Practical Neurology

Lessons and pitfalls of whole genome sequencing

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Lessons and pitfalls of whole genome sequencing

Christopher J Record, Mary M Reilly

Centre for Neuromuscular Diseases, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK

Lead corresponding author: Professor Mary M Reilly, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK. m.reilly@ucl.ac.uk

Secondary corresponding author: Christopher J Record, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK. Chris.record@ucl.ac.uk

4545/4400

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

ABSTRACT

Whole genome sequencing (WGS) has recently become the first line genetic investigation for many suspected genetic neurological disorders. Whilst its diagnostic capabilities are innumerable, as with any test, it has its limitations. Clinicians should be aware of where WGS is extremely reliable (detecting single nucleotide variants), where its reliability is much improved (detecting copy number variants and small repeat expansions) and where it may miss/misinterpret a variant (large repeat expansions, balanced structural variants or low heteroplasmy mitochondrial DNA variants). Bioinformatic technology and virtual gene panels are constantly evolving, and it is important to know what genes and type of variants are being tested; the current NHS Genomic Medicine Service WGS offers more than early iterations of the 100,000 Genomes Project analysis. Close communication between clinician and laboratory, ideally through a multidisciplinary team meeting, is encouraged where there is diagnostic uncertainty.

INTRODUCTION

Whole genome sequencing (WGS) and the use of genomic testing in neurology, including consent, indications and results, have recently been expertly reviewed in Practical Neurology.(1,2) The success of the Genomics England 100,000 Genomes Project (100KGP), sequencing patients with cancer and rare-diseases, has led to the introduction of WGS with virtual panels into routine clinical practice for many neurological diseases via the National Health Service Genomic Medicine Service (NHS-GMS, https://www.england.nhs.uk/genomics/nhs-genomic-med-service/). The theoretical benefits of WGS are clear; sequencing the entire genome (many orders of magnitude more DNA than previous routine testing, at comparable costs) wherein, provided the clinical diagnosis of a genetic disorder is correct, the molecular diagnosis should lie. However, as with every new technology, WGS has its limitations. This article aims to outline the diagnostic utility of WGS, but also where caution needs to be exerted. A critical step in any patient's diagnostic journey is the decision to request WGS. Where appropriate, especially in sporadic cases, acquired diseases should be excluded first. Key points to consider before requesting WGS are highlighted in Figure 1

IS CHARCOT-MARIE-TOOTH DISEASE A GOOD DISEASE PROTOTYPE FOR UNDERSTANDING WGS?

Charcot-Marie-Tooth (CMT) disease is an umbrella term for inherited neuropathies but is a clinically and genetically heterogenous group of diseases. The clinical sub-types of CMT include demyelinating sensory and motor neuropathy (CMT1), axonal sensory motor neuropathy (CMT2), sensory and motor neuropathy with intermediate conduction slowing (upper limb motor conduction velocity between 25 and 45 m/s, CMTi), hereditary sensory neuropathy (HSN), and hereditary motor neuropathy (HMN).(3,4)

The diagnostic utility of WGS for an individual lies in its ability to detect vast numbers, and theoretically different types, of genetic variant. Figure 2 illustrates the features of a disease group that make it suitable for considering WGS testing.

One down-side of CMT as a disease prototype, is that functional validation of novel variants/genes is challenging but this underpins how important WGS is in CMT clinical practice. Gold standard functional evidence would be ex-vivo human diseased tissue demonstrating absent, deficient or dysfunctional protein contributing to pathology. This is theoretically possible with peripheral nerve biopsies, but this is an invasive procedure requiring technical expertise. Alternatively, RNA sequencing can be used to demonstrate aberrant transcripts in appropriate tissues; Schwan cells are clearly easier to study than

dorsal root ganglia or anterior horn cells. Overall, we feel CMT is an excellent disease to demonstrate the lessons and pitfalls of WGS and will explore these herein.

WHOLE GENOME SEQUENCING TECHNOLOGIES

A basic understanding of the molecular techniques involved in WGS is important to appreciate its potential pitfalls. First, WGS when used in common medical parlance, refers to 'short-read' WGS (srWGS). Some useful terminology is highlighted in Table 1. Other forms of genomic sequencing exist, and although currently used mostly in the research setting, their use is increasing in diagnostic genetic laboratories worldwide. Long-read WGS (lrWGS), as suggested in the description, continuously sequences long molecules of DNA, typically tens of kbp in length, but up to many hundreds of kbp depending on the sequencing technology used. The major benefit of lrWGS is the ability to accurately detect and size repeat expansions, and detect complex, balanced structural variants (SVs). The drawbacks include the cost and longer sequencing time, and its error rate on an individual nucleotide level which, when combined with low read depth, affects its ability to reliably detect single nucleotide variants (SNVs) or insertion-deletions (indels).(5) Optical genome mapping (OGM) is another form of genomic interrogation, and more appropriately termed 'genome imaging'. Its uses have been compared to those previously investigated with karyotyping (i.e. large SVs) but with the benefit of up to 20,000-fold higher resolution. DNA molecules are enzymatically labelled, and the resultant ligated DNA then 'imaged' for its pattern of periodically spaced fluorescent signals. Its ability to detected large SVs (0.5 – 1 Mbp) is superior to srWGS and lrWGS, and it is less costly to get higher coverage. As with srWGS and lrWGS it cannot detect aneuploidy (an abnormal number of chromosomes), although this is less relevant in the setting of non-developmental disorders. Another potential drawback of OGM is the requirement for DNA extraction from a fresh blood sample.(6) As neither lrWGS nor OGM are used in standard NHS testing, from this point forward we will not discuss them further, and we will refer to srWGS simply as WGS.

TABLE 1 USEFUL TERMINOLOGY

Alignment	Process by which the sequenced reads of the individual are matched to the appropriate region in the reference genome
Bioinformatic pipeline	Series of computational steps processing raw WGS data resulting in variants for analysis; includes alignment, filtering, variant calling
Copy number variant	A type of structural variant, where there is an increase or decrease in dosage across a region of the genome e.g. deletion or duplication
Insertion-deletion (indel)	A small variant, usually <50bp, where a series of consecutive nucleotides are altered/inserted/deleted

Long-read	A single strand of DNA is sequenced continuously, depending on the technology, up to hundreds of kbp in length
Paired-end reads (read-pair)	A fragment of DNA is sequenced from both ends to create paired-end reads, or a read pair. Information can be gleaned when the pairs are aligned to the reference if they are unexpectedly too close, too far, on separate chromosomes or if the one half of the pair is 'unmatched'
Read	The basic molecular output of next generation sequencing; a read is a single consecutively sequenced strand of DNA, before alignment to the reference genome
Read depth	The number of cumulative reads aligned at a particular genomic locus i.e. how many times a particular nucleotide been sequenced
Read length	The number of nucleotides in a single read
Repeat expansion	Repeating nucleotide motif (e.g. CAG_n – the common polyglutamate expansion, or $AAGGG_n$, the most common configuration in *RFC1* CANVAS) where n is the number of repeats. The number of repeats that is considered pathogenic varies widely between diseases.
Short-read	A single strand of DNA is consecutively sequenced between 75-300 bp in length
Single nucleotide variant (SNV)	The alteration/insertion/deletion of a single nucleotide
Structural variant (SV)	Medium to large (typically 100s to Mbp in length) variants including duplications, deletions, insertions, balanced translocations and more complex rearrangements
Variant calling	The process by which alterations (variants) in the individual's genome are identified by comparing to the reference genome

NEXT-GENERATION SEQUENCING TECHNOLOGY

WGS uses next-generation sequencing (NGS) technology, also known as high-throughput or massively-parallel sequencing. NGS has been used for many years in clinical diagnostic laboratories for the sequencing of disease-specific gene panels and whole exome sequencing. There are a number of sequencing platforms,[7] but the dominant provider worldwide is Illumina, which is also used by NHS-GMS, and the process described hereafter. A flow diagram of the process involved is shown in Figure 3.[8]

The first step is **library preparation** (Figure 3A); the genomic DNA library is a series of short fragments ready for sequencing. The DNA (typically extracted from leukocytes in blood; purple EDTA tube) is fragmented and then each fragment amplified. Fragments are then **sequenced** in a process called 'sequencing by synthesis', whereby fluorescently tagged nucleotides are added to a linear single strand of DNA complementary to the fragment; the resultant fluorescent DNA strand is known as a 'read' and can be sequenced by its characteristic spectral emission (one wavelength for each of the four nucleotides, Figure 3B). The fragment is sequenced from both ends forming 'paired-end reads',

allowing additional information to be gleaned when the reads are aligned. Data is then fed into the **bioinformatic pipeline** (Figure 3C). The millions of reads are **aligned** to the reference genome, which when visually represented, form piles of overlapping reads. The overall coverage of the WGS describes what proportion of the reference genome is sequenced to a satisfactory read depth. Figure 4A shows in detail how an unmutated fragment is sequenced and aligned to the reference.

Variant calling is the process of identifying variants i.e. variation in an individual's genome when compared with the reference. The basic output of a WGS bioinformatic pipeline is the identification of small variants; alteration/insertion/deletion of single nucleotides (SNVs, Figure 4B) or a small number of consecutive nucleotides (indels). The universal final output for the millions of variants generated is a .vcf file. Other types of genetic variant can also be detected including SVs (both copy number variants, and balanced rearrangements; the latter where there is no change in dosage at a particular locus), repeat expansions and mitochondrial DNA (mtDNA) variants, but their detection and calling is variable (Figure 4 and see *When WGS might not be the correct test*).

After variant calling, the variants are **filtered** according to specified criteria (see *Filtering and prioritisation*). Application of a **virtual panel**(s) may yield possible candidate variants, which are **interpreted** by clinical scientists (Figure 3D). If there is ambiguity or uncertainty, results are ideally discussed at a multidisciplinary team (MDT) meeting, following which a genetic report can be issued.

VIRTUAL PANELS

Although WGS theoretically allows analysis of variants from an individual's entire genome, this is neither desirable (incidental unwanted findings) nor practical (a human genome contains approximately five million SNVs) therefore virtual panels are essential to refine the search. In the NHS, clinicians are required to select virtual gene panel(s) when requesting WGS. The NHS-GMS PanelApp (https://nhsgms-panelapp.genomicsengland.co.uk/panels) is a publicly available resource that utilises genetic expertise through crowdsourcing to curate disease-specific gene panels.(9) For a gene to be included it needs to be approved as 'green' by a number of verified experts; a green gene is broadly one in which plausible disease-causing variants have been found in three or more unrelated individuals/families. However, the panels can only be as correct and up-to-date as their reviewers and the current available evidence. For example, *SORD* was discovered as a common, and potentially treatable, cause of CMT in 2020,(10) but was not approved as a green gene until November 2022. Panels are periodically updated, and previous iterations can be found on PanelApp. Genes that cause a complex phenotype which include the disease group of interest e.g. *ABHD12* causing polyneuropathy, hearing loss, ataxia, retinitis pigmentosa, and cataract (PHARC) syndrome, are often not included if the panel specifies an isolated phenotype; it is not a green gene on the current

'Hereditary Neuropathy or pain disorder' panel (R78 version 3.24). Similarly, novel, rare genes may not meet green inclusion criteria. It is therefore important to have an understanding of which genes are tested in a specific panel, and if there is a particular gene of interest in a clinical case, this should be discussed with the genetic laboratory. It is currently recommended that broad rather than narrow use of panels is applied to maximise chances of identifying causative variants.

FILTERING AND PRIORITISATION

Refining the vast number of variants detected through WGS requires filtering strategies. The two most powerful tools are the allele frequency of the variant in reference databases (the most commonly used is gnomAD; https://gnomad.broadinstitute.org/, Box 1) and in family studies, the inheritance pattern, as defined by relative disease status.

Population allele frequency

Historically the upper limit for the population allele frequencies was set at < 1 in 100 for autosomal recessive, and < 1 in 1000 for autosomal dominant (AD) disease, however we know that for most rare diseases these thresholds are far too high. A useful online calculator for the estimation of a disease-specific population allele count and frequency is found at https://cardiodb.org/allelefrequencyapp/. It is important to remember that if a variant seemingly occurs at too high an allele frequency, it will be filtered by the bioinformatic pipeline, and not considered for interpretation. The most common variant c.757delG in *SORD*-related CMT is present in a highly homologous non-functioning pseudogene *SORD2P* in 95% of controls; the two variants can be challenging to delineate bioinformatically and therefore the *SORD* variant is potentially inappropriately filtered.(10) This problem with this particular variant has been overcome but was a barrier to its discovery.

One must also be wary of regional 'hotspots' for particular variants. The *GNE* variant p.Val696Met (previously p.Val727Met) causing the rare recessive hereditary inclusion body myopathy/Nonaka myopathy is exceedingly common in the South Asian population where the majority of the disease is seen.(11) The overall quoted allele frequency appears too high for the prevalence of the disease in the UK, and may result in the variant being discounted. Only when the regional breakdown is examined, can it be appreciated that the variant is very rare in European populations, in keeping with disease prevalence.

Reference genome

The current human reference genome, denoted GRCh38, originates from the genomes of 20 anonymous volunteers from the USA. It has been shown that two-thirds is comprised of the genome

of a single individual of mixed European and African descent.(12) It is widely recognised that the current reference genome has significant limitations; it contains some gaps (~5%), has regions of unreliable coverage (e.g. around the centromere), and reflects a very narrow ancestry. The Human Pangenome Reference consortium have set out to rectify the flaws in the current reference by creating a new reference built from 350 human genomes, and have recently published a draft from 47 individuals from diverse backgrounds.(13) Until the 'Pangenome' comes into routine clinical practice, clinicians must we aware that patients from certain ethnic backgrounds (e.g. the Indian subcontinent) may have variants missed because the reference does not reflect their ancestry.

Family studies and relative disease status

Variant segregation through family studies (WGS in more than one family member that are subsequently analysed together) enhances diagnostic success.(14) At recruitment, participants are assigned as affected, unaffected or unknown. Downstream in the process, if a dominant variant is detected in the affected proband and a reportedly unaffected parent, it will be disregarded or deprioritised. Therefore, caution should be exerted, when the disease has an adult-onset or a variable presentation, that relatives' disease status is appropriately assigned.

Human phenotype ontology terms

As part of the process of requesting WGS through the NHS-GMS, the clinician is required to include Human Phenotype Ontology (HPO) terms (https://hpo.jax.org/app/, Box 2). These phenotypic descriptors can be used to prioritise variants using Exomiser, a programme utilised by NHS-GMS.(15) For example, a male patient with a demyelinating neuropathy and upper motor neurone signs underwent WGS in the 100KGP with the Hereditary Neuropathy virtual panel applied. There were no candidate variants from the panel, but because the HPO terms included 'demyelinating neuropathy' and 'Babinski sign', a variant in *ABCD1*, known to cause X-linked myeloneuropathy, was identified. Subsequent discussion at our MDT, and further clinical and laboratory assessments, confirmed this to be the causative gene. This gene is not present in the current Hereditary Neuropathy panel.

VARIANT INTERPRETATION AND REPORTING

Every candidate variant is classified according to established criteria. UK laboratories use the American College of Medical Genetics and Genomics (ACMG) and Association for Clinical Genomic Service (ACGS) guidelines.(16,17) Any given variant, with no supporting data, starts as a 'variant of uncertain significance' (VUS). Evidence is combined, from different categories (including data on allele frequency, functional studies, segregation and prior literature reports) to upgrade the variant as likely

pathogenic or pathogenic, or downgrade to likely benign or benign (Figure 5). As with gene panels, variant interpretation is reliant upon the available evidence, and its application, and therefore variant classification may differ between laboratories. Ideally, clinicians will have access to an MDT (with clinical scientists) to discuss WGS results of unsolved cases, cases with unexpected pathogenic variants, or those with a very typical phenotype for a particular gene, in which no variants have been reported. There is a criterion within the ACMG/ACGS (PP4) criteria that uses phenotype specificity to upgrade variants e.g. absence of dystrophin in a muscle biopsy in a male patient with muscular dystrophy phenotype, when considering a variant in *DMD*. Without the communication of clinical information from clinician to laboratory, the variant might remain a VUS.

Historically, relevant VUSs were listed as an addendum to genetic reports. However, NHS-GMS have adopted guidance from the ACGS that VUSs should not be reported unless exceptional circumstances apply, after a discussion at an MDT meeting. This change is critical for practising clinicians to be aware of. The rationale is that reporting a VUS may lead to confusion on the part of referring clinician or patient, misinterpretation and potentially misdiagnosis. Even when a VUS is likely to be causative, family screening for the variant would still need careful discussion and counselling, and preimplantation genetic testing or entry into a clinical drug trial would only be considered in exceptional circumstances.

However, we have experience that transparent reporting of VUSs to clinicians with genetic expertise, has been vital in clinching a genetic diagnosis with the passage of time. A 'warm' VUS may be upgraded to pathogenic following, for example, a new publication implicating the gene/variant in disease. Without information about VUSs made available on a genetic report, such cases may remain unsolved.

Another example of the need for careful reporting is the presence of a single pathogenic variant in a recessive gene, a so called 'single hit'. If reported, it should be made clear that the diagnosis is not confirmed, but a single pathogenic variant has been detected. With a suggestive phenotype, a 'single hit' will often trigger a discussion with the laboratory to look on the other allele for deep intronic variants (that might affect splicing or create pseudoexons), or SVs (i.e. deletion of a portion of the gene), or explorative analysis of the genome in a research setting.(18)

WHEN WGS MIGHT NOT BE THE CORRECT TEST

The essential first step for genetic testing is ensuring the right test is sent. Jain *et al*. have previously discussed this in detail.(1) In the UK clinicians must consult the NHS Genomic Test Directory https://www.england.nhs.uk/publication/national-genomic-test-directories/. Many neurological diseases, including some that are treatable, have their molecular basis in non-SNV genetic variation.

Huntington's disease (CAG trinucleotide repeat expansion in *HTT*), genetic motor neurone disease/frontotemporal dementia (GGGGCC hexanucleotide repeat expansion in *C9ORF72*), spinal muscular atrophy (biallelic deletion of exon 7 +/- 8 in *SMN1*), fragile X syndrome (CCG trinucleotide repeat expansion in *FMR1*) and Duchenne muscular dystrophy, (~60% caused by exon-level deletions in the X-linked *DMD*) are all caused by either repeat expansion or SVs. More than 50% of CMT is caused by a duplication of *PMP22*, and the remainder, a mixture of genetic variant types.

The limitation of WGS to accurately detect SVs and repeat expansions lies in the read length. Put simply, it is difficult to quantify a variant with genomic size potentially orders of magnitude larger than the unit of measurement. Figure 4 details the use of paired-end reads in the sequencing and alignment process, and how they can be used to detect non-SNV variants. When the DNA fragment is sequenced from both ends, the two paired-end reads contain markers that identify them as a pair. If, when the reads are aligned to the reference genome, they align too far apart or too close together, this can be bioinformatically detected. Similarly, if a read aligns without a 'mate' (the other part if the pair cannot satisfactorily align to the reference), this can also be flagged. This approach for detecting non-SNV variants is shown in Figure 4Ci and ii and is known as a 'paired-end' (or 'read-pair') approach to detecting SVs. Similarly, the 'split-read' approach uses information that a single read is disrupted, or split, by a SV. The read depth or 'depth of coverage' approach replies upon algorithms detecting regions where there is a significant increase or decrease in coverage (Figure 4D). All of these computational approaches have their limitations for different SVs, and the best algorithms combine more than one approach.(19) Structural variation on a chromosomal level e.g. aneuploidy or ringed chromosomes, will not be detected by WGS and karyotyping should be requested separately.

Repeat expansions, where the number of repeats is critical to the diagnosis, can be challenging to size through WGS; large repeat expansions will be longer than the read, or read-pair (Figure 4Civ). ExpansionHunter is a tool that estimates the repeat size at the loci of known expansions, which when paired with visual inspection, was sensitive and specific for correctly sizing expansions in the 100KGP when the expansion size was less than the read length.(20)

However, there are three important caveats to the above. First, as with virtual panels, if the gene and specifically the expansion (if that is the diagnostic question) is not on the virtual panel, non-SNVs will not be tested. Second, when the expansion is larger than the read length (as seen in *FMR1*, *C9orf72*, *DMPK* (myotonic dystrophy type 1) and *FXN* (Friedreich's Ataxia), although an expansion could be identified, it was often significantly underestimated by ExpansionHunter (Figure 4Civ). Although *RFC1*, the gene recently identified as causing cerebellar ataxia, neuropathy and vestibular areflexia syndrome (CANVAS) through biallelic pentanucleotide repeat expansions, was not examined by Ibañez

et al., the same would apply; the expansion is typically >1000 repeats (>5000 nucleotides).(21) In NHS laboratories *RFC1* is currently tested using non-WGS methods. Third, early iterations of the 100KGP pipeline did not routinely analyse for any non-SNVs, and many were missed and not reported.

MtDNA sequencing is currently requested as a separate test to sequencing of the nuclear genome. Studies have shown that with a satisfactory read depth, WGS can detect mtDNA variants at a heteroplasmy level down to 10%,(22) but if there is a significant suspicion for mitochondrial disease, mtDNA sequencing should be requested separately. Other types of genetic mechanism including epigenetic factors such as DNA methylation or imprinting will not be detected using WGS and should have separate testing requested. Lastly, in the NHS, if a rapid result is critical to guide management, the R14 'Acutely unwell children with a likely monogenic disorder' WGS can be requested for critically ill children and adults, with a turnaround time of 2-3 weeks.(23)

EXAMPLES FROM THE CLINIC

Key to our diagnostic success in the 100KGP was access to the data in the research environment, and regular review of cases at a dedicated clinical-research-genetic MDT. We illustrate with clinical cases practical examples of potential pitfalls discussed above.

Case 1

A woman in her late teens presented with a subacute history of sensory changes in her hands, a few weeks following a viral illness. She developed progressive weakness and wasting of intrinsic hand muscles. At initial assessment she also had mild sub-clinical distal lower limb weakness (Figure 6A-E). There was no family history of neuromuscular illness and parents were non-consanguineous. Initial neurophysiology showed a patchy, widespread, conduction slowing neuropathy. She was treated in her local hospital with intravenous immunoglobulin for presumed chronic inflammatory demyelinating neuropathy. Subsequent CSF examination showed normal constituents, nerve roots were markedly thickened and pathologically enhancing on MRI, and nerve biopsy demonstrated a chronic demyelinating neuropathy without inflammation (Figure 6G). She progressed slowly despite treatment; initial genetic testing, including CMT1A with multiplex ligation-dependent probe amplification (MLPA), and a 14 gene panel in 2015, was negative. She was enrolled into the 100KGP with her parents, with no primary findings. Through a research collaboration we identified the variant c.4271C>T p.(Thr1424Met) in *ITPR3,* a gene only reported in three families and not included in the virtual panel.(24,25) Additionally, the variant was maternally inherited (Figure 6F). Clinically the mother had no symptoms and a completely normal neurological examination, but neurophysiology showed a clear conduction slowing neuropathy. The diagnosis is CMT, with remarkable variability in

severity, due to an *ITPR3* variant. This case highlighted the importance of the assigned affected status; segregation was confirmed but only by neurophysiology. Similarly, research access to the 100KGP data was essential to identify a gene not on the virtual panel but in the literature.

Case 2

A man in his late 60s was referred for a diagnostic opinion. He had a progressive sensory and motor neuropathy since his 20s. Neurophysiology was clearly demyelinating with a median nerve motor conduction velocity of 22 m/s. The family history was of AD disease. His 100KGP primary findings report was negative. We examined the aligned sequence data and discovered 1.5x the read depth in the region of *PMP22* compared with the rest of the genome (Figure 6H). MLPA confirmed the 17p.22 duplication; the diagnosis was CMT1A. The bioinformatic pipeline did not call this common copy number variant seen in CMT. We have now seen 3 cases of CMT1A referred for a diagnostic opinion where the chromosome 17 duplication was either missed or not looked for as clinicians were not aware that next-generation sequencing gene panels and WGS in the 100KGP did not reliably detect the duplication.(26) Despite the panel name 'Hereditary Neuropathy NOT PMP22 copy number', the current WGS panel does now include the *PMP22* duplication, but the first line test in conduction slowing neuropathies should still be 'R77 Hereditary Neuropathy – PMP22 copy number' (MLPA).

Case 3

A man presented in his early 50s with a four-year history of progressive unsteadiness, particularly in the dark, and reduced sensation in his distal limbs. He had a longstanding cough. Examination revealed a sensory ataxia and large and small fibre sensory loss, without weakness. Neurophysiology showed a severe pure sensory axonal neur(on)opathy. Extensive investigations including antibody testing, neuroaxis imaging, positron emission spectroscopy scan, nerve and lip biopsy excluded inflammatory, nutritional, and malignant causes. A 56-gene CMT panel, *FXN* and *POLG* sequencing and 100KGP testing was negative. We examined the aligned WGS sequence data of *RFC1* in the research environment and found a complete drop of read depth within intron 2 (Figure 6I). Subsequent repeat-primed polymerase chain reaction confirmed biallelic AAGGG repeat expansions in *RFC1*, and a diagnosis of CANVAS. This case highlights a missed large intronic repeat expansion, still not reliably called on WGS. Currently *RFC1* testing must be requested separately.

Case 4

A man in his early forties presented with a ten-year history of progressive walking difficulties due to distal lower limb weakness. There was no family history. Examination revealed a length-dependent motor neuropathy; this was confirmed on neurophysiology and there was no slowing or conduction

block. Lead and hexosaminidase A levels were normal. Testing for *AR* expansion, 32-gene CMT2/distal HMN panel and 100KGP were negative. Review in the research environment identified a heterozygous variant in *MME* (c.202C>T p.(Arg68Ter)), a gene known to cause adult onset recessive, motor predominant CMT.(27) The single variant is classed as pathogenic when in *trans* with a second pathogenic variant; this was a single hit in a recessive disease. We then examined the aligned sequence data and identified a 9kbp drop in read depth in *MME*, consistent with a deletion including exons 15 and 16, predicted to be pathogenic (Figure 6J). Both variants were confirmed in the diagnostic laboratory. The diagnosis was distal HMN due to compound heterozygous variants in *MME*; one that was missed because a single recessive variant was not reported, and the SV was not identified by the analysis pipelines.

Case 5

A man in his late teens was assessed as he transitioned to the adult neuropathy clinic. He had a normal birth but began walking with in-turning feet aged four. His feet then began to slap as he developed slowly progressive weakness. His father had mild symptoms compatible with CMT. Examination of the proband revealed relatively mild, length-dependent motor deficits (Figure 6K-L). His neurophysiology showed a sensory and motor demyelinating neuropathy; a clinical diagnosis of CMT1 was made. A 56-gene CMT panel was negative and the 100KGP project had no primary findings. Review of genes not included in the virtual panel used by the 100KGP in the research environment revealed a paternally inherited, previously reported pathogenic variant in the myelin protein gene *PMP2*, confirming the genetic diagnosis (Figure 6M).(28) Despite *PMP2* being established as a cause of CMT in 2016, the gene was not included in the 100KGP panel.(29)

CONCLUSIONS

The diagnostic opportunities through WGS are clear and are reflected in the introduction of WGS into routine NHS diagnostic testing. However, caution must be taken when reading a 'negative' report. WGS has its technical limitations; it very reliably detects SNVs and small indels, and although bioinformatic algorithms are now confidently detecting copy number variants, this wasn't always the case, and detecting balanced SVs and sizing large repeat expansions remains unreliable. Variants are prioritised according to the information provided by the requesting clinician; a detailed phenotypic description and, if applicable, broad use of virtual panels, increases the chances of a correct genetic diagnosis. Family studies increase the diagnostic yield but rely upon correct assignment of disease status of relatives. If a negative report is received but there is high diagnostic suspicion, we encourage discussion with the genetic laboratory and/or an MDT meeting to consider further focused analysis.

Provided the diagnosis of a genetic disorder is correct (excluding mtDNA disorders), although the answer should in theory like within 'whole' genome sequencing, WGS is not always the correct test to request. Lastly, all the cases in this review were diagnosed through research access to 100KGP data; there will always be unsolved and novel causes for neurological disease and the authors feel strongly that clinical genomic researchers should, where their patient has consented, have access to their data to ensure we continue to increase genetic diagnoses for individuals and their families, and advance the field as a whole. Access to research data is not universal, and if after discussion with the local genetic laboratory there is no diagnosis, clinicians should consider referring to a specialist centre.

KEY POINTS

- WGS is the first line test for many, but not all, suspected genetic neurological disorders. Before requesting WGS, clinicians should first ensure relevant initial single genetic tests are negative (e.g. *PMP22* duplication in CMT).

- Gene panels are constantly evolving, and it is important to check which genes and/or type of genetic variant is offered, particularly if there is a specific genetic diagnosis in mind.

- Accurate phenotype information, via HPO terms, and correct assignment of relative affected status, are critical to maximise diagnostic yield. Relative testing is desirable, and sometimes essential.

- Discussion, ideally in an MDT setting, with the genetics laboratory is recommended for selected unsolved cases and where there are unexpected or uncertain results. Where VUSs remain unreported, communication of specific phenotype data may be the key to their reclassification to pathogenic.

FURTHER READING

- 100,000 Genomes Project Pilot Investigators; Smedley D, Smith KR, *et al*. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care – Preliminary Report. N Engl J Med. 2021 Nov 11;385(20):1868-1880.

- Ibañez K, Polke J, Hagelstrom RT, *et al*. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. Lancet Neurol. 2022 Mar;21(3):234-245.

- Moore AR, Yu J, Pei Y, Cheng EWY, *et al*. Use of genome sequencing to hunt for cryptic second-hit variants: analysis of 31 cases recruited to the 100 000 Genomes Project. J Med Genet. 2023 Nov 27;60(12):1235-1244.

- Pipis M, Rossor AM, Laura M, Reilly MM. Next-generation sequencing in Charcot-Marie-Tooth disease: opportunities and challenges. Nat Rev Neurol. 2019 Nov;15(11):644-656.

Box 1 GnomAD

The Genome Aggregation Database (gnomAD, pronounced nō͵mad) is the most widely used population database of genomic variation. Launched in 2014 as the Exome Aggregation Consortium (ExAC), it is now in its fourth iteration (gnomAD v4, released in November 2023). The open access online database contains genomic data from around 730,000 exomes and 76,000 genomes (up to 1.6 million alleles), derived from more than 100 studies in more than 25 countries. The major output is variant frequency data i.e. how many times has a particular variant been observed in this dataset – 'the population'? The genomic data is broadly derived from a mixture of case-control studies, and large biobanks, including more than 400,000 individuals from the UK Biobank; this is not a healthy control database and will contain affected individuals, with a frequency probably no higher than the disease prevalence.

Box 2 Human Phenotype Ontology (HPO) terms

The concept of HPO is straightforward; to standardise the description of a clinical phenotype. HPO terms can include symptoms, examination findings, syndromes, investigation results, disease severity and onset. The NHS-GMS WGS request form requires inputting of at least one, but ideally several, HPO terms for the patient in question. This can be time consuming and seem unnecessary, but detailed clinical information maximises the chances of WGS finding an answer for the patient. Consider the scenario of a patient deemed by the neurologist to have a unique phenotype of ophