., 2001). Therefore, it is possible that incomplete penetrance may also be seen with potential PRNP copy number cases that have been diagnosed as sCJD.

7.1.4 PRNP copy number hypothesis

Two discrete hypotheses relating altered PRNP copy number to disease susceptibility were tested in this chapter:

(i) elderly Fore women (aged >50 years old in 2000), with acute exposure to kuru through attendance at multiple mortuary feasts, who were genotyped as codon 129 homozygotes, were actually hemizygous which afforded relative protection against kuru.

(ii) sporadic CJD patients who were heterozygous at codon 129 and should have essentially been less susceptible to disease, actually had a duplication of the PRNP129m allele

7.2 Quantitative Real Time PCR probe and primer design for PRNP

Selection of PRNP PCR primers and MGB probe sequences for quantitative PCR (qPCR) was performed as described in Materials and Methods 2.5.3. The probe was designed to anneal 5' to codon 129 in a region of PRNP without any known sequence variation (Figure 7.1). Primers identical to those used for codon 129 genotyping by allele discrimination were used and a methionine allele specific PRNP probe was also used. Ideally, the probe should have straddled an intron/exon boundary to amplify genomic DNA only (and not mRNA) however, by spanning codon 129 probe specificity could be tested by comparing detection of M and V codon 129 alleles.

RNA contamination was minimised by treating DNA samples with RNAse either during extraction from blood or (for samples obtained as DNA) following the identification of RNA impurities in
DNA by spectrophotometry. The house-keeping gene beta-actin (*β-Actin*; OMIM: 102630) was used as an endogenous control and *β-Actin* probe and primer sequences were taken from the literature (Suarez-Merino et al., 2005).

**Figure 7.1** *PRNP* and *β-Actin* quantitative real time PCR probes and primers

Probes are indicated in yellow boxes and forward and reverse primers are in grey boxes. The methionine specific *PRNP*<sub>129M</sub> probe is indicated by the black line. *All* *PRNP* probes were major groove binders (MGB) and were labelled with VIC fluorophore reporters. *β-Actin* probes were MGB and labelled with FAM fluorophore reporters.

To determine changes in *PRNP* copy number in relation to *β-Actin*, the comparative $C_T$ method ($2^{-\Delta\Delta C_T}$) was used (Livak et al., 2001). Although this method is usually used to quantify changes in gene expression it is also useful for quantifying genomic template changes and standardising assays across different experiments/plates. In brief, the difference between the threshold cycle of the *PRNP* target assay and the *β-Actin* endogenous reference assay ($ΔC_T$) is calculated for both the test and calibrator samples:

Equation 7.1 $\Delta C_T = C_T$ (target gene) - $C_T$ (endogenous control)

The difference between the average $ΔC_T$ value of a test sample and the average $ΔC_T$ for the calibrator sample is calculated ($ΔΔC_T$) and the expression fold value by:

Equation 7.2 $ΔΔC_T$ (test sample) = $ΔC_T$ (test sample) - $ΔC_T$ (calibrator sample)

Equation 7.3 Expression fold value = $2^{-ΔΔC_T}$

### 7.3 Validating qPCR probe efficiencies

For the $ΔΔC_T$ calculation to be valid, the amplification efficiencies of the target and reference assays had to be approximately equal. A sensitive method for assessing if two amplicons have the same efficiency is to look at how $ΔC_T$ varies with template dilution. *PRNP*<sub>129M</sub> and *PRNP* probe

---

* $C_T$ or Threshold Cycle reflects the cycle number at which the fluorescence generated within a reaction crosses a baseline threshold of background fluorescence.
efficiency was validated by serially diluting a DNA sample of known concentration over a 32-fold range (approximately 200ng/µl to 6ng/µl – the working range of the majority of the samples used for the study). Assays were performed as described in Materials and Methods 2.6.7 and 5 replicates were conducted for each dilution. The average ΔCₜ was calculated as explained above (Equation 7.1).

Figure 7.2 Relative efficiencies for PRNP and PRNP₁₂₉M quantitative PCR probes
The efficiency of amplification of the PRNP (A) and PRNP₁₂₉M (B) targets and endogenous control β-Actin by CT (top) and ΔCT (bottom). The data were fitted using least-squares linear regression analysis (N=5).

The relative efficiencies of the target and endogenous control probes, as calculated by the slope of the ΔCₜ against graph, were approximately equal (Figure 7.2) and compared well to the acceptable efficiency of <0.1 established by the manufacturer of the real-time system (Applied Biosystems, 1997).

Figure 7.3 Average ΔCₜ differences between PRNP codon 129 MM and MV CEPH samples
Error bars are ± 95% SEM. VV homozygote not shown. (N=11)

To ensure the assay was sensitive enough to detect a two-fold difference in template concentration, the methionine-specific PRNP₁₂₉M probe was used to evaluate 23 CEPH samples blinded to codon
129 genotype (11 MM, 11 MV and one VV control sample). A highly significant difference between the $\Delta C_T$ of the MM and MV samples (T-test $p<<0.0001$) was seen and, as expected, the valine homozygote controls failed to amplify with the $PRNP_{129M}$ probe (Figure 7.3). This result demonstrated that a two fold increase in genomic template (one cycle difference) could be detected.

7.4 Copy number variation in elderly Fore females

The 104 elderly Fore women were screened with both the $PRNP_{129M}$ and $PRNP$ copy number probes. The samples were comprised of 17 MM homozygotes, 64 MV heterozygotes and 18 VV homozygotes from the North Fore, South Fore and Gimi. The mean age was 59 years (ranging from 50 to 82 years, with a median age 57 years). In addition, 25 healthy controls from the North and South Fore were screened to identify deletion/duplication polymorphisms at a detectable frequency >2%. All samples were screened in triplicate as described in Chapter 2.7.3, with 5 identical control samples included to standardise across plates (the control sample with the least within-plate variation was used as the calibrator), and the mean $2^{\Delta C_T} \pm 1.96 \times$SEM calculated (see Equation 7.2 and Equation 7.3).

Initially, whilst waiting for reagents to arrive, 20 multiple-exposure samples were screened with the $PRNP_{129M}$ probe and three control samples (one MV, MM and VV). Samples were calibrated by an MM homozygote control. Methionine homozygotes generally fell within the $2^{\Delta C_T}$ range 1.04 – 1.08 and heterozygotes between 2.04 – 2.14, however two MM samples were observed to not conform to the heterozygote group: sample PDG2711 (1.99 ± 0.07) and PDG9466 (1.57 ± 0.03) (Figure 7.4). Both of these samples were suggestive of a heterozygous methionine deletion.

On closer inspection it was recognised that both of these samples possessed a glycine to valine heterozygous mutation at codon 127, which had been recently found to occur in several samples from the Fore populations. The mutation was a G→T transversion at the second position within codon 127 (GGC to GTC). As these samples had been genotyped as heterozygous at additional loci linked to codon 129 (data not shown), it was clear that the suspected deletion was a spurious result, most likely a consequence of the 127 mutation lying directly within the $PRNP_{129M}$ probe binding region.

* The data from replicates was tested for conformation to a normal distribution, proving that the use of statistical tests based on the normal distribution was valid.
Figure 7.4 *PRNP* 129M copy number analysis of 25 Fore women over 50 years old repeatedly exposed to kuru

Samples calibrated by an MM homozygote. Red and blue lines give average homozygote and methionine heterozygote $2^{\Delta \Delta C_T}$ values respectively with dashed lines ±1.96xSEM.

When examined with the general *PRNP* probe, these samples were seen to lie within the population range. The difference in the $2^{\Delta \Delta C_T}$ values of PDG2711 and PDG9466 from the MM average may reflect a stochastic effect of the 127 mutation spuriously affecting probe binding to the template. Although this result was due to an oversight in assembling the sample set, it also provided a precedent for the assay to detect sequence variations.

---

Figure 7.5 *PRNP* copy number analysis of Fore women over 50 years old repeatedly exposed to kuru

Samples calibrated by an MM homozygote control. 104 samples were screened with 5 control samples. The blue line gives average $2^{\Delta \Delta C_T}$ value ±1.96xSEM. Error bars are ±1.96xSEM.

The 104 multiple kuru-exposure samples were screened as described above, with 5 control samples to ensure data were standardised between experiments (Figure 7.5). Mean $2^{\Delta \Delta C_T}$ was 0.98 ± 0.02,
with values within a normal range of 0.8 - 1.2. All samples were observed to fall within the normal range and close to the population mean. The 25 healthy control samples also showed no difference in copy number (data not shown). The general \textit{PRNP} probe was not affected by the G\rightarrow T mutation at codon 127 (possibly due to the position of the probe/template mismatch generated by the mutation).

### 7.5 Copy number variation in sporadic CJD samples

To examine copy number changes of \textit{PRNP} as a pathogenic mechanism in sCJD, 28 brain-derived sCJD samples were screened with the \textit{PRNP} copy number probe, with 4 healthy control samples from the UK. All samples were RNAse treated and qPCR was undertaken in triplicate as described above. As Figure 7.6 illustrates the majority of samples fell within the population mean of 0.95 ± 0.05 (±SEM) however, one sCJD sample (PDG2271) with average $2^{\Delta\Delta CT}$ value 3.77 ± 0.3 was clearly distinct from the others. PDG2271 was taken form an anonymous patient with confirmed sporadic CJD who was heterozygous at codon 129. As a consequence of anonymity of the sample, the patient’s clinical information was not available. The average copy number of PDG2271 was approximately 4, which suggested that the patient would have possessed a triplication of both \textit{PRNP} alleles.

![Figure 7.6 PRNP copy number analysis of sporadic CJD samples](image)

\textit{Figure 7.6 PRNP copy number analysis of sporadic CJD samples}

A total of 28 sCJD samples and 4 control samples are shown (samples 29 to 32 are controls). Red line gives average $2^{\Delta\Delta CT}$ value ±1.96xSEM. Error bars are 1.96xSEM.

The DNA quality of PDG2271 was rechecked as described in Materials and Methods 2.3.3 and the $260/280$ ratio was found to be sub-optimal at 1.64. A standard curve was constructed using a 32-fold serial dilution of PDG2271 (as described above) to dilute any impurities that may have affected the experiment. The standard curve produced for PDG2271 was within the tolerances described previously (data not shown) and therefore DNA quality could be excluded.
7.6 Discussion

This chapter describes the development of a robust assay to examine PRNP copy number as a novel mechanism for disease susceptibility in prion disease and extends the development of the PNG sample collection as a model to detect novel susceptibility loci. The PRNP and β-Actin probes and primers designed for this assay were demonstrated to effectively discriminate two or more fold differences in PRNP and PRNP

\( _{129M} \) copy number, from genomic DNA varying in concentration from 6ng/μl to 200ng/μl. This assay was applied to test two hypotheses, both based on the observation that codon 129 status alone is insufficient to explain disease susceptibility in several human prion diseases.

7.6.1 Hypothesis 1 - PRNP deletion mediates protection against kuru

The first hypothesis related to the elderly Fore women of the Eastern Highlands of PNG, who were repeatedly exposed to kuru at multiple mortuary feasts but failed to develop disease. Based on data from the previous chapter, ~30% of elderly kuru-exposed women possessed a susceptible codon 129 homozygous genotype and yet had not developed disease. The mechanism regulating protection may involve the deletion of a PRNP allele to effectively reduce PrP production to a level where the clinical disease is not seen. The assay was applied to 104 samples and 25 controls and failed to detect any deletions of PRNP.

Using the methionine specific assay, two samples, which had been genotyped as MM homozygotes, were seen to deviate from the average \( 2^{ΔΔC_T} \) copy number value. These two samples were found to harbour a known polymorphism at codon 127, within the probe-binding region. The resulting single-base mismatch may have affected specificity of probe binding and thus deviating \( C_T \) values. Indeed the sensitivity of MGB probes to mismatches has been described in the literature, see for example (Yao et al., 2006). It is intriguing however that only the binding of the methionine specific probe was affected by the mismatch, despite both probes spanning the same region. It is likely that the type and relative position of the mismatch influences the kinetics of probe binding, such as altering melting temperatures, which can explain differential results such as these (Kutyavin et al., 2000).

The absence of PRNP deletions in the elderly Fore females was a disappointing result however, the limit of detection of PRNP CNPs was 0.4%, implying that it is unlikely that PRNP deletion is a common protective mechanism against kuru.
7.6.2 Hypothesis 2 – PRNP duplication mediates susceptibility of codon 129 heterozygotes to sCJD

The hypothesis that those individuals heterozygous at codon 129 but who had succumbed to sCJD may have possessed a duplication of a PRNP allele (which would have gone undetected by genotyping codon 129 alone) was examined. Of the 28 UK sCJD samples tested, one potential triplication of PRNP was found. This was an intriguing result not least because the possible triplication had occurred on both alleles. The triplication of both PRNP alleles as a de novo event is incredibly unlikely (de novo CNP of two allelic loci is restricted to a few circumstances such as polyglutamine tract expansions). As witnessed in Parkinson’s disease, triplications often result in an earlier age of onset and pseudo-autosomal inheritance (Lesage et al., 2006; Lucking et al., 2001). The inheritance of two triplicated loci is a possible mechanism, although without population frequency data it can only be assumed that the probability of acquiring such a genotype through random mating is small, implying therefore that the patient belonged to a consanguineous pedigree. Unfortunately, historical anonymisation of sCJD samples restricted the availability of clinical and familial data which would have helped to confirm this potential result.

Due to time limitations, additional work could not be undertaken to explore this sample further. However, future work to ensure that the result is genuine should explore the following avenues:

- the result should be replicated - the sample was originally extracted from autopsied brain material and should therefore be re-extracted to ensure that the result is replicated and not merely a product of inherent, undetected contamination.
- the assay could be repeated with different probes, both for the target and endogenous control, to ensure that unknown polymorphisms in the binding regions of the primers and probes are excluded and that the copy number of the beta actin gene does not vary.
- a second independent technique could be used to identify a duplication, such as FISH or extended homozygosity/heterozygosity.

7.6.3 Limitations of this approach and future work

There are limitations to the approach used in this chapter to detect CNPs and demonstrate that these are phenotypically relevant.

7.6.3.1 PRNP gene coverage

Despite the sensitivity of the copy number assay, it was limited to resolving information on relative copy number in a restricted region of the gene. PRNP spans approximately 15kb across chromosome 20 with a 12.5kb intron separating exons one and two. The assay amplified ~100bp of
exon two which therefore omitted information on, amongst other regions, the promoter. With regards to dosage effects and expression levels in neurodegenerative diseases, the promoter is an important region of the gene to assay – and with the promoter and other upstream elements implicated in sCJD (Mead et al., 2001) hint. Therefore it may be necessary, for future work to design probes/primers that span the *PRNP* genomic locus representing in particular, the promoter region.

### 7.6.3.2 *PRNP* copy number and kuru

One attractive hypothesis would be to investigate potential duplications in individuals who succumbed to kuru despite being heterozygous at codon 129. Currently the availability of good quality DNA for such tests is limited although efforts are currently being pursued at the MRC Prion Unit to consolidate samples using whole genome amplification – although it is not known how gene copy numbers will be affected by these procedures.

### 7.6.3.3 A test for somatic mutation

The somatic mutation hypothesis is a likely mechanism to explain the majority of sCJD and the role of CNPs in these somatic mutations will only be elucidated by a sensitive test (such a sensitivity test was planned but limited time dictated that the test was not conducted). The copy number assay developed here should be tested for detection of less than two-fold increases in template. Brain tissue displaying somatic mosaicism is likely to comprise of subtle differences in templates and so the ability to detect a range of template concentrations should be investigated.

### 7.6.3.4 A test for other diseases associated with *PRNP*

The development of a copy number assay for the prion gene could also inform on the pathogenesis of other diseases in which the *PRNP* has been implicated. Inclusion body myositis (IBM) is one such disease, which was beyond the scope of this study (MIM: 147421). IBM is one of the most common inflammatory myopathies of late onset and is defined histologically by inclusions which are immunoreactive for several proteins including beta amyloid fragments. PrP immunoreactivity has also been noted within IBM inclusions and an association between codon 129 genotype and IBM has been proposed but not replicated (Lampe et al., 1999; Orth et al., 2000). Although the consequence of PrP deposition within inclusions in IBM is not known, a similar dosage mechanism to that hypothesised here for human prion diseases may exist and therefore, such a copy number assay may be able to elucidate the genetic mechanism responsible.
8 General discussion

The work presented in this thesis has attempted to apply evolving techniques and analyses used to investigate complex diseases to two different neurodegenerative diseases. As each set of experiments has been discussed individually, the purpose of this chapter is to review this work in context.

8.1 Evolving tools and techniques

A common thread throughout this thesis has been the rapid advance in our ability to analyse complex traits. Many of the tools and techniques evolved during or after the work in this thesis was carried out and therefore too late to impact this work. However, where this has happened, the implications of these changes have been discussed. The effect of large scale project and new resources such as the HapMap, on research carried out concomitantly is not a new phenomenon. Prior to the sequencing of the human genome, methods to localise Mendelian disease genes were extremely difficult and very expensive. For example, localising the cystic fibrosis gene cost several hundred million dollars because of the huge amounts of unknown DNA which needed cloning and careful analysis.

8.2 DYN1H1 and ALS

Mutation screening of DYN1H1 The preliminary aim of the work presented in the first results chapter was to elucidate the genomic structure of DYN1H1, a candidate gene for ALS. The initial evidence implicating a role for DYN1H1 in ALS pathogenesis came from mouse models of neurodegeneration in which mutations were positionally cloned to the mouse homologue Dynclhl. More recently, mutations in FALS families have been identified in a subunit of dynactin, a binding partner of cytoplasmic dynein. Using in silico methods, a 78 exon genomic structure of DYN1H1 spanning 86.6kb was elucidated.

Subsequent work in this chapter aimed to screen DYN1H1 exons 8, 13 and 14 for disease-associated mutations in index cases from FALS, HSP and SMA families and controls. These exons were homologous to those harbouring mutations in the Loa and Cra1 mouse models of disease. Although two SNPs were identified in exons 8 and the intervening intron between exons 13 and 14, neither was seen to be associated with disease. This was essentially a direct association study and therefore unknown variants within the gene were not tested.

DYN1H1 association study The second results chapter aimed to conduct an indirect association study, to investigate genetic variation across DYN1H1 for an association with SALS. SNPs were
ascertained, the underlying pattern of LD elucidated and tSNPs selected. Two tSNPs were found to be sufficient to tag the majority of variation across *DYNC1H1* and these were genotyped in SALS cases and controls. No significant association was seen.

During investigation of *DYNC1H1* the pattern of genetic variation – specifically, reduced diversity – was suggestive of an evolutionary signal of selection. Several indicators of selection were investigated including heterozygosity, $F_{ST}$, haplotype frequencies and LD decay, in northern Europeans, Cameroonian and Japanese populations. Although the results were suggestive of selection at this locus, due to ascertainment bias, no conclusive answers could be reached.

**Genetic analysis of the cytoplasmic dynein subunit families** The aim of this third results chapter was to conduct a phylogenetic study to clarify the relationships between the known cytoplasmic dynein subunits, clarify their mapping positions, identify novel members of the subunit families and clarify discrepancies in their nomenclature. In the absence of an association of *DYNC1H1* with both FALS and SALS, several additional components of the cytoplasmic dynein-dynactin complex were considered for further study. However, the available data on these subunits was found to be confusing and even erroneous. Interspecies phylogenetic comparisons and an exhaustive search for novel human and mouse homologs have clarified many existing discrepancies.

### 8.2.1 Is there a role for *DYNC1H1* in ALS?

The role of *DYNC1H1* in ALS remains unclear. The studies conducted in this thesis have identified no association of *DYNC1H1* with either FALS or SALS however, as previously discussed, it is highly likely that these studies were underpowered due to the lack of available samples and factors intrinsic to the study design.

Evidence is accumulating that suggests *DYNC1H1* remains a good candidate for ALS and potentially other neurodegenerative diseases by virtue of its importance in retrograde axonal transport. Defective axonal transport has been observed in models of ALS by several investigators (Jablonka *et al.*, 2004; Rao *et al.*, 2003; Williamson *et al.*, 1999) and defects in anterograde axonal transport are one of the earliest pathologies observed in SOD1 mice (Williamson *et al.*, 1999). The work of Hafezparast and colleagues directly linked Dynclhl mutations in both Loa and Cra1 mice with inhibited retrograde axonal transport and a progressive motor neuron degeneration phenotype (Hafezparast *et al.*, 2003). More recently, compelling data for the involvement of *DYNC1H1* in ALS pathogenesis has come from a surprising observation by Kieran and colleagues in a compound heterozygous mouse created by crossing a SOD1<sup>G93A</sup> mouse with a *Loa* heterozygous mouse (Kieran *et al.*, 2005). Kieran and colleagues noted that whilst heterozygous littermates of both
mutations displayed a normal disease phenotype with impaired axonal transport and motor neuron
death, the compound heterozygote showed a 28% increase in lifespan (compared to SOD1 \textsuperscript{G93A}
mice), increased motor neuron survival and rescued axonal transport deficits. The mechanism by
which the dynein mutation induces amelioration in Loa/SOD1 \textsuperscript{G93A} mice is still under investigation,
however Kieran postulates that amelioration of disease in Loa/SOD1 \textsuperscript{G93A} mice may result from the
restoration of axonal homeostasis as the balance between anterograde and retrograde transport in
double-heterozygote motor neurons is restored. In addition, the dynein mutation may result in
abnormal intracellular transport, which in turn may change the interaction of mutant SOD1 with
organelles such as mitochondria, thus delaying cell death.

8.2.1.1 The dynein-dynactin complex and ALS

Beyond \textit{DYN1H1} there is substantial evidence linking the cytoplasmic dynein-dynactin complex
with ALS pathogenesis. Mutations in at least three different subunits of the complex, \textit{Dynclhl},
dynamitin (p24) and dynactin (p150), have all been shown to result in motor neuron degeneration
phenotypes (Hafezparast \textit{et al.}, 2003; LaMonte \textit{et al.}, 2002; Munch \textit{et al.}, 2004; Munch \textit{et al.}, 2005;
Puls \textit{et al.}, 2003; Puls \textit{et al.}, 2005). In addition, the role of this complex as a retrograde motor
protein makes it a good candidate in other axonal transport-mediated neurodegenerative diseases.
For example, \textit{DYN1H1} has been found to colocalise with amyloid plaques in the brains of AD
patients and transgenic mouse models of AD (Liao \textit{et al.}, 2004). In addition, disrupted axonal
transport has been witnessed in Huntington’s disease in which the expanded mutant huntingtin
protein is thought to disrupt axonal transport by interaction with the p150\textsuperscript{Glued} subunit of dynactin
indirectly, via the huntingtin-associated protein HAP1 (Engelender \textit{et al.}, 1997). The expanded
glutamine repeats in huntingtin have been proposed to disrupt the integrity of the dynein/dynactin
complex, thus reducing the efficiency of neurotrophic factor transport and contributing to neuronal
death (Gauthier \textit{et al.}, 2004).

There is clearly some degree of genetic heterogeneity in the cytoplasmic dynein-dynactin complex
and its associated neurodegeneration phenotype. Clarification of the genetic relationships of the
cytoplasmic dynein subunits was undertaken with a view to conduct a future association study in
which multiple members of the dynein-dynactin complex may be investigated in a ‘candidate
pathway’ approach. Kasperavičiūtė and colleagues recently performed such a study, analysing 1277
putative functional and tSNPs in 134 genes, including 17 members of the dynein-dynactin complex,
in 822 British SALS samples and 872 controls (Kasperaviciute \textit{et al.}, 2007). 19 SNPs showed a
trend of association in the initial screen and were genotyped in a replication sample of 580 German
sporadic ALS patients and 361 controls, after which no strong evidence of association was seen.
Despite this negative result, this study was only able to capture common SNPs (>5%) with
predicted moderate effect size. By genotyping only a subset of SNPs in the replication sample variants with lower effect size may have been missed and functional variants which were not represented well by the initially genotyped set of SNPs may have also been missed. Additionally, some variants are known to have population-specific effects, as is illustrated by the association of the VEGF gene promoter polymorphisms and ALS. These polymorphisms have been positively associated with disease in three studies (Sweden, Belgium, Birmingham) which has not been replicated in another four (London, Sheffield, The Netherlands, North America) (Brockington et al., 2005; Chen et al., 2006; Lambrechts et al., 2003; Van Vught et al., 2005). Therefore, the genes of the dynein-dynactin complex may be associated with disease in additional populations.

8.2.2 Are association studies the right approach for complex neurodegenerative diseases?

The application of association studies to complex diseases has had mixed success. Non-replication of association findings is common for complex diseases however, few association studies have been found to be consistently replicable (Table 8.1). The success of an association study is dependent on several factors including: sample size and sample homogeneity (phenotypic and therefore genetic homogeneity and population homogeneity); genotyping accuracy; statistical analyses (corrections for multiple candidate genes or genome-wide analyses) and study design /genetic architecture to name a few (Abou-Sleiman et al., 2004; Hattersley et al., 2005; Healy, 2006).

<table>
<thead>
<tr>
<th>Disease</th>
<th>Gene</th>
<th>Polymorphism</th>
<th>Associated allele -Freq</th>
<th>-OR</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thrombophilia</td>
<td>F5</td>
<td>Leiden Arg506Gln</td>
<td>0-03</td>
<td>4</td>
<td>(Bertina et al., 1994)</td>
</tr>
<tr>
<td>Crohn's disease</td>
<td>CARD15</td>
<td>3 SNPs</td>
<td>0-06</td>
<td>4-6</td>
<td>(Hugot et al., 2001)</td>
</tr>
<tr>
<td>Alzheimer's</td>
<td>APOE</td>
<td>e2/3/4</td>
<td>0-15</td>
<td>3-3</td>
<td>(Corder et al., 1993; Farrer et al., 1997; Rubinsztein et al., 1999; Saunders et al. 1993)</td>
</tr>
<tr>
<td>Osteoporotic fractures</td>
<td>COL1A1</td>
<td>Sp1 restriction site</td>
<td>0-19</td>
<td>1-3</td>
<td>(Mann et al., 2001; Mann et al., 2003)</td>
</tr>
<tr>
<td>Age-related macular deeneration Type 2 diabetes</td>
<td>HFE/CfH</td>
<td>Tyr402His</td>
<td>0-30</td>
<td>2-5</td>
<td>(Edwards et al., 2005; Hageman et al., 2005; Haines et al., 2005; Klein et al., 2005; Zare-Darsi et al., 2005)</td>
</tr>
<tr>
<td></td>
<td>KCNJ11</td>
<td>Glu23Lys</td>
<td>0-36</td>
<td>1-23</td>
<td>(Gloyon et al., 2003)</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>CTLA4</td>
<td>Thr17Ala</td>
<td>0-36</td>
<td>1-27</td>
<td>(Marron et al., 1997; Ueda et al., 2003)</td>
</tr>
<tr>
<td>Graves' disease</td>
<td>CTLA4</td>
<td>Thr17Ala</td>
<td>0-36</td>
<td>1-6</td>
<td>(Chistiakov et al., 2003)</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>INS</td>
<td>Variable 5' tandem repeats</td>
<td>0-67</td>
<td>1-2</td>
<td>(Bennett et al., 1996)</td>
</tr>
<tr>
<td>Bladder cancer</td>
<td>GSTM1</td>
<td>Null (gene deletion)</td>
<td>0-70</td>
<td>1-28</td>
<td>(Engel et al., 2002)</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>PPARG</td>
<td>Pro12Ala</td>
<td>0-85</td>
<td>1-23</td>
<td>(Altshuler et al., 2000)</td>
</tr>
</tbody>
</table>

Table 8.1 Consistent associations with complex disease

Approximate disease associated polymorphism or haplotype frequency shown (~Freq) and odds ratio (~OR) shown
8.2.2.1 Sample size

Sample size is of particular concern in neurodegenerative disease and is often a limiting factor in association studies of these diseases including the DYNCIHI study described in this thesis. Sample size is a direct determinant of study power and therefore the study outcome. However, many neurodegenerative diseases have low disease prevalence with short survival times, which restricts the development of the large sample collections required to detect associated variants with moderate-to-low effect size or of low population frequency. As our understanding of the importance of sample size continues to develop, more multicentre and international collaborations have been undertaken. For example, the recent genome-wide association study in ALS undertaken by Kasperavičiūtė and colleagues was comprised of over 800 samples from collections across the UK and resultantly, their study had sufficient power to detect a causal variant with OR of 1.47 (Kasperaviciute et al., 2007). In addition, WGA of these study samples was undertaken to increase DNA stock 300-fold and ensure that these limited resources were not depleted without any detrimental impact to genotyping quality or efficiency.

8.2.2.2 Statistical analyses

The availability of cheaper and more efficient genotyping platforms has recently made genome-wide association studies of complex diseases a reality. Several different platforms are now available which either employ fixed marker sets, such as the Affymetrix GeneChip arrays, or custom marker sets, such as the GoldenGate assays (reviewed in Syvanen, 2005) and these sets can also comprise HapMap tSNPs. However, the multiple testing of 100,000 or more SNPs increases the type I error rate and therefore the significance threshold must be corrected to reflect this. Several different methods of achieving an appropriate genome wide significance threshold exist but they all mean that low $P$-values ($\alpha$-value) are required to achieve significance. This in turn affects the sample size required to detect an association at various susceptibility allele frequencies (Table 8.2). Genome-wide studies of rare neurodegenerative diseases may therefore require unfeasibly large sample sizes.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Susceptibility allele frequency in controls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>0.05</td>
<td>13599</td>
</tr>
<tr>
<td>0.01</td>
<td>19258</td>
</tr>
<tr>
<td>0.001</td>
<td>27055</td>
</tr>
<tr>
<td>5×10^{-8}</td>
<td>36869</td>
</tr>
<tr>
<td>5×10^{-9}</td>
<td>58678</td>
</tr>
</tbody>
</table>

Table 8.2 Effect of differing statistical significance levels on sample size

Numbers indicate sample size needed to detect significant association (power=90%) for different values of $\alpha$, assuming allelic odds ratio of 1.3, given differing allele frequencies for predisposing allele or haplotype. (From Hattersley et al., 2005)
8.2.2.3 Rare diseases

Both the frequency and penetrance of causal alleles affect the statistical power to detect these alleles; power increases with increasing frequency and increasing penetrance. The power to detect an allele therefore depends on what is ultimately the most relevant measure of a genetic variant's contribution: the proportion of the phenotypic variance in the population that is explained by a particular variant. This means that rare variants with modest effects will be difficult to detect by any method because they explain only a small fraction of the variance in a trait.

Rare alleles might also be more difficult to detect by association for other reasons. Even if rare alleles have strong effects, they might be difficult to detect by association methods because they are less well represented in SNP databases and because tag SNP approaches are currently designed to tag common SNPs (usually with frequencies >5%). However, population-genetic considerations indicate that most rare alleles with frequencies <5% are likely to have arisen relatively recently (because old alleles tend to either disappear or become common), so there will have been less time for recombination and mutation to disrupt the haplotype on which they arose. Therefore, rare variants are expected to be on single, long haplotypes, as has been observed (Kamatani et al., 2004).

However, as previously discussed, new techniques such as the exhaustive searching of all haplotypes employed in the EATDT approach could greatly increase power to detect rare variants (de Bakker et al., 2005; Lin et al., 2004). In addition, Vermeire and colleagues have shown that the alleles of CARD15 with MAF<5% that contributes to IBD could have been detected indirectly with haplotypes composed of common variants (Vermeire et al., 2002) and a rare (<3%) haplotype in the 11β-hydroxysteroid dehydrogenase type 1 gene (HSD11β1) has been shown to be associated with a 6-fold increase in risk for sporadic AD.

8.3 Identifying kuru susceptibility loci

Several different aims were investigated in the first kuru results chapter:

(i) to perform a comprehensive analysis of PRNP codon 129 genotype data

HWE analysis of 147 kuru samples corroborated the correlation of homozygosity at codon 129 with earlier kuru onset and conversely, the protective effect of heterozygosity at codon 129. Although only age of onset was available as a comparable phenotype, it is more likely that codon 129 status affected kuru incubation time of these cases, however, without knowing approximately when transmission occurred and at what dose, incubation time would be impossible to estimate. However, a recent study of long incubation time kuru deaths (deaths post-1960) by Collinge and colleagues identified a significant excess of codon 129 heterozygotes, suggesting that heterozygosity at codon 129 is associated with long incubation times (Collinge et al., 2006). The protective effect of codon
129 was illustrated by analysis of 125 elderly Fore women who despite being exposed repeatedly to kuru at multiple mortuary feasts, did not succumb to disease. The codon 129 genotype in this cohort was highly enriched for heterozygotes and found to be extremely significant when compared to the genotype proportions expected under HWE.

(ii) to identify signatures of selection at codon 129 as a paradigm for other candidate loci
Several different tests of selection were undertaken at codon 129 and across the PRNP locus. The co-variation of these signatures of selection with the intensity of the selective force (i.e. kuru) was examined. HWE and heterozygosity analyses based on linguistic groups identified deviations from HWE in those groups known to have the highest exposure to kuru but not in those with moderate exposure to kuru. In addition, despite a world-wide cline in codon 129 valine allele frequency, no cline within the Eastern Highlands was seen. The establishment of an exposure index to account for the exposure of each individual village within a linguistic group consolidated the HWE data and identified a significant valine frequency cline between the highly exposed and unexposed Highland populations and non-Highland populations.

(iii) to develop a PNG sample panel as a resource to test candidate genes
The refinement of the PNG samples according to kuru exposure has provided a powerful resource for testing candidate genes against. The significant variation of valine allele frequency from non-Highland to Highland samples and within the Highlands from kuru-unexposed to kuru-exposed samples illustrates how potential susceptibility loci may be identified. Such clines in allele frequency or magnitude of other measures such as deviation from HWE or $F_{ST}$ at candidate loci may be indicative of a susceptibility allele.

(iv) to assess the feasibility of genome-wide genotyping on archived kuru samples
The analysis of 7 kuru samples which had undergone WGA and 7 multiple exposure samples on Affymetrix NspI GeneChip arrays permitted several observations to be made. Despite the PCR banding patterns obtained following amplification of the WGA kuru DNA, fragmentation of the samples appeared relatively normal. The optimum rank score threshold for analysis of the WGA samples, compared to a baseline analysis at $P=0.1$, was found to be $P=0.19$. At this threshold, the greatest gain in SNP calls was seen with a minimal increase in discordance of calls. The analysis of pairwise LD identified that LD was insensitive to the rank score threshold used to analyse the data, possibly because the discordance seen at lower stringencies comprised a small fraction of the overall pairwise comparisons. The inflation of pairwise LD in a small sample size was seen to affect long-range LD to a greater extent compared to smaller pairwise intervals (230% inflation for 100kb comparisons compared to 113% inflation for 5kb comparisons). LD decay over distance for the 7
WGA kuru samples was markedly different to that of 7 UK and 7 multiple exposure samples, which probably reflected the failure of the WGA DNA samples to genotype accurately. Interestingly, the decay of LD in the 7 multiple exposure samples was 1.25-fold greater than that of the UK samples, with a similar distribution implying that LD is extends over a greater distance in these PNG samples.

8.3.1 **Signatures of selection at PRNP**

8.3.1.1 **Evolutionary studies require good epidemiological evidence**

All tests conducted for signatures of selection and for the influence of codon 129 on kuru incubation/onset have relied on good epidemiological data. For example, the protective effect of heterozygosity at codon 129 in women aged >50 years in 2000 required the documentation of the date of sample collection, age data and knowledge of the approximate cessation of the practice of endocannibalism. Together, these data allowed a powerful analysis of those women who lived during the peak of the kuru epidemic but who had not succumbed to kuru. Similarly, the HWE analysis of the surviving Eastern Highland population required knowledge of these data and appreciation of the selection model (i.e. sex specific balancing selection). These data permitted the stratification of data to provide a clearer signal of selection and resultantly, the HWE analysis recapitulated the known epidemiology of the disease.

The importance of precise epidemiological data was also highlighted in the investigation of a valine allele cline across the Eastern Highlands. Analysis of a valine allele cline, covarying with kuru exposure was undertaken within linguistic groups. This led to the observation of an equilibrium frequency for valine alleles across the Eastern Highlands. However, on application of the exposure index for each village within a linguistic group, a significant cline in valine allele frequency was seen.
8.3.1.2 Was evidence for selection seen at $PRNP$?

There were 3 key pieces of evidence that supported the role of selection at $PRNP$:

(i) a highly significant increase in codon 129 heterozygosity was seen in those exposed to kuru but not succumbing to disease.

(ii) a significant cline in valine allele frequency was seen with the highest value seen in the kuru exposed group suggesting that kuru may have been responsible for the increase.

(iii) although the microsatellite data did not yield any useful information regarding differences in LD between exposed and unexposed groups, marginal significance was seen with microsatellite $F_S$ on the valine chromosome, implying that a sudden increase in the valine haplotype may have resulted in a selective sweep of surrounding alleles.

8.3.1.3 $PRNP$ selection study caveats

There are several caveats to the work carried out to identify selection at $PRNP$.

A necessarily limited amount of polymorphism data was analysed. Although it would have been preferable to have sequenced the $PRNP$ ORF and flanking sequences to test for selection using measures based on the five main signatures of selection described by Sabeti and colleagues (Sabeti et al., 2006) and summarised in Chapter 1, financial and time constraints dictated that available codon 129 data and microsatellite data was used. This restricted the analysis largely to single locus tests which omits data from surrounding polymorphisms.

The cline in valine frequency may be a demographic effect. The increase in valine frequency from costal PNG populations to within the Eastern Highlands could reflect the stochastic effects of genetic drift through migration and population bottlenecks. Unlinked loci should have been typed to ascertain if the increase in valine allele frequency was due to demography rather than selection, as following a population bottleneck a reduction in diversity will be seen across the genome at unlinked loci. Whereas following selection, only the genetic diversity related to the selected locus will be affected. Unlinked markers were not typed in this study as a whole-genome study in these samples was anticipated.

A population bottleneck cannot explain however, the significant difference seen in valine frequency between the Highland exposed and unexposed populations. Analysis by linguistic groups indicates that both of these cohorts have an equilibrium frequency however when grouped according to exposure index, a significant difference is seen.
**Are the assumptions of Hardy-Weinberg equilibrium violated?** Several of the tests applied to the PRNP data in this thesis and other published work have interrogated codon 129 data for departures from HWE and have attributed departures to a violation of the assumption of an absence of selection. However, as there is no written history in the Eastern Highlands, we cannot be clear about how the populations of the Eastern Highlands are structured, if migration occurs amongst them and if mating is truly random – at least prior to oral history being obtained by interview. Violation of these assumptions should be stated as a caveat to these data.

8.3.2 **G127V is an additional PRNP susceptibility factor**

The novel G127V polymorphism was found to be highly geographically restricted to the Purosa Valley, where it was a common variant occurring at ~7%. Analysis of the multiple exposure women and exposed men against the kuru samples illustrated that the polymorphism confers resistance to kuru in the heterozygous state. The mechanism of protection associated with this polymorphism is unclear: it may result steric hinderance from the addition of a bulky amino-acid side-chain which resultanty impedes beta-sheet formation and protect against prion formation.

The possibility that, rather than protecting against kuru, the 127V polymorphism might have instead triggered the epidemic cannot be completely excluded. Given that life expectancy in the kuru region is 42 years, it is possible that 127V is associated with a late-onset and low penetrant inherited prion disease which might have triggered the kuru epidemic. There are few examples of mutations causative of autosomal dominant neurodegenerative disease that achieve the polymorphic frequency observed in the Purosa Valley (Wexler et al., 2004). A wealth of data indicates that prion transmission is generally more efficient where there is complementarity between the primary PrP sequence of donor and host, implying that if 127V was pathogenic it would be expected to confer susceptibility to the corresponding acquired prion disease (Collinge, 2001). Additionally, the glycoform type of PrPSc in kuru, although restricted to a small number of autopsy samples, closely resembles sCJD rather than point mutation inherited prion disease (Parchi, 2000). Finally, if 127V had triggered kuru, the localization to the southeastern part of the Fore linguistic group would be inconsistent with oral history that the first kuru patient was observed in the Keiagana linguistic group located to the northwest of the Fore. Future work should included the modelling of the 127 mutation in transgenic mouse lines and in cell-models to identify how this change may protect against disease.
8.3.3 Is selection a valid method for mapping loci in all neurodegenerative diseases?

The use of identifying a signature of selection for identifying loci involved in the pathogenesis of complex disease is an interesting concept but one that is not yet supported by concrete examples.

There are several reasons why signatures of selection may not be applicable to all complex neurodegenerative diseases. Perhaps the most significant is that pointed out by J.B.S. Haldane who recognised that late-onset genetic diseases in humans, such as Huntington's disease, encounter only weak selection (Haldane, 1941). Haldane's observation recognises that the principle driver of evolution is differential fitness or fecundity, meaning that an organism's ability to reproduce must be affected by mutation. However, the majority of cases of complex neurodegenerative diseases are late-onset, presenting after reproduction is completed and therefore are likely to be under little selection pressure.

For Mendelian diseases, which are usually caused by rare strongly deleterious mutations, a mutation-selection-balance model can be evoked, in which disease alleles are continuously generated by mutation and eliminated by purifying selection before reproduction. This model of purifying selection can also be extended to common late-onset diseases, by assuming that although onset occurs after reproduction is complete, the susceptibility genotypes still have weakly deleterious effects on fitness (Pritchard, 2001; Reich et al., 2001b). Weak purifying selection is actually a pervasive mechanism underpinning patterns of variation in human populations, as non-synonymous variants occurring at evolutionarily conserved positions, tend to occur at lower frequencies compared with synonymous polymorphisms (Hughes et al., 2003). Empirical data have also corroborated this model. In analysing nucleotide sequences of genes known to be mutated in Mendelian diseases and replicated genes in complex disease, Thomas and Kejariwal identified a preponderance for Mendelian disease-associated non-synonymous SNPs to occur at highly conserved positions in proteins compared to their complex disease counterparts (Thomas et al., 2004). This result strongly suggests that, on average, the molecular effects of non-synonymous SNPs in complex diseases will be more subtle than the severe functional changes associated with most non-synonymous SNPs in Mendelian disease. In addition, under this mutation-selection-balance model multiple rare new alleles increasing risk are predicted to be found at trait loci. This has been observed at loci such as the ABCA1 gene responsible for low HDL cholesterol (Cohen et al., 2004).
In contrast, some mutations may have pleiotropic effects, which are beneficial in youth, but yield a greater subsequent risk of neurodegenerative disease. These mutations would then be incorporated into the population by selection, which will act more strongly on the early, beneficial effect. This can be considered an extension of James Neel’s “thrifty gene” hypothesis which attempts to explain a tendency of certain populations towards obesity and diabetes (Neel, 1962). The hypothesis postulates that certain human genes have evolved to maximize metabolic efficiency which would thereby confer a survival advantage in times of food shortages. However, in times of abundance, the thrifty genotype no longer confers a survival advantage, instead predisposing carriers to diseases caused by excess nutritional intake, such as obesity and diabetes. This hypothesis has been used to explain ethnic variation in AD prevalence (Farrer et al., 1997) linked to variation in the APOE ε4 allele frequency (Reviewed in Laws et al., 2003). Briefly, APOE ε4 is associated with increased levels of dietary cholesterol and hypercholesterolaemia may play a role in the mechanism by which APOE ε4 increases the risk for AD. Therefore, a change in dietary environment for populations possessing this “thrifty gene” may lead to an increased risk of AD. This phenomenon has been suggested to explain why Africans in Nigerian and East Africa show no association of APOE ε4 with AD whereas in African-Americans an association has been reported (Reviewed in Laws et al., 2003).

A very strong signature of selection was found in the Fore, related to kuru. Perhaps those neurodegenerative diseases where signatures of selection may be of most use are diseases similar to kuru – epidemics with high mortality rates and high selection coefficients. ALS with Parkinsonism and dementia seen in Guam and a similar disease in the Kii peninsula of Japan of ALS with dementia are at first glance, good candidates, However despite having a historical prevalence ~400 per 100,000 population (Galasko et al., 2000) and ~194 per 100,000 population (Yase et al., 2001) respectively, onset and death occur after reproductive age is reached: mean age of onset is 55 years (Galasko et al., 2000) in the Guamanian disease and 56.5 years in the Kii form of ALS (Yoshida et al., 1998). The unique feature of prion disease – it’s transmissibility – makes kuru exceptional amongst neurodegenerative diseases. K may be our best example of a neurodegenerative disease affected by selection and provide the first clues to novel susceptibility loci but it remains to be seen if evolutionary analyses will have utility in other neurodegenerative diseases.

8.4 Future directions for complex genetic analyses of neurodegenerative disease

It is likely that future studies of complex diseases will not be based solely on association studies or studies to identify evolutionary selection. A paper published within the last few weeks by David Reich and his group has illustrated the statistical power available to detect disease variants by combining these two approaches (Ayodo et al., 2007). Reich and colleagues observed that power to
detect risk variants for malaria was increased by several orders of magnitude if case-control association studies were conducted with variants showing allele frequency differentiation between differently exposed populations. The applicability of this type of test to the majority of neurodegenerative diseases remains to be seen, however, kuru, which has been shown in this thesis to have caused codon 129 allele frequency differentiation in differently exposed populations, may be a good place to begin.

In conclusion, the work presented in this thesis aimed to investigate two complex neurodegenerative diseases by means of two very different methods – a candidate gene case-control association study and analyses based on the signatures of selection. The work presented in this thesis reflects the methods and study designs accepted and widely used at the time and therefore, this work is very much “of its time”. No association of DYN1I with FALS or SALS was seen. In the human prion disease kuru, strong balancing selection imposed by the disease was witnessed at codon 129 of PRNP and a new protective polymorphism was identified. And finally, the role of PRNP copy number polymorphisms was investigated as a potential pathogenic mechanism, in kuru and sCJD. From the data collected, it is likely that PRNP CNPs could account for a small minority of disease cases.
References


Al Chalabi, A., Andersen, P.M., Chioza, B. et al. (1998). Recessive amyotrophic lateral sclerosis families with the D90A SOD1 mutation share a common founder: evidence for a linked protective factor. *Hum Mol Genet.* 7, 2045-2050


Blair, I.P., Bennett, C.L., Abel, A. et al. (2000). A gene for autosomal dominant juvenile amyotrophic lateral sclerosis (ALS4) localizes to a 500-kb interval on chromosome 9q34. Neurogenetics. 3, 1-6


Bueler, H., Aguzzi, A., Sailer, A. et al. (1993). Mice devoid of PrP are resistant to scrapie. Cell. 73, 1339-1347


Chapman, J.M., Cooper, J.D., Todd, J.A. et al. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered.* 56, 18-31


Cohen, J.C., Kiss, R.S., Pertsemlidis, A. et al. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science.* 305, 869-872


Combarros, O., Sanchez-Guerra, M., Llorca, J. et al. (2000). Polymorphism at codon 129 of the prion protein gene is not associated with sporadic AD. Neurology. 55, 593-595


Daly, M.J. and Rioux, J.D. (2004). New approaches to gene hunting in IBD. *Inflamm Bowel Dis.* 10, 312-317


Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika.* 80, 3-26


Ghani, A.C., Donnelly, C.A., Ferguson, N.M. et al. (2003). Updated projections of future vCJD deaths in the UK. *BMC Infect Dis.* 3, 4-


Gong, W., Gottlieb, S., Collins, J. et al. (2001). Mutation analysis of TBX1 in non-deleted patients with features of DGS/VCFS or isolated cardiovascular defects. *J Med Genet.* 38, E45-


Honig, L.S., Chambliss, D.D., Bigio, E.H. et al. (2000). Glutamate transporter EAAT2 splice variants occur not only in ALS, but also in AD and controls. *Neurology. 55*, 1082-1088


Hubbard, T., Barker, D., Birney, E. et al. (2002). The Ensembl genome database project. *Nucleic Acids Res. 30*, 38-41


Kamatani, N., Sekine, A., Kitamoto, T. et al. (2004). Large-scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP Maps, of 199 drug-related genes in 752 subjects: the analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs. Am J Hum Genet. 75, 190-203


Khoris, J., Boukaftane, Y., Moulard, B. et al. (1997). Familial amyotrophic lateral sclerosis associated with either homozygous or heterozygous Asp90Ala Cu/Zn-superoxide dismutase mutation. *Proceedings of the 8th International Symposium on ALS/MND, Abstract Booklet.* 46-


King, S.M., Barbaresi, E., Dillman, J.F., III et al. (1996b). Brain cytoplasmic and flagellar outer arm dyneins share a highly conserved Mr 8,000 light chain. *J Biol Chem.* 271, 19358-19366

King, S.M., Barbaresi, E., Dillman, J.F., III et al. (1996a). Brain cytoplasmic and flagellar outer arm dyneins share a highly conserved Mr 8,000 light chain. *J Biol Chem.* 271, 19358-19366


King, S.M. and Patel-King, R.S. (1995). The M(r) = 8,000 and 11,000 outer arm dynein light chains from Chlamydomonas flagella have cytoplasmic homologues. *J Biol Chem.* 270, 11445-11452

243


Kondo, K. (1996). Rising prevalence of neurodegenerative diseases worldwide. Intern Med. 35, 238-


Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell. 44, 283-292


Lambrechts, D., Storkebaum, E., Morimoto, M. et al. (2003). VEGF is a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death. Nat Genet. 34, 383-394


Lampe, J.B., Gossrau, G., Herting, B. et al. (2003). HLA typing and Parkinson's disease. Eur Neurol. 50, 64-68


246


McCormack, J.E., Baybutt, H.N., Everington, D. et al. (2002). PRNP contains both intronic and upstream regulatory regions that may influence susceptibility to Creutzfeldt-Jakob Disease. Gene. 288, 139-146


MGI. (2004). Dnch1 MGI Entry: mKIAA0325 synonym. Mouse Genome Informatics.


Nagase, T., Ishikawa, K., Nakajima, D. et al. (1997). Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro. DNA Res. 4, 141-150


NCBI. (2004e). DNCL11 LocusLink Entry: Dynein Light Chain A synonym. NCBI LocusLink.


NCBI. (2004g). Mouse D2LIC entry: D2LIC, mD2LIC, MGC7211, MGC40646, 493340401Rik synonyms. GenBank.


Nebert, D.W., Sophos, N.A., Vasiliou, V. et al. (2004). Cyclophilin nomenclature problems, or, 'a visit from the sequence police'. Hum Genomics. 1, 381-388


251


Parton, M.J., Broom, W., Andersen, P.M. *et al.* (2002). D90A-SOD1 mediated amyotrophic lateral sclerosis: a single founder for all cases with evidence for a Cis-acting disease modifier in the recessive haplotype. *Hum Mutat.* 20, 473-


Robakis, N.K., Sawh, P.R., Wolfe, G.C. et al. (1986). Isolation of a cDNA clone encoding the leader peptide of prion protein and expression of the homologous gene in various tissues. *Proc Natl Acad Sci U S A.* 83, 6377-6381

Robberecht, W., Aguirre, T., Van Den, B.L. et al. (1996). D90A heterozygosity in the SOD1 gene is associated with familial and apparently sporadic amyotrophic lateral sclerosis. *Neurology.* 47, 1336-1339


258


Veldink, J.H., Kalmijn, S., Van der Hout, A.H. et al. (2005). SMN genotypes producing less SMN protein increase susceptibility to and severity of sporadic ALS. Neurology. 65, 820-825


Windl, O., Giese, A., Schulz-Schaeffer, W. et al. (1999). Molecular genetics of human prion diseases in Germany. *Hum Genet.* 105, 244-252


Young, J.H., Chang, Y.P., Kim, J.D. et al. (2005). Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. PLoS Genet. 1, e82-


Zareparsi, S., James, D.M., Kaye, J.A. et al. (2002). HLA-A2 homozygosity but not heterozygosity is associated with Alzheimer disease. Neurology. 58, 973-975

Zhang, J., Rowe, W.L., Clark, A.G. et al. (2003). Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. Am J Hum Genet. 73, 1073-1081


Zhao, J.H. and Sham, P.C. (2002). Faster haplotype frequency estimation using unrelated subjects. Hum Hered. 53, 36-41

## 10 Appendices

<table>
<thead>
<tr>
<th>Number</th>
<th>ID</th>
<th>Position (bp)</th>
<th>Alleles</th>
<th>Amplicon size (bp)</th>
<th>Genotype</th>
<th>Minor allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>rs2180511</td>
<td>27670</td>
<td>A/T</td>
<td>535</td>
<td>AA</td>
<td>6.0%</td>
</tr>
<tr>
<td>2</td>
<td>rs1678034</td>
<td>30289</td>
<td>A/C</td>
<td>566</td>
<td>AA</td>
<td>6.3%</td>
</tr>
<tr>
<td>3</td>
<td>rs2180513</td>
<td>61497</td>
<td>C/T</td>
<td>622</td>
<td>AB</td>
<td>0%</td>
</tr>
<tr>
<td>4</td>
<td>rs1956297</td>
<td>74179</td>
<td>C/T</td>
<td>453</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>5</td>
<td>rs2474677</td>
<td>75613</td>
<td>C/T</td>
<td>360</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>6</td>
<td>rs2474678</td>
<td>75617</td>
<td>A/G</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>7</td>
<td>rs2448241</td>
<td>76261</td>
<td>C/T</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>8</td>
<td>rs2093025</td>
<td>76319</td>
<td>C/T</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>9</td>
<td>rs2474679</td>
<td>76323</td>
<td>C/T</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>10</td>
<td>rs1044838</td>
<td>76846</td>
<td>A/C</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>11</td>
<td>rs2861339</td>
<td>76858</td>
<td>C/T</td>
<td>510</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>12</td>
<td>rs2273434</td>
<td>76920</td>
<td>C/G</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>13</td>
<td>rs2273435</td>
<td>77049</td>
<td>A/G</td>
<td>15</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>14</td>
<td>rs2448240</td>
<td>77419</td>
<td>C/G</td>
<td>360</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>15</td>
<td>rs2448239</td>
<td>79006</td>
<td>A/G</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>16</td>
<td>rs1741151</td>
<td>81081</td>
<td>C/T</td>
<td>14</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>17</td>
<td>rs2403015</td>
<td>89054</td>
<td>A/G</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>18</td>
<td>rs1622866</td>
<td>89874</td>
<td>A/G</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>19</td>
<td>rs1741155</td>
<td>88892</td>
<td>A/G</td>
<td>12</td>
<td>BB</td>
<td>0.3%</td>
</tr>
<tr>
<td>20</td>
<td>rs2749911</td>
<td>90216</td>
<td>A/G</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>21</td>
<td>rs2720193</td>
<td>90225</td>
<td>A/G</td>
<td>460</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>22</td>
<td>rs2720194</td>
<td>90295</td>
<td>A/G</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>23</td>
<td>rs2448242</td>
<td>91338</td>
<td>A/G</td>
<td>440</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>24</td>
<td>rs2273438</td>
<td>107101</td>
<td>C/T</td>
<td>390</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>25</td>
<td>rs2273439</td>
<td>107195</td>
<td>C/T</td>
<td>380</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>26</td>
<td>rs2720218</td>
<td>109449</td>
<td>A/C</td>
<td>190</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>27</td>
<td>rs2749890</td>
<td>114824</td>
<td>A/G</td>
<td>200</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>28</td>
<td>rs2749894</td>
<td>119522</td>
<td>A/C</td>
<td>211</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>29</td>
<td>rs2720206</td>
<td>119787</td>
<td>C/T</td>
<td>260</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>30</td>
<td>rs2720213</td>
<td>124949</td>
<td>C/G</td>
<td>160</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>31</td>
<td>rs2251644</td>
<td>129341</td>
<td>A/G</td>
<td>280</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>32</td>
<td>rs2749896</td>
<td>132256</td>
<td>A/G</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>33</td>
<td>rs2749897</td>
<td>138855</td>
<td>A/G</td>
<td>380</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>34</td>
<td>rs2720197</td>
<td>138455</td>
<td>A/G</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>35</td>
<td>rs3742426*</td>
<td>147536</td>
<td>A/G</td>
<td>12</td>
<td>BB</td>
<td>0.3%</td>
</tr>
<tr>
<td>36</td>
<td>rs2180510</td>
<td>147694</td>
<td>A/G</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>37</td>
<td>rs2749899</td>
<td>150639</td>
<td>A/G</td>
<td>453</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>38</td>
<td>rs2749900</td>
<td>151540</td>
<td>C/G</td>
<td>365</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>39</td>
<td>rs2720198</td>
<td>151863</td>
<td>A/C</td>
<td>440</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>40</td>
<td>rs2403016</td>
<td>153341</td>
<td>C/T</td>
<td>359</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>41</td>
<td>rs2295445</td>
<td>157537</td>
<td>C/G</td>
<td>367</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>42</td>
<td>rs754596</td>
<td>160726</td>
<td>A/G</td>
<td>498</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>43</td>
<td>rs941632</td>
<td>163482</td>
<td>C/T</td>
<td>275</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>44</td>
<td>rs2093024</td>
<td>167571</td>
<td>A/C</td>
<td>530</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>45</td>
<td>rs941793</td>
<td>168955</td>
<td>C/T</td>
<td>260</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>46</td>
<td>rs110605</td>
<td>171447</td>
<td>C/G</td>
<td>420</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>47</td>
<td>rs104345</td>
<td>172220</td>
<td>C/T</td>
<td>327</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>48</td>
<td>rs13749</td>
<td>175721</td>
<td>C/T</td>
<td>430</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>49</td>
<td>rs227366</td>
<td>175929</td>
<td>A/C</td>
<td>16</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>50</td>
<td>rs941792*</td>
<td>176207</td>
<td>C/T</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>51</td>
<td>rs1127284</td>
<td>176475</td>
<td>C/T</td>
<td>380</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>52</td>
<td>rs1004903</td>
<td>176009</td>
<td>A/T</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>53</td>
<td>rs1013238*</td>
<td>180364</td>
<td>C/T</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>54</td>
<td>rs1190610</td>
<td>180468</td>
<td>A/G</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>55</td>
<td>rs1190613</td>
<td>180283</td>
<td>A/G</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>56</td>
<td>bp183293*</td>
<td>180293</td>
<td>A/T</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>57</td>
<td>rs1190614</td>
<td>180402</td>
<td>A/T</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>58</td>
<td>rs11849604*</td>
<td>180573</td>
<td>A/G</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>59</td>
<td>rs1203482</td>
<td>193407</td>
<td>C/T</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
<tr>
<td>60</td>
<td>rs1190618*</td>
<td>193460</td>
<td>C/G</td>
<td>12</td>
<td>BB</td>
<td>0%</td>
</tr>
</tbody>
</table>

**Appendix 1. Full genotypes for 60 SNPs across DYNC1H1**

* SNPs previously absent from the SNP databases which have since been added
† rs3742426 was merged with rs12895291 in May 2006
‡ no information available for this SNP in SNP databases
<table>
<thead>
<tr>
<th>Primer name</th>
<th>Forward sequence</th>
<th>Reverse sequence</th>
<th>$T_m$ (°C)</th>
<th>Amplicon size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1741155</td>
<td>ggaatttcacgggagggga</td>
<td>gacctgtgcctccagccaa</td>
<td>60</td>
<td>604</td>
</tr>
<tr>
<td>rs2720193</td>
<td>gtcgacgctgggtggtgctt</td>
<td>ggcacatcaccactcagag</td>
<td>60</td>
<td>752</td>
</tr>
<tr>
<td>rs2720194</td>
<td>gtcgacgctgggtggtgctt</td>
<td>ggcacatcaccactcagag</td>
<td>60</td>
<td>752</td>
</tr>
<tr>
<td>rs2273438</td>
<td>cagtagtgcctgctgaacctaa</td>
<td>cggctcactagggcttacc</td>
<td>60</td>
<td>444</td>
</tr>
<tr>
<td>rs2749890</td>
<td>agggaggagagaccttgtgta</td>
<td>gctgtgctgtcgtcgtggtc</td>
<td>60</td>
<td>767</td>
</tr>
<tr>
<td>rs2251644</td>
<td>ggagccctgtcattgtggtt</td>
<td>acaglggaagctgagggagac</td>
<td>60</td>
<td>401</td>
</tr>
<tr>
<td>rs3742426</td>
<td>agcataaagtgtacagctgtaa</td>
<td>gacaccacctgctcatacctt</td>
<td>60</td>
<td>589</td>
</tr>
<tr>
<td>rs2180510</td>
<td>agcataaagtgtacagctgtaa</td>
<td>gacaccacctgctcatacctt</td>
<td>60</td>
<td>589</td>
</tr>
<tr>
<td>rs941793</td>
<td>ggaaccaccttagggagatgctc</td>
<td>tgctcactctcaggaacacc</td>
<td>60</td>
<td>503</td>
</tr>
<tr>
<td>rs13749</td>
<td>tccccagtgtgtgactcactca</td>
<td>tggagacactgctcagaggtt</td>
<td>60</td>
<td>633</td>
</tr>
<tr>
<td>rs941792</td>
<td>acacaagctcggattcaagt</td>
<td>clgtgcacgltggtctcctt</td>
<td>60</td>
<td>850</td>
</tr>
<tr>
<td>rs1004903</td>
<td>acacaagctcggattcaagt</td>
<td>clgtgcacgltggtctcctt</td>
<td>60</td>
<td>850</td>
</tr>
<tr>
<td>rs10135238</td>
<td>cctgagctcaaggtgttctc</td>
<td>ttctcggggcatcaggtct</td>
<td>60</td>
<td>677</td>
</tr>
<tr>
<td>rs1190610</td>
<td>cctgagctcaaggtgttctc</td>
<td>ttctcggggcatcaggtct</td>
<td>60</td>
<td>677</td>
</tr>
<tr>
<td>rs1190613</td>
<td>ttttcagctgccaggttaa</td>
<td>ggggtgccgaaacagctgltc</td>
<td>60</td>
<td>562</td>
</tr>
<tr>
<td>rs1190618</td>
<td>cattgcccagcactatggttt</td>
<td>tctgggaaacagagggagccctc</td>
<td>60</td>
<td>458</td>
</tr>
<tr>
<td>rs2251644CBDay</td>
<td>tgaagcagaccgtcaggtgatg</td>
<td>aaaaacggcgagccgttcatacg</td>
<td>60</td>
<td>392</td>
</tr>
<tr>
<td>DYNC1H1 Exon 8</td>
<td>ttaagccctctgttgtgtcag</td>
<td>agctctgctattacagtgctt</td>
<td>60</td>
<td>1190</td>
</tr>
<tr>
<td>DYNC1H1 Exon 13 and 14</td>
<td>gttgglgaaagacatqagaaccttc</td>
<td>ctggccgtgtgctactaggtgtg</td>
<td>60</td>
<td>655</td>
</tr>
</tbody>
</table>

Appendix 2. DYNC1H1 informative primers
Publications
No association with common Caucasian genotypes in exons 8, 13 and 14 of the human cytoplasmic dynein heavy chain gene (DNCHC1) and familial motor neuron disorders

Azlina Ahmad-Annuar,¹ Paresh Shah,¹ Majid Hafezparast,¹ Holger Hummerich,¹ Abi S Witherden,¹ Karen E Morrison,² Pamela J Shaw,³ Janine Kirby,³ Thomas T Warner,⁴ Andrew Crosby,⁵ Christos Proukakis,⁴ Philip Wilkinson,⁵ Richard W Orrell,⁴ Lloyd Bradley,⁴ Joanne E Martin,⁶ and Elizabeth MC Fisher¹
No association with 3 DNHC1 exons and ALS/MA/HSP
No association with 3 DNCHC1 exons and ALS/SMA/HSP
association with 3 DNCHC1 exons and ALS SMA HSP
**Genetic Analysis of the Cytoplasmic Dynein Subunit Families**

K. Kevin Pfister*, Paresh R. Shah, Holger Hummerich, Andreas Russ, James Cotton, Azlina Ahmad Annuar, Stephen M. King, Elizabeth M. C. Fisher

**ABSTRACT**

Cytoplasmic dyneins, the principal microtubule minus-end-directed motor proteins of the cell, are involved in many essential cellular processes. The major form of this enzyme is a complex of at least six protein subunits, and in mammals all but one of the subunits are encoded by at least two genes. Here we review current knowledge concerning the subunits, their interactions, and their functional roles as derived from biochemical and genetic analyses. We also carried out extensive database searches to look for new genes and to clarify anomalies in the databases. Our analysis documents evolutionary relationships among the dynein subunits of mammals and other model organisms, and sheds new light on the role of this diverse group of proteins, highlighting the existence of two cytoplasmic dynein complexes with distinct cellular roles.

**Introduction**

Dyneins are large multi-subunit protein complexes that undertake a wide range of roles within the cell. They are adenosine triphosphate (ATP)-driven, microtubule minus-end-directed molecular motors that can be divided, based on function, into two classes: axonemal and cytoplasmic dyneins [1-7] (reviewed in [8,9]). Axonemal dyneins are responsible for the movement of cilia and flagella. Two cytoplasmic dynein complexes have been identified. The most abundant cytoplasmic dynein complex, cytoplasmic dynein 1, is involved in movements as diverse as spindle-pole organization and nuclear migration during mitosis, the positioning and functioning of the endoplasmic reticulum, the Golgi apparatus, and the nucleus, and also the minus-end-directed transport of vesicles, including endosomes and lysosomes, along microtubules and retrograde axonal transport in neurons. A second cytoplasmic dynein complex, cytoplasmic dynein 2, has a role in intracellular and intravascular transport (IFT), a process required for ciliary/flagellar assembly (reviewed in [10]).

The core of the cytoplasmic dynein 1 complex is a homodimer of two heavy chain polypeptides and associated intermediate, light intermediate, and light chain polypeptides, which are defined and named by their molecular mass and mobility in SDS-PAGE gels (Figure 1A). The protein subunits are encoded by families of at least two genes, and the expression patterns of the individual family members are different in various cell types. At least one of the light chains, DYNLL1 (LC8), has multiple cellular roles independent of its participation in a dynein complex. Cytoplasmic dynein 1 interacts with various other proteins including a second multimer, dynactin, to form the dynein-dynactin complex. Dynactin is comprised of at least seven different proteins, which together act as an adaptor that connects the cytoplasmic dynein motor to a range of cargoes (for review, see [11]). Interaction with dynactin also increases dynein motor processivity [12]. Furthermore, dynactin functions independently of dynein, anchoring microtubules at the centrosome [13]. Current evidence suggests that the second cytoplasmic dynein complex, cytoplasmic dynein 2, is also a homodimer of a distinct heavy chain, DYN2CHL1, with associated light intermediate chain, DYN2L1LI1 (Figure 1B). No other subunits have yet been identified for this complex, and it does not appear to interact with the dynactin complex [14-16].

The cytoplasmic dynein proteins are fundamental to the functioning of all cells, and have recently been shown to be causally mutated in forms of neurodegeneration [17-19]. They are thus of great interest for mammalian genetic and other studies. We therefore sought to examine the role of cytoplasmic dynein subunits from a genetic perspective.

During this analysis, we noted considerable confusion in the human and mouse gene and protein names and mapping positions. Therefore, we reexamined the mapping locations for the subunit genes and clarified and updated entries in the various sequence databases. In doing so, we utilized the revised consensus nomenclature developed for the cytoplasmic dynein subunits and their genes (Table 1). We also defined, as far as possible with current data, homologous genes in model organisms, including *Drosophila, Caenorhabditis*

---


**DOI:** 10.1371/journal.pgen.0020001

**Copyright:** © 2006 Pfister et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Abbreviations:** ATP, adenosine triphosphate; BBSRC, Biotechnology and Biological Sciences Research Council; IFT, intracellular transport; JTT, Jones, Taylor, and Thornton; MGI, Mouse Genome Informatics; Mr, relative mobility; NCBI, National Center for Biotechnology Information; NIH, National Institutes of Health; nNOS, neuronal nitric oxide synthase; PSI-BLAST, position-specific iterative BLAST; RefSeq, Reference Sequence

K. Kevin Pfister is in the Department of Cell Biology, School of Medicine, University of Virginia, Charlottesville, Virginia, United States of America. Paresh R. Shah is in the Department of Neurodegenerative Disease and in the MRC Prion Unit, Institute of Neurology, London, United Kingdom. Holger Hummerich is in the MRC Prion Unit, Institute of Neurology, London, United Kingdom. Andreas Russ is in the Genetics Unit, Department of Biochemistry, University of Oxford, Oxford, United Kingdom. James Cotton is in the Department of Zoology, the Natural History Museum, London, United Kingdom. Azlina Ahmad Annuar and Elizabeth M. C. Fisher are in the Department of Neurodegenerative Disease, Institute of Neurology, London, United Kingdom. Stephen M. King is in the Department of Molecular, Microbial, and Structural Biology, University of Connecticut Health Center, Farmington, Connecticut, United States of America.

* To whom correspondence should be addressed. E-mail: kpflw@virginia.edu
elegans, Chlamydomonas, and yeast. To further our understanding of the function of cytoplasmic dynein subunits, we also briefly examined mutations in this group of proteins in a variety of model organisms. We do not discuss dynein-binding proteins such as dynactin, LIS1, or various kinases, which while important for dynein function, have not yet been shown to be stoichiometric components of the cytoplasmic dynein complex.

Human and mouse cytoplasmic dynein subunit genes. The subunits of the cytoplasmic dynein complexes are resolved into subunit polypeptides of ~530 kDa (heavy chains), ~74 kDa (intermediate chains), ~33–59 kDa (light intermediate chains), and ~10–14 kDa (light chains) in SDS-PAGE gels (Figure 1A). Research on the cytoplasmic dynein subunits has been undertaken in a wide range of organisms from yeast to humans. The nomenclature of the mammalian genes encoding these proteins has drawn on homologs in other organisms and, consequently, a number of aliases have been found for any given human or mouse cytoplasmic dynein subunit. Much of the early research into dynein genetics was conducted in the biflagellate green alga Chlamydomonas on the dyneins found in the flagellar axoneme, and therefore some cytoplasmic dynein nomenclature derives from these studies. For example, mammalian members of the cytoplasmic light chain families DYNLRB and DYNLL have commonly been referred to as LC7 and LC8, respectively, which are the names of homologous Chlamydomonas axonemal dynein subunits.

Nomenclature. The revised classification system for mammalian cytoplasmic dynein (Table 1) recognizes the two distinct dynein complexes, cytoplasmic dynein 1 and cytoplasmic dynein 2, and the fact that cytoplasmic dynein light chains are shared with some axonemal dyneins. Cytoplasmic dynein subunits are also classified into polypeptide families according to sequence similarity within groups of similarly sized proteins; thus there is sequence similarity within the dynein gene families (and when cytoplasmic and axonemal members of the same gene families are compared) but not among them.

This nomenclature has been approved by the Human Genome Organization Nomenclature Committee (20) and the International Committee on Standardized Nomenclature for Mice. In accordance with their policy, the designation of each unique cytoplasmic dynein subunit starts with DYNC for dynein, cytoplasmic, followed by the specific dynein complex subtype 1 or 2; for example, cytoplasmic dynein 2 is designated DYNC2. The shared light chains start with DYN. Each subunit is designated with a letter(s) for the size of the polypeptides, H for the heavy chain, I for the intermediate chain, L1 for the light intermediate chain, and L for the light chain. Additional letters (T, RB, and L) are used to distinguish the three distinct light chain families as described in the text. Individual members of the gene families are assigned numbers. Standard human and mouse gene nomenclature is used: italicized upper case for human gene symbols (for
<table>
<thead>
<tr>
<th>Cytosolic Dynine Gene Family</th>
<th>Official or Proposed Aliases</th>
<th>Location: Human (Hs)</th>
<th>Entrez Gene ID* (Human and Mouse)</th>
<th>mRNA (NCBI RefSeq Accession Numbers)</th>
<th>Protein (NCBI/ SwissProt** Accession Numbers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytosolic dynine 1 heavy chain</td>
<td>Human DYNC1H1</td>
<td>DNCHC1 [177]; DNECL [178]; Hp22 [43]; pm7 [98]; Rk3-8 [98]; DHC1 [23]; DHC1a [61]; MAPIC [3]; DNCL [20]; HL-3 [98]; KIAA0325 [21]; AB002333 [179]; Dyh1 [180]; Cdc4 [44]; DYNH_HUMAN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse Dynch1</td>
<td>Dnch1 [98]; Cdc4 [44]; Loa [181]; Rk3-3 [98]; Dnc1 [205]; Dnc2 [182]; MAPIC [41]; mKIAA0325 [183,184]; DYNH_Mouse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytosolic dynine 2 heavy chain</td>
<td>Human DYNCH2H1</td>
<td>DHC2 [23,74]; DHC1b [74]; DLP4 [56]; DYN1B [8]; Dhy2 [180]; hDhc1 [185]; FLJ11756 [186]; DYNHC2 Mouse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse Dynch2H1</td>
<td>Dnc2H1 [187]; Mdhc11 [185]; DYNHC2_Mouse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytosolic dynine 1 intermediate chain</td>
<td>Human DYNC1I1</td>
<td>IC1 [79]; IC7 [74]; IC4-1 [40]; D1 IC7 [14]; DH IC1-1 [188]; Dnc1I1 [94]; DYN1I1_HUMAN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse Dynch1I1</td>
<td>Dnic1 [187]; Dnic1I [94]; DYN1I1_Mouse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytosolic dynine 1 light intermediate chain</td>
<td>Human DYNC1I2</td>
<td>IC2 [79]; IC7 [74]; IC7-2 [189]; DYN1I2_HUMAN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse Dynch1I2</td>
<td>Dnic2 [187]; Dnic2I [94]; DYN1I2_Mouse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytosolic dynine 2 light intermediate chain</td>
<td>Human DYNC2I1</td>
<td>D2LIC [14]; UC3 [15]; CG60 [15,192]; DKS3P64AA033 [192]; D2LIC_HUMAN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse Dynch2I1</td>
<td>4933400401Rik [193]; D2LIC [193]; mD2LIC [193]; MGCL211 [193]; MGCL0064 [193]; D2LIC_Mouse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytosolic dynine 3 light chain</td>
<td>Human DYNLT1</td>
<td>TCTE1 [194]; Tectx1 [114,116]; Protein CW-1 [195]; DYNLT_HUMAN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse Dynlt1</td>
<td>Tctex1 [116]; Tctex1 [114,116]; DYLX_MOUSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytosolic dynine 4 light chain Roadblock</td>
<td>Human DYNLRB1</td>
<td>MGC15113 [197]; DNLCA2 [142] (bithoraxoid-like protein [135,198]; Bith [199]; Khc2 [22]; mLc7-1 [144]; Robb2 [143]; Robb2 [22]; Robb2 [144]; DLYA_HUMAN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse Dynrb1</td>
<td>Hnrc2a [187]; Hnrd2a2 [187]; mLc7-1 [144], 2010012151Rik [196]; 20103201M1Rik [196]; DLYA_Mouse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytosolic dynine 5 light chain LC8</td>
<td>Human DYNLL1</td>
<td>DCL1 [139]; DNLCL1; DCL1 [202]; M, 8000 LC [104]; M, 8000 DCL [147]; DCLB [152] [LC (22); LC [48]; PIN [153]; hDc [162]; D2K [203]; DYN1I1_HUMAN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse Dynll1</td>
<td>Dnc1 [187]; DYNLL1_Mouse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytosolic dynine 6 light chain LC8</td>
<td>Human DYNLL2</td>
<td>DCL2 [149]; LcBb [148]; MGC7810 [204]; DCL2 [187]; 1700464A15Rik [205]; 6720463E02Rik [205]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse Dynll2</td>
<td>Dnc2 [187]; 1700464A15Rik [205]; 6720463E02Rik [205]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*NCBI* preffices are from the National Center for Biotechnology Information (NCBI) as part of an initiative to identify and sequence *C. elegans* genes containing a full-length open reading frame for human, mouse, and rat (206). "CGD" prefixes are assigned by the Comparative Gene Identification study, which uses the C. elegans proteome as an alignment template to assist in novel human gene identification from human EST databases (207). "Rs" suffixes are gene aliases assigned by the Riken Genomic Sciences Center (http://genome.gsc.riken.jp). Source: Pfister et al (2010).
example, DYNC1H1), italicized initial upper case and then lower-case letters for mouse (DynClH1), and for proteins of both species, the same symbols in upper case, upright (DYNC1H1). In accordance with the International Union of Pure and Applied Chemistry standards, isoforms of the intermediate chain gene products are referred to with letters. This nomenclature system can be expanded to other subunits as appropriate. We refer to mapping positions using the prefixes Hsa (Homo sapiens) for human and Mmu (Mus musculus) for mouse, followed by the chromosomal localization e.g. Hsa2q11, Mmu11.

Table 1 lists the aliases, map position, and protein/DNA-sequence accession data for each known mouse and human cytoplasmic dynein gene. The greatest number of aliases was observed for the cytoplasmic dynein 1 heavy chain 1 (DYNC1H1) for which we identified 15 different names. Some alternative cytoplasmic dynein gene names have come from large-scale gene and transcript identification efforts such as the partial DYNC1H1 clone KIAA01925 and its mouse homolog “mKIAA00925,” generated by the Kazusa cDNA project [21]. A small number of gene names have been derived from the names of DNA markers and cDNA clones used to identify the genes, for example, cytoplasmic dynein 2 light intermediate chain 1, DYNCL2L1, was named DKEF2054A03S after the cDNA sequence and clone of the same name. The heavy chain gene DYNC1H1 has also been referred to by the name of a marker, Hp22, generated from its human cDNA sequence, as well as the rat-derived marker Rk3-8 and a cDNA clone named HL-3.

Cytoplasmic Dynein Heavy Chain Gene Family (DYNC1H1, DYNCL2H1)

Figure 2A shows the phylogenetic relationships amongst the dynein heavy chain protein sequences from various organisms. The heavy chain sequences fall into two distinct clades, and the relationships within each clade are generally consistent with known evolutionary distances between the organisms shown. We note that our phylogeny fits well with and extends previous phylogenetic analyses of the heavy chain proteins [22,23]. This analysis indicates that the partial human sequence DNAH12, (AA809729) [23], is unlikely to be a cytoplasmic dynein.

Cytoplasmic dynein heavy chain 1, DYNC1H1. DYNC1H1, cytoplasmic dynein 1 heavy chain 1, is the largest cytoplasmic dynein subunit, having ~4,600 residues and a molecular weight of >530 kDa. First identified in rat spinal cord and brain and termed Microtubule Associated Protein 1C (MAP1C) [24], DYNC1H1 is a distant member of the AAA family of ATPases and is the cytoplasmic counterpart to axonemal dynein heavy chains [3,25]. DYNC1H1 associates as a homodimer within the cytoplasmic dynein complex and effects the contact and translocation of the dynein complex along microtubules via its large motor domain [8,26] (Figure 1B).

The C-terminal region of DYNC1H1 is the motor domain of the dynein complex and is conserved in all cytoplasmic and axonemal dynein heavy chains. This region is arranged as a heptamer ring with six AAA domains and a seventh domain, the identity of which remains a matter of discussion (Figure 1B) [12,25,27,28]. AAA domains are regions of ATP binding and hydrolysis, and thus they generate the energy required for translocation [29–31]. While the first AAA domain is essential for motor activity [32], reviewed in [30], the first four AAA domains are potentially capable of binding and hydrolysing ATP [33–35]. Contact of the heavy chain with a microtubule is established via an ~15-nm projection that extends between the fourth and fifth AAA domains [28,56]. The N-terminal region of DYNC1H1 is known as the stem, and force production, and therefore translocation, is thought to be achieved through the contact and shift of a 10-nm fold of the stem closest to the first AAA domain [37]. DYNC1H1 dimerization also occurs in the stem, and the intermediate chains and light intermediate chains bind in this region as well [38,39]. The three light chains bind to the intermediate chains [40]. Collectively the five smaller dynein subunits that bind to the N-terminus of DYNC1H1 make up the cargo-binding portion of the dynein complex.

The sequence of full-length mammalian DYNC1H1 was first obtained in rat and mouse [41,42]. Human DYNC1H1 was identified by screening an adenocarcinoma library with a partial human cDNA [25,43]. As yet, the only mutations reported in mammalian heavy chains have been in the mouse: the Lsa and Crod strains have allelic point mutations in Dync1h1 that cause late-onset motor neuron degeneration in heterozygotes and neuronal apoptosis in homozygotes [17]. The loss of both copies of Dync1h1 has been shown to be lethal during early embryonic development, with disorganization of the Golgi complex, improper distribution of endosomes and lysosomes, and defects in cell proliferation; no phenotype has yet been reported for heterozygote knock-out mice [44].

In Drosophila, the dynein heavy chain gene, Dhc64C, functions in oogenesis [45,46], oocyte differentiation [47], centrosome attachment during mitosis [48], eye development, cell development in thorax, abdomen, and wing [45], and axonal transport [49]. Homozygous mutations induced by the mutagen ethyl methane sulfonate in Dhc64C are larval-pupal lethal, whilst heterozygotes have defects in bristle formation, eye development, and fertility [45]. In C. elegans, dynein heavy chain (dhe-1) is an essential gene, also known as let-354 (LEThal) [50]. Extensive mutational analysis has been conducted on dhe-1 to produce a range of variants from recessive/dominant lethals to temperature-sensitive mutants. The resultant phenotypes invariably include embryonic lethality, spindle orientation defects, polar body abnormalities, and excessive blebbing in the early embryo [51–54].

In the yeast Saccharomyces cerevisiae, heavy chain function ensures the alignment and orientation of mitotic spindles. Mutation of the S. cerevisiae heavy chain gene dyn1, which has 50% similarity (28% identity) to DYNC1H1 over 80% of the protein’s length, has been shown to disrupt spindle orientation and reduce the fidelity of nuclear segregation during mitosis [55,56]. Despite this phenotype, dyn1 mutations remain viable, although dyn1 and kinesin double mutants are lethal [57]. This observation suggests some functional redundancy for dynein by kinesin motors in yeast. No cytoplasmic dynein 1 heavy chain 1 homolog has unambiguously been identified in Chlamydomonas, and neither have dyneins been found in either the Arabidopsis or rice genomes [58,59] (reviewed in [60]). There are many dynein heavy chains in the Chlamydomonas genome. However, with the exception of DYN2H1, they appear to be components of the axonemal dyneins.
**Cytoplasmic dynein 2 heavy chain 1, DYNC2H1.** The cytoplasmic dynein 2 heavy chain, DYNC2H1, was originally identified in sea urchin embryos by Gibbons and colleagues and was termed DYN1b [61]. It is much less abundant than DYN1H1 and does not appear to heterodimerize with DYN1H1; biochemical analyses suggest that DYNC2H1 is a homodimer [16]. DYN2H1 contains regions characteristic of cytoplasmic dyneins, for example, human DYNC1H1 and DYNC2H1 sequences are similar within both the motor region and around the light intermediate chain–binding site [15]. However, the expression of its mRNA increases during embryonic reciliation, a property typical of axonemal dyneins, suggesting a flagellar role for an otherwise cytoplasmic-like dynein heavy chain. The flagellar properties of DYNC2H1 were clarified with its identification as the motor responsible for retrograde (tip to base) IFT, in *Chlamydomonas*, a process required for assembly and maintenance of the eukaryotic cilia/flagellum [6,22].

### Figure 2. Panel Showing the Protein-Based Phylogenies of the Cytoplasmic Dynein Subunit Families

Species names are shown with NCBI/GenBank gene/protein names. NCBI/GenBank protein-sequence accession numbers are given in Table S1. Orthologous human, mouse, and rat gene names use the revised systematized consensus nomenclature (e.g. DYN1H1 in humans, mouse, and rat).

Relationships amongst dynein sequences of different species do not necessarily reflect the evolutionary relationships amongst species; see [208] and [209] for further details. Named clades are indicated in the right margins. Bayesian and maximum-likelihood bootstrap values are shown as percentages (top and bottom, respectively), adjacent to branch points. Asterisks denote bootstraps below 50%. Filled circles denote bootstraps at 100%. Scale-bar represents evolutionary distance (estimated numbers of amino-acid substitutions per site).

- **A** Cytoplasmic dynein light chain 2 family. *Chlamydomonas* outer arm heavy chain (ODA11) is used as the outgroup. DNAH12frag is the partial axonal heavy chain fragment taken from [23]. For mouse DNAH21, XP_35830, only partial protein sequence (336aa) was available in the GenBank database. Adding this partial sequence to our analysis resulted in spurious clustering, therefore we obtained an extended, putative sequence by using BLAST (BLASTN) against the mouse genome (Build 32) with human and rat sequences XP_370652 and NP_075413, respectively. Incomplete mouse genomic assembly at the DNAH21 locus yielded a truncated sequence 3455 amino acids in length, 85% the length of human DNAH21.
- **B** Cytoplasmic dynein intermediate chain family. The *Chlamydomonas* IC2 (ODA6) is used as the outgroup.
- **C** Cytoplasmic dynein light intermediate chain family. There does not appear to be a sufficiently distant homolog in *Chlamydomonas* to be used as an outgroup in this analysis, therefore ODA11 (Q39610, a heavy chain protein) was chosen as the outgroup for this tree.
- **D** Cytoplasmic dynein light chain family. The *Chlamydomonas* LC2 light chain is used as the outgroup.
- **E** Cytoplasmic dynein light chain Roadblock family. The *Chlamydomonas* outer arm dynein LC7a, is used as the outgroup.
- **F** Cytoplasmic dynein light chain family. The *Chlamydomonas* Q39579 sequence is used as an outgroup. This phylogeny is poorly resolved, with low bootstrap support values and posterior clade probabilities, most likely due to there being little variation amongst the ingroup sequences. We found good support for the LC8 light chain 1 clade, and some support for the LC8 light chain 2 clade, of four vertebrate sequences. The relationships of the two sequences, *C. elegans* and *Takifugu* were poorly resolved, and therefore we have not included these in the LC8 light chain 2 clade.
DYNC2H1 is also important in modified ciliary structures such as nematode mechanosensory neurons [62] and vertebrate photoreceptors [63,64]. In C. elegans, the DYNC2H1 homolog, che-3, is expressed in ciliated sensory neurons which are thought to be involved in odorant chemotaxis [65]. Mutations of che-3 affect IFT, the establishment and maintenance of sensory cilia, which are stunted and swollen in the mutants, [62,66], chemotactic behavior [67], and formation of the third larval stage, dauer formation [68].

The first mammalian DYNC2H1 gene was described in rat, designated DLP4 [69], and full-length sequence has been obtained [15]. Genetic and biochemical studies suggest that DYNC2H1 associates with a member of the light intermediate chain family, DYNC2LI1 (Figure 1B, and see discussion of the light intermediate chain family below) [14,16,70–73] and possibly also with DYNL1L1 (LC8) light chain [72]. In mice Dync2h1, mRNA is abundant in the olfactory epithelium and comprised of vertebrate species only. An alternative dynein intermediate chain encoded by the ODA6 locus supports against 51% support). In view of this and with all cytoplasmic dynein 1 protein phylogeny. The intermediate chain sequences fall into two distinct clades, intermediate chains 1 and 2, non-vertebrate species falling outside these clades, the data indicate a role for DYNCI in targeting the dynein complex to various cargo, including membranous organelles and kinetochores [76,79,89]. These data also support a role for DYNCl in the generation and maintenance of mammalian cilia [14,16].

Intermediate chains are present in axonemal and/or cytoplasmic dyneins from yeast to mammals (Figure 2B). Protein-sequence data demonstrate evolutionary distant relationships between axonemal and cytoplasmic dynein intermediate chains; for example, rat DYNC1I1 has 48% similarity to the Chlamydomonas IC2 axonemal outer arm dynein intermediate chain encoded by the ODA6 locus [75,76]. Figure 2B shows the dynein intermediate chain protein phylogeny. The intermediate chain sequences fall into two distinct clades, intermediate chains 1 and 2, comprised of vertebrate species only. An alternative placement of a Takifugu sequence, as a member of the intermediate chain 1 clade, is almost as well supported by the data as the placement shown in Figure 2B (49% bootstrap support against 51% support). In view of this and all non-vertebrate species falling outside these clades, the data suggest a recent evolutionary origin for the split into intermediate chain gene 1 and intermediate chain gene 2, perhaps as part of a "2R" event of genome duplication (see [77] for review). The absence of an amphibian (Xenopus) intermediate chain 1 protein may be due to the current paucity of X. laevis sequences in the GenBank sequence database (http://www.ncbi.nlm.nih.gov/Genbank).

The cytoplasmic dynein 1 intermediate chains have a molecular weight of ~74 kDa [5] and associate in the cytoplasmic dynein complex with a stoichiometry of two intermediate chains per complex [40,78]. DYNC1I1 and DYNCl2 proteins are thought to help assemble the cytoplasmic dynein complex and to bind various cargoes. The intermediate chains interact with the dynein activator, dynactin, via their conserved N-termini [79]. The DYNCl C-termini contain a WD repeat domain [76,80,81] that is conserved between cytoplasmic and axonemal intermediate chains and is important for intermediate chain-binding to the heavy chains [76,82]. The dynein light chains, DYNLL1 (LC8) and DYNLT1 (Tctex1), bind near the N-termini of the intermediate chains [83–85], and the DYNLRB (Roadblock) light chains bind just upstream of the WD repeat region [40]. The DYNCl are phosphorylated, and phosphorylation at one site regulates DYNCl2 interaction with the p150 subunit of dynactin [86,87].

The Chlamydomonas IC2 axonemal intermediate chain was localized to the base of the dynein heavy chain dimer by immunoelectron microscopy [88]. Steffen and colleagues identified a similar location for the cytoplasmic dynein intermediate chain and found that antibodies to it block dynein binding to membrane-bound organelles [89,90]. These data also suggest a role for DYNCl in targeting the dynein complex to various cargo, including membranous organelles and kinetochores [76,79,89]. In Drosophila, mutations in dynein intermediate chain, Cdic (also referred to as Cdic and Dic), lead to larval lethality, demonstrating that this intermediate chain provides an essential function. Cdic mutations dominantly enhance the rough-eye phenotype of Glued, a dominant mutation in the p150 subunit of dynactin [91].

The Chlamydomonas IC2 axonemal intermediate chain was localized to the base of the dynein heavy chain dimer by immunoelectron microscopy [88]. Steffen and colleagues identified a similar location for the cytoplasmic dynein intermediate chain and found that antibodies to it block dynein binding to membrane-bound organelles [89,90]. These data also suggest a role for DYNCl in targeting the dynein complex to various cargo, including membranous organelles and kinetochores [76,79,89]. In Drosophila, mutations in dynein intermediate chain, Cdic (also referred to as Cdic and Dic), lead to larval lethality, demonstrating that this intermediate chain provides an essential function. Cdic mutations dominantly enhance the rough-eye phenotype of Glued, a dominant mutation in the p150 subunit of dynactin [91].

We note that in Drosophila, the Cdic gene lies in the 19DE region of the X chromosome, adjacent to several dynein intermediate chain-like sequences. These sequences are derived from a 7-kb duplication/deletion event involving Cdic and its proximal gene annexin X, which encodes a cell-surface-adhesion protein [92]. The duplication/deletion of this 7-kb region resulted in the formation of a de novo coding sequence, under the control of a testes-specific promoter, called sperm-specific dynein intermediate chain gene (Sdc) [93]. The de novo region has undergone at least 10-fold tandem duplication, which has given rise to a multi-gene family comprising at least four classes of Sdc gene, of which more than one class is functional [93].

Cyttoplasmic dynein 1 intermediate chain 1, DYNCl1.

Multiple DYNCl1 isoforms exist in mammals. They are the products of alternative splicing of the N-terminal region of a single DYNCl1 gene and phosphorylation [76,79,86]. In humans, alternate splicing may arise from cryptic splice-acceptor sites located within exon 4 of this 17-exon gene [94]. Two DYNCl1 isoforms were found in rat brain and DYNCl1 mRNA, and protein isoform expression is regulated during rat brain development, and a single DYNCl1 isoform is found in testis. DYNCl1 expression is also cell-specific; cultured rat neurons express at least two DYNCl1 alternative splicing variants and their phosphorylated isoforms, while cultured glial astrocytes do not express any DYNCl1 gene products [95–97]. In the mouse, expression of Dync111 has been shown to be restricted primarily to the brain, with weak expression in testis [94], further supporting possible neuronal specificity for Dync111. As with the other dynein subunits, the isoform diversity of the intermediate chain is thought to result in specific populations of dynein...
molecules that have specific functions; for example, both DYNC1I1 isoforms are components of cytoplasmic dynein found in the slow component of axonal transport in the optic nerves [95]. Multiple isoforms of the *Drosophila* intermediate chains are also produced by alternative splicing of the single gene [92].

**Cytoplasmic dynein 1 intermediate chain 2, DYNC1I2.** Vaughan and Vai Lee used a partial human cDNA sequence with identity to the already known DYNClI gene as a probe to isolate a rat DynClI cDNA; predicted human and rat DYNC1I2 sequences are 94% identical [79], and the existence of two genes was supported by mapping data that placed DynClI and DynClI2 at distinct loci within the mouse genome [98]. Like DynClI, DynClI2 produces different splice isoforms: alternative splice sites lie at two positions within the N-terminal region. The expression of DynClI2 isoforms is ubiquitous with the rat DYNC1I2C isoform being expressed in all tissues and cells examined [79,94,96,97]. During rat brain development, DYNC1I2C is the only isoform found before E14 (embryonic day 14) and is often the only isoform observed in cultured cells [96,99]. During nerve growth-factor stimulation of PC12 cell differentiation and neurite extension, there is a change in relative expression levels of the DYNC1I2 isoforms [100]. In the rat optic nerve, it has been shown that the DYNC1I2C isoform is the only intermediate chain involved in the fast component of anterograde transport to the axon tip [95,99].

The strong expression of DynClI2 in the mouse developing limb bud led to the suggestion that DYNC1I2 may play a role in limb development and digit patterning and/or in establishing cell polarity [94]; dynein may not do this directly, but may mediate these processes by orientating intracellular components correctly [101].

**Cytoplasmic Dynein Light Intermediate Chain Gene Family (DYNC1LI1, DYNC1LI2, DYNC2LI1)**

Figure 2C shows the phylogenetic relationships amongst the dynein light intermediate chain protein sequences from various organisms. The light intermediate chains can be separated into three distinct groups: the two light intermediate chains that are components of cytoplasmic dynein 1, DYNC1LI1 and DYNC1LI2, are more closely related to each other than to the cytoplasmic dynein 2 light intermediate chain, DYNC2LI1. Hughes and colleagues first proposed the name light intermediate chains for these subunits [102], although these polypeptides were also referred to as light chains [103] prior to the discovery of the smaller light chains [104]. The mammalian cytoplasmic dynein complex contains four species with molecular masses of 50–60 kDa that resolve into numerous isoforms on 2D gels [86,102]. The multiple isoforms observed in 1D and 2D gels are thought to be the result of post-translational phosphorylation, although the possibility of alternate splicing has not been eliminated [86,102,103]. A third gene, *Dync2li1*, has recently been described which encodes a protein that appears to exclusively associate with DYNC2H1 in the cytoplasmic dynein 2 complex [14–16,71]. Unlike the other subunits of cytoplasmic dynein, homologs of the DYNC1IIs have not yet been identified in the axonemal dyneins [105]. The function of the DYNClIIs has yet to be determined, although it has been suggested that they may regulate the interactions of dynein with dynactin, or with sub-cellular cargoes of dynein-mediated motility. DYNC1LI1 and DYNC1LI2 form only homo-oligomers, and their mutually exclusive binding to the N-terminal base of the dynein heavy chain is consistent with a role in cargo binding [38].

*C. elegans* appears to have one light intermediate chain (*dli-l*) for cytoplasmic dynein (DYNC1LI1-based complexes), and one (*xbr-1*) for cytoplasmic dynein 2 (DYNC2H1-based complexes) [79]. *dli-l* is required for dynein function during mitosis, promonuclear migration, centrosome separation, and centrosome association with the male pronuclear envelope [106], as well as retrograde axonal transport. Mutations in *dli-1* lead to an accumulation of cargo at axonal terminals [52]. Disruption of *xbr-1* results in ciliary defects and causes behavioral abnormalities that are observed in other cilia mutants [14]. Binding of *dli-1* to ZYG-12 is thought to be the mechanism for dynein binding to the nuclear envelope [107].

**Cytoplasmic dynein 1 light intermediate chain 1, DYNC1LI1.** DYNC1LI1 was cloned from rat [38] and found to have a P-loop motif, which is one of the major conserved motifs making up the nucleotide-binding domain found in numerous proteins, including ATPases and kinases [108]. DYNClI1, however, lacks other essential motifs associated with ATPase activity, which itself has not been assayed. Tyan showed that pericentrin, a known dynein cargo, binds DynClI1 and not DYNC1LI2 [38]. DynClI1 and its phospho-isoform are exclusively found with dynein in the slow component of axonal transport in rat optic nerves [95]. In HeLa cells, DynClI1 localizes to the microtubule organizing centre and mitotic spindle, co-localizing with the GTPase Rab4a (which interacts with the central domain of DYNClI1) [109]; thus DYNClI1 may be implicated in the regulation of membrane-receptor recycling. Phosphorylation of the *Xenopus* DynClI1 has been implicated in regulation of dynein binding to membrane-bound organelles [110]. It is thought that *Xenopus* melanosomes contain a distinct dynein light intermediate chain protein, possibly a version of DYNClI1 [111]. In the chicken (Gallus gallus) DYNClI1 has been called DLC-A, as part of the DLC-A group of light chains [103].

**Cytoplasmic dynein 2 light intermediate chain 1, DYNC2LI1.** DYNClI2 is paralogous to DYNClI1 and is also thought to be post-translationally modified by phosphorylation [86,102,103]. DYNClI2 is found in both the fast and slow components of axonal transport in rat optic nerves, although its phospho-isoforms are found only in the slow component of axonal transport. During nerve growth-factor stimulation of PC12 cell differentiation and neurite extension, DynClI2 expression is up-regulated [112], and phosphorylation of both DYNClI1 and DYNClI2 is increased [100]. Like DynClI1, the chicken (G. gallus) DYNClI2 has also been termed DLC-A, as part of the DLC-A group of light chains [103].

**Cytoplasmic dynein 2 light intermediate chain 1, DYNC2LI1.** DYNClI2 is a light intermediate chain that was identified in mammals by two groups and was originally designated D2LIC [14] and LIC3 [15]. DYNClI1 is the light intermediate chain that associates with DYNClH1 in the cytoplasmic dynein 2 complex: Grissom and colleagues observed that DYNClI1 co-immunoprecipitated specifically with DYNClH1 and co-localized with DYNClH1 at the Golgi
apparatus. Mikami and coworkers [15] found the 350-amino acid LC3 polypeptide (AAD34055) had a 24% similarity to rat DYNCL1B but failed to observe Golgi localization. DYNC2LI1 has been identified in mouse, C. elegans, Drosophila, and Chlamydomonas [14, 16]. A targeted deletion of Dyncl211 in mouse affects development, in particular ventral cell fates and axis establishment in the early embryo [113]. In Chlamydomonas, DYNCL2LI1 (DlBLC) is essential for retrograde IFT [71]. As mentioned above, DYNC2LI1 appears to bind exclusively with DYNCL2H1; in agreement with this, we find DYNC2LI1 homologs in species that have DYNC2LI1. The exclusive association of DYNC2LI1 and DYNC2LI1 with one another, and not with any of the other cytoplasmic dynein subunits, emphasizes the distinct cellular identities and roles of these separate DYNC1LI1 and DYNC2H1 dynein complexes.

**Cytoplasmic Dynein Light Chain Gene Families**

There are three known dynein light chain gene families that are components of cytoplasmic dynein 1: (1) the t-complex-associated family (DYNLT1, DYNLT2), (2) the Roadblock family (DYNLRB1, DYNLRB2), and (3) the LC8 family (DYNLL1, DYNLL2). The gene families are named according to their original discovery, through the effect of mutations in mouse (t-complex associated, Tctex1) and Drosophila (Roadblock), or according to the size of the protein in Chlamydomonas (LC8) as discussed below. We present each family by molecular weight, starting with the largest light chain protein family, the t-complex-associated family (~115 amino acids), through to the Roadblock family (~96 amino acids) and the smallest light chain proteins, the LC8 family (~89 amino acids). As described below, some of the light chains have cellular functions that are independent of their role in the cytoplasmic dynein 1 complex.

**Cytoplasmic Dynein Light Chain Tctex1 Family Gene (DYNLT1, DYNLT3)**

Figure 2D shows the phylogenetic relationships amongst the dynein light chain Tctex1-family protein sequences from various organisms. Our phylogeny shows distinct clades for DYNLT1-like and DYNLT3-like sequences. Tctex2-like sequences lie closer to the outgroup than they do to the DYNLT1 and DYNLT3 clades (not shown).

**Cytoplasmic dynein light chain Tctex1, DYNLT1.** Tctex1 (t-complex testis-expressed) gene was originally identified within the mouse t-complex (a 30- to 40-Mb region of Mmu17) as a candidate for one of the "distorter" proteins responsible for the non-Mendelian transmission of variant haplotypes [114]. Lader et al. [114] and O'Neill and Arzt [115] found evidence of four copies of Dintit (Tctex1) in the mouse genome; we found that the current genomic sequence databases appear to contain only one such locus that maps to Mmu 17, although a processed pseudogene has also been described on Mmu8. Subsequently, DYNLT1 was found to be an integral component of cytoplasmic dynein [116], and has since also been identified within axonemal inner and outer arm dyneins [117, 118]. DYNLT1 binds to the N-terminus of the intermediate chain DYNCL1 [85]. Many studies have identified DYNLT1 as a binding partner for various cellular proteins, and it has been suggested that it may attach specific proteins or cellular components to cytoplasmic dynein; for example, DYNLT1, but not its homolog DYNLT3 (see below), binds to the C-terminal domain of rhodopsin and is required for the trafficking of this visual pigment within photoreceptors [119]. The two DYNLT1 polypeptides in the cytoplasmic dynein complex dimerize, and their dimer structure is similar to that of the DYNLL1, LC8, dimer [116, 120–122]. The evidence suggests that the same Dynth1 gene product is a component of both axonemal and cytoplasmic dyneins in mouse [117]. The binding site on DYNCL1 for DYNLT1 has been mapped to a 19-amino acid region at the N-terminus [85].

The Schizosaccharomyces pombe DYNLT-like gene SPAC1805.08 (also referred to as Dlc1) is involved in movement of nuclear material during meiotic prophase and is expressed in astral microtubules and microtubule-anchoring sites on the cell cortex. The Dlc1 localization pattern is similar to that of cytoplasmic dynein heavy chain Dhc1 [123]. Dlc1 null mutants are viable but have irregular nuclear movement during meiosis and defects in sporulation, recombination, and karyogamy [123]. Genetic analyses in Drosophila, which appears to have only one member of the DYNLT family, suggest that DYNLT1 is not essential for cytoplasmic dynein function, as the null mutation is not lethal. However, the mutants do have sperm-motility defects, suggesting they do have an essential role in axonemal dynein [124, 125]. In Chlamydomonas, Tctex1 is an axonemal inner arm dynein component [117], and recently a variant form has been identified in axonemal outer arm dynein (DiBella et al., in press).

**Cytoplasmic dynein light chain, DYNLT3.** Closely related to DYNLT1 is DYNLT3, also known as tpc3 because it was initially a candidate for causing X-linked retinitis pigmentosa type 5 [126]. However, the actual gene that is defective in this disease was later identified as a guanine nucleotide exchange factor that is unrelated to DYNLT3 [127]. Subsequently, King and colleagues found that DYNLT3 is a cytoplasmic dynein light chain that is differentially expressed in a cell- and tissue-specific manner [78, 116]. Interestingly, while many proteins have been identified as binding partners for DYNLT1, none have been identified as binding exclusively to DYNLT3, though recently the *Herpes simplex* virus capsid protein VP26 has been shown to bind both DYNLT1 and DYNLT3 [128]. There is no evidence that DYNLT3 is a component of axonemal dyneins.

**Axonemal dynein light chain, Tctex2.** To avoid confusion with Tctex2, an axonemal dynein subunit, the DYNLT2 designation is not used: a third human t-complex testis-expressed gene, originally characterized by Rappold and colleagues [129, 130], was given the name Tctex2, and is also known as LC2, TCTE3, and Ted3. Patel-King and colleagues demonstrated that it has 35% identity to the 19,000-M<sub>r</sub>, (relative mobility) axonemal outer arm dynein light chain (LC2) of Chlamydomonas [131], and that it is distantly related to cytoplasmic light chains DYNLT1 and DYNLT3 [116, 120]. LC2 is essential for outer arm dynein assembly [132]. There is evidence that Tctex2 may interact substoichiometrically with cytoplasmic dynein, but there has not yet been a definitive demonstration that it is a cytoplasmic dynein subunit.

In mice, expression of Tctex2 is testis-specific, particularly in later spermatogenic stages, and isoforms are thought to be generated by alternative splicing [130]. As yet, isoforms of the human homolog have not been identified, and its expression...
is restricted to tissues containing cilia and flagella [133].

Mutations in Tctex2 have been implicated in the autosomal recessive disorder primary ciliary dyskinesia, which results in the impairment of ciliary and flagellar function, although these mutations are thought not to be the primary cause of the disorder [133].

Mouse Tctex2 lies within the Mmu17 t-complex in a central region containing the distorter/sterility locus Tid3 [134]. Human Tctex2 maps to the long arm of Chromosome 6 [129] and, interestingly, is a neighbor of the two genes, TCP1 and TCP10, which are also homologs of mouse t-complex loci found adjacent to mouse Tctex2. This conservation of gene order suggests that the region of Chromosome 6q containing these genes is syntenic to the homologous central region of mouse Chromosome 17. In contrast, DYNTL1 and DYNTL3 are located on human Chromosome 6p and show synteny to the distal portion of the mouse t-complex, suggesting that the middle and distal portions of the mouse t-complex are syntenic to the long and short arms of human Chromosome 6, respectively [129].

**Cytoplasmic Dynein Light Chain Roadblock Gene Family (DYNLRB1, DYNLRB2)**

The first Roadblock gene was identified in *Drosophila* through mutational analyses, and from biochemical and sequence comparisons with the *Chlamydomonas* outer arm dynein LC7a light chain [135,136]. *Drosophila* has at least six Roadblock homologs, including bithoraxoid, which has been implicated in thoracic and abdominal parasegment development. These proteins belong to an ancient family that has been implicated in NTPase regulation in bacteria [137]. Mutations in the Roadblock genes result in the accumulation of axonal cargoes, mitotic defects, female sterility, and either larval or pupal lethality [135]. Roadblock mutations also affect neuroblast proliferation and result in reduced dendritic complexity, as well as in defects in axonal transport [138]. Mutational analysis in *Chlamydomonas* suggests that DYNLKB (LC7a) is involved in axonemal outer arm dynein assembly, and a related protein (LC7b) is associated with dynein regulatory elements [136,139].

Figure 2E shows the phylogenetic relationships amongst the dynein light chain Roadblock-family protein sequences from various organisms. The Roadblock sequences are remarkably well conserved between different organisms, with 96% of pair-wise sequence comparisons amongst all sequences shown in Figure 2E demonstrating an identity greater than 50% (data not shown). The high conservation of Roadblock-family sequences presumably arises from functional constraints on the proteins. We note that genes in mammals and in other species incorporate conserved and complete Roadblock sequences (known as Roadblock domains) within their coding regions [135,137]. However, these genes are not thought to be cytoplasmic dyneins; for example, MAPBPIP in human and mouse appears to function mainly in the endosomes/lysosome pathway [140]. Both DYNLKB polypeptides are found in mammalian cytoplasmic dynein, but it is not yet known if just one, or both, are utilized in mammalian axonemal dyneins.

**Cytoplasmic dynein light chain Roadblock1, DYNLRB1.**

Database searches [135,141] (Figure 2E) revealed there are two Roadblock-related proteins in mammals, DYNLRB1 and DYNLRB2 (also termed DYNLC2A and DYNLC2B) [142].

Biochemical studies suggest that in mammals both Roadblocks exist as homo- and heterodimers that associate with cytoplasmic dynein [143] through specific binding sites on the intermediate chains, distinct from those for the DYNNL (LC8) and DYNTL (Tctex1) light chains [40]. Expression studies in humans have identified tissue-specific differences in the expression of the two human Roadblock-like genes, with strong expression of DYNLRB1 in heart, liver, and brain, and up-regulation in hepatocellular carcinoma tissues [142]. In a role that may be independent of its association with cytoplasmic dynein, TGFb phosphorylation of human DYNLRB1 (termed mLCT-14kDa by Tang and colleagues [144]) results in the human DYNLRB1 binding to the TGFb receptor that mediates TGFb responses including JNK activation, c-JUN phosphorylation, and growth inhibition.

**Cytoplasmic Dynein Light Chain LC8 Gene Family (DYNNL1, DYNNL2)**

Cytoplasmic dynein light chain LC8 1, DYNNL1. DYNNL (the light chain that has been known as LC8, as well as LC8a and PIN) is a component of many enzyme systems, and it has a long and somewhat confusing history. This protein was originally identified, using biochemical methods, as a light chain of the *Chlamydomonas* axonemal outer arm dynein [145,146]. The term LC8 derives from the observation that this component migrates at ~8 kDa in SDS-PAGE gels, and it is also the smallest of the eight light chains then known within this *Chlamydomonas* axonemal dynein. It was first cloned from *Chlamydomonas*, and closely related sequences were identified in mouse and nematode along with more distantly related proteins in higher plants [147,148]. Using biochemical and immunochemical methods, DYNNL was also identified as an integral component of brain cytoplasmic dynein [104]. Only recently, it has been realized that mammals have two closely related DYNNL genes, and that the protein products of both genes are components of cytoplasmic dynein [148,149]. Thus, most of the studies on the cellular roles of DYNNL do not distinguish between the two DYNNL polypeptides.

Another factor complicating efforts to elucidate the role of the DYNNL polypeptides in dynein function was the realization that large amounts of DYNNL1 in brain, and presumably cells in general, are not associated with the dynein complex [104]. In fact, the DYNNL polypeptides have other important functions unrelated to their role in axonemal and cytoplasmic dyneins. DYNNL1 is a subunit of the flagellar radial spokes which are involved in control of axonemal dynein motor function [150]. DYNNL1 is also a substrate of a p21-activating kinase, and its interaction with the kinase may be important for cell survival [151]. A DYNNL is an integral component of the actin-based motor myosin V [152]. Immunostaining shows that a DYNNL is concentrated in dendritic spines and growth cones, and it is proposed that this is due to its association with the actin-based motor.
myosin V [149]. DYNLL1 was identified within neuronal nitric oxide synthase (nNOS) [14] and named “PIN” for “protein inhibitor of nNOS” [153]. However, it is unclear whether it is actually an inhibitor of nNOS or is merely a component of the nNOS complex, as DYNLL1 appears to be required for the stability of various multimeric complexes. DYNLL1 has been found to interact with a wide variety of other cytoplasmic components, including the pro-apoptotic factor Bim [154], Drosophila swallow [155,156], and rabies virus P protein [157], and it may act to attach them to the dynein and/or myosin-V molecular motors. In addition, there are many other DYNLL-interacting proteins not mentioned here that have been identified using yeast two-hybrid screens and other methods.

There are two copies of DYNLL in the cytoplasmic dynein complex, and the crystal and NMR structures of the DYNLL dimer with bound peptide are known [104,158-160]. Both monomers contribute to the formation of two symmetrical grooves in the dimer that are the binding sites for two DYNCII polyepitides reviewed in [161]. Adding DYNLL to an N-terminal polyepitide of DYNCII in vitro increases the structural order of DYNCII, suggesting that DYNLL is important for the assembly of a functional dynein complex [84].

Figure 2F shows the phylogenetic relationships amongst the dynein light chain LC8-family protein sequences from various organisms. Our phylogeny shows that the mammalian LC8 light chain family falls into two distinct clades containing DYNLL1- and DYNLL2-like genes. DYNLL is highly conserved from alga and humans, and homologs are required for sensory axon projection and other developmental events in Drosophila [162,163], nuclear migration in Aspergillus [164], and retrograde IFT in Chlamydomonas [72]. The phenotype of partial loss-of-function mutants in Drosophila revealed a wide array of pleiomorphic developmental defects; the total loss-of-function mutation was embryonic lethal [162]. The Drosophila dynein light chain 1 (Cdlc1, also known as dclc1 and “cut-up”) is ubiquitously expressed during development and in adult tissue, and is required for proper embryogenesis and cellular differentiation. Mutations in this gene result in female sterility, which may be due to the severely disordered cytoskeletons of ovarian and embryonic cells [162]. A high degree of sequence similarity (92%) exists between Drosophila Cdcl1 and the 8-kDa flagellar outer arm dynein light chain from Chlamydomonas, and with human and C. elegans light chain 1 (91%), suggesting this gene has been under strong selective pressure [162]. S. pombe has a single known DYNLL homolog, SPAC926.07c (also referred to as Dlc2); it is transcribed during the vegetative phase, induced at low level in the sexual phase, and is enriched in the nuclear periphery [123]. A Dlc2 null mutant has been described with marginally reduced recombination in meiosis, but no other reported phenotype [123].

During the course of homology searches for this paper, we noted that DYNLL1 has related sequences in several locations in the human genome (data not shown); none of these appear to be associated with expressed sequences and thus may be pseudogenes. There was also a discrepancy in the likely mapping position of DYNLL1, and therefore we carried out a sequence analysis of DYNLL1-related genomic loci and show that the cognate human locus lies on Hsa12q24.31 (data not shown), which agrees with the mouse mapping result of Dynll1 on Mmu5.

Cytoplasmic dynein light chain LC8 2, DYNLL2. DYNLL2, also known as DYNLL2 and LC8b, is the second member of this light chain family. It was identified by micro-sequencing of polyepitides from purified brain cytoplasmic dynein [149] and a yeast two-hybrid screen [149]. Mammalian DYNLL1 and DYNLL2 have 93% identity, differing by only six amino acids out of 89. Indicative of the extraordinary conservation of these proteins, the amino acid sequences of both DYNLL1 and DYNLL2 from human, mouse, rat, pig, and cow are identical [148]. Human DYNLL2 was identified in a yeast two-hybrid screen using the guanylate kinase-associated protein (GKAP) as bait and may mediate the interaction between GKAP and actin- and microtubule-based motors, allowing GKAP and its associated proteins to be translocated as a cargo, although DYNLL1 also binds to GKAP [149]. DYNLL2 binds the pro-apoptotic factor Bmf, which binds Bcl2, neutralizing its antiapoptotic activity, a role comparable to that reported for the binding of Bim to DYNLL1 [165]. However, it has also been observed that Bim and Bmf have identical binding affinities for both DYNLL1 and DYNLL2 [166].

It has further been proposed that DYNLL1 binds specifically to the dynein intermediate chain DYNCII, while DYNLL2 binds to the myosin-V heavy chain. However, DYNLL2 co-purifies with cytoplasmic dynein from various rat tissues [148], and DYNLL1- and DYNLL2-GST are equally effective in binding myosin V [149]. Furthermore, DYNLL1 and DYNLL2 bind with equal affinity to DYNCII in pair-wise yeast two-hybrid studies (K. W. Lo and K. K. Pfister, unpublished data). It is not yet known if one, or both, of the DYNLL polypeptides are associated with axonemal dyneins; however, DYNLL1 is enriched in testes and lung—tissues that have large numbers of cilia or flagella [148].

Human and Mouse Cytoplasmic Dyneins: Nomenclature, Map Positions, and Sequences

To create Table 1, we cataloged, by literature searches, all known gene and protein names for the cytoplasmic dyneins in mouse and human. In addition, aliases were recorded from the single-query interface LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink) and the Mouse Genome Informatics (MGI) website (http://www.informatics.jax.org). We also included aliases previously approved by the HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/ nomenclature) as well as aliases referenced in sequence submissions to the GenBank (http://www.ncbi.nlm.nih.gov/ Genbank) and Entrez (http://www.ncbi.nlm.nih.gov/entrez) sequence databases [167].

Human and mouse orthologs in Table 1 are taken from the literature and databases. Human and mouse chromosomal locations were obtained from the literature and from the MGI and LocusLink databases. The OMIM numbers given for gene and disease loci in humans refers to the unique accession numbers in the On-line Mendelian Inheritance in Man database (http://www.ncbi.nlm.nih.gov/entrez/query. fcgi?db=OMIM). Nucleotide and protein sequences (prefix NM___ and NP___ respectively) are National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq) and Swiss-Prot accession numbers (http://www.ncbi.nlm.nih.gov/RefSeq) and SwissProt accession numbers (http://www.ebi.ac.uk/swissprot), respectively [167]. The NCBI RefSeq project provides a non-redundant and comprehensive collection of nucleotide and
protein sequences drawn from the primary-sequence database GenBank. RefSeq collates and summarizes primary-sequence data to give a minimal tiling path for individual transcripts, using available cDNA and genomic sequence whilst removing mutations, sequencing errors, and cloning artifacts. Sequences are validated in silico by NCBI's Genome Annotation project to confirm that any genomic sequence incorporated into a RefSeq cDNA matches primary cDNA sequences in GenBank, and that the coding region really can be translated into the corresponding protein sequence. Accession numbers beginning with the prefix XM_ (mRNA) and XP_ (protein) are RefSeq sequences of transcripts and proteins that are annotated on NCBI genomic contigs; these may have incomplete cDNA-tiling-sequence data or contig sequences [168]. For dyneins with known isoforms, isoform-sequence accession numbers available within nucleotide and protein databases are given. Included in the heavy chains that we found were "Cell Division Cycle 23, yeast homolog" (CDC23) and "Cell Division Cycle 22, yeast homolog" (CDC22), which are GenBank aliases for human and mouse dynein heavy chain 1, respectively. We found no evidence in the literature to support the "CDC" designation of these genes and their products in terms of either "Cell Division Cycle" or "Cytoplasmic Dynein Chain". We compared mouse and human heavy chain 1 cDNA and protein sequences with mouse, human, and yeast CDC23 and CDC22 sequences and found no similarity to support this designation (data not shown). We concluded that the synonym CDC had most likely been attributed in error, and we contacted NCBI who, in agreement with our findings, removed the CDC designation from the sequences involved.

**Human/Mouse Homology Searches**

Homology searches of human cytoplasmic dynein subunit genes were conducted using position-specific iterative BLAST (PSI-BLAST) [169] at NCBI (http://www.ncbi.nlm.nih.gov/BLAST; Table 2). The PSI-BLAST program identifies families of related proteins using an iterative BLAST procedure [170]. In an initial search, a position-specific scoring matrix is constructed from a multiple sequence alignment of the highest scoring hits. Subsequent iterations using the position-specific scoring matrix are performed in a new BLAST query to refine the profile and find additional related sequences. We used nucleotide and protein sequences from each known human dynein gene to query the human and mouse non-redundant sequence databases at GenBank, using default parameters and the BLOSUM-62 substitution matrix, which has been shown to be the most effective substitution matrix to identify new members of a protein family [171]. Where dynein isoforms were present, the longest sequence was used to search the databases.

**Phylogenetic Analysis**

To establish gene family groupings, we investigated the phylogenetic relationships between dynein protein homologs in various organisms. Homologous sequences were identified by searching the GenBank non-redundant protein database, with the human protein using PSI-BLAST with default parameters and the BLOSUM-62 substitution matrix. Searches of pufferfish sequence *Takifugu rubripes* (commonly known as *Fugu rubripes*), for which little transcribed sequence exists although a usable genome assembly is present, were performed using the BLAST (TBLASTN) feature at the Ensembl Fugu Genome Browser (version 2.0; http://www.ensembl.org/Fugu__rubripes), searching with human protein sequence against a translated nucleotide database.

Protein sequences were aligned for comparison across their full lengths using the multiple sequence alignment program CLUSTALW [172] (http://www.ebi.ac.uk/clustalw) and applying the GONNET250 matrix as default. The GONNET250 is a widely used matrix for performing protein-sequence alignments, allowing 250 accepted point mutations per 100 amino acids, using scoring tables based on the PAM250 matrix [173]. Two different phylogenetic methods were used to analyse the dynein gene family alignments. Maximum-likelihood trees were inferred under the Jones, Taylor, and Thornton (JTT) empirical model of amino-acid substitution using PHYML version 2.4.3 [174], as was non-parametric bootstrapping using 100 resampled alignments for each gene family. Bayesian analyses were performed using MrBayes version 3.0B4 [175], using the default Bayesian priors on tree topologies and branch lengths. Two different sets of analyses were performed for each gene family, the first allowing the Markov-chain Monte-Carlo algorithm to move between the 11 different amino-acid substitution models available in MrBayes, and another specifying the JTT model. The first analysis allows the chain to take into account uncertainty in the substitution process. For all analyses performed here, the posterior probability of the JTT model was at least 99%, confirming that this model best describes the evolution of the dynein sequences—so only results from the fixed-JTT model analyses are shown here.

For each analysis, three chains of 1,000,000 generations each were run, sampling parameters every 100 generations and discarding the first 100,000 generations as a burn-in period. Running these multiple independent chains allowed visual confirmation that the chains had reached a stationary state by ensuring that all three chains were moving around a region of similar likelihood. For one of the gene families (cytoplasmic dynein heavy chain), the three chains had reached different likelihood values after 1,000,000 generations, suggesting failure to converge. Running another three independent chains resulted in five out of six chains agreeing on the likelihood values, suggesting that only one chain had not converged properly. In all cases, the phylogeny presented is the majority-rule consensus of the posterior sample of tree topologies from all three Markov chains, drawn using TreeView [176] with posterior clade probabilities and maximum-likelihood bootstrap values shown for each clade on these trees.

**Searching for Function and Mutant Phenotypes**

As well as literature searches, information on protein function was taken from the Gene Ontology database (http://www.geneontology.org), which provides data on function and processes associated with a search protein. Mutant-phenotype data were obtained from the literature and the following sources: Online Mendelian Inheritance in Man at NCBI for human (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM); MGI for mouse (http://www.informatics.jax.org); FlyBase for *Drosophila* (http://www.flybase.org); and WormBase for *C. elegans* (http://www.wormbase.org).
Conclusions

In this paper, we have provided an overview of the two cytoplasmic dynein complexes, cytoplasmic dynein 1 and cytoplasmic dynein 2, from a genetic perspective. We have highlighted the unique subunit compositions and cellular functions of the two cytoplasmic dyneins, and we have emphasized the unique role of cytoplasmic dynein 2 in IFT.

We have described the different mammalian dynein gene families, and have shown the phylogenetic and functional relationships between members of individual families. We carried out initial database searches and clarified and corrected anomalous data. We have also discussed known functions and mutations of these proteins, and we have highlighted both their fundamental importance to the cell and the fact that much research remains to be carried out to define the roles of individual proteins.

Supporting Information

Table SI. Species Names, NCBI/GenBank Protein-Sequence Accession Numbers, and NCBI/GenBank Gene/Protein Names for Figures 2A–F

Accession Numbers

The Entrez Gene database (http://www.ncbi.nlm.nih.gov/entrez) accession numbers for the proteins discussed in this paper are Cdc19 (also referred to as Cdc19 and Dic1) (44160); che-3 (DYNC2H1 homolog) (172593); DYNClH1 (172598); DYNClH2 (172597); DYNClH3 (51433); DYNClH2, (1783); DYNClH1 (197659); DYNClH2 (51628); DYNLL1 (LC8) (8655); DYNLL2 (also known as DYNLL2 and LC8b) (140735); DYNLL1 (Tctex1) (6991); MAP8 (28956); MAP8P (28955); MAP180 (28956); MAP180s (28956); SPAC 1805.08 (also referred to as Dic1) (3561491).

The Entrez Gene database (http://www.ncbi.nlm.nih.gov/entrez/) accession numbers for the genes discussed in this paper are Dhc64C (also known as LC2, LC8 sequence) (6991); mouse Dhc64C (13426); Chlamydomonas dli-1 (498422); Cdlcl (525075); DynclH1 (Dyncl1) (13426); DynclH2 (Dyncl2) (13427); DynclH1 (Tctex1) (6991); mouse Tctex2 (19647); xbo-1 (184080).

The Entrez Protein database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=protein&searchterm=accession numbers) accession numbers for the proteins discussed in this paper are C. elegans LC8 sequence (49822), C. elegans light chain 1 (498422), Dic1 (520576); Chlamydomonas 19,000-Mr axonemal outer arm dynein light chain (LC2) (AABS5933); rat DYNClH1 (102107); rat DYNClH2 (142289).

The OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) accession numbers for the proteins discussed in this paper are autosomal recessive disorder primary ciliary dyskinesia (264260); Bino (605266); X-linked retinitis pigmentosa type 3 (300359).

The Ensemble (http://www.enscmbl.org/Fugu_UBR1/queries/textview) accession number for the Takifugu LC8 sequence is SIFRUP000015498.

The SwissProt (http://ca.expasy.org/sprot/) accession numbers for the Chlamydomonas 8-kDa flagellar outer arm dynein light chain and the Chlamydomonas LC2 light chain used as the outgroup in Figure 2 are Q39580 and T06816, respectively.

Acknowledgments

PRS, HH, and EMCF are supported by the UK Medical Research Council, the Motor Neurone Disease Association, and the American Amyotrophic Lateral Sclerosis Association. KKP is supported by a grant from the National Institute of Neurological Disorders and Stroke, at the National Institutes of Health (NIH). SMK is supported by grants (GM51293 and GM63548) from the National Institutes of General Medical Sciences, NIH, and is an investigator of the Patrick and Catherine Welldon Donaghy Medical Research Foundation. JC is supported by the Biotechnology and Biological Sciences Research Council (BBSRC), grant 40G18588. AR is supported by grants from the BBSRC and the Royal Society. We are most grateful to Lois Maltais of the Mouse Genomic Nomenclature Committee and to Mathew Wright of the Human Genome Organization Gene Nomenclature Committee for their help and support in preparing for this manuscript. We thank Ray Young for supplying graphics.

References


43. Schrauwen J, Joshi PJ,Inserted4, Dwyer ND, Bargmann CJ, et al. (1999) Role of a class D HCII dynein in retrograde transport of IFT motors and IFT raft particles along cilia, but not dendrites, in chemo sensory neurons of living G a r m u d e s o m e e l e g a n s . J Cell Biol 147: 515-520.


No association of DYNCH1 with sporadic ALS in a case-control study of a northern European derived population: A tagging SNP approach

PARESH R. SHAH1,2, AZLINA AHMAD-ANNUAR1, KOUROSH R. AHMADI3, CARSTEN RUSS4, PETER C. SAPP5, H. ROBERT HORVITZ5, ROBERT H. BROWN JR4, DAVID B. GOLDSSTEIN3 & ELIZABETH M. C. FISHER1
Lack of association of DYNC1H1 and sporadic ALS
Lack of association of DYNC1H1 and sporadic ALS
Lack of association of DYNC1H1 and sporadic ALS
Lack of association of DYNC1H1 and sporadic ALS
Lack of association of DYNC1H1 and sporadic ALS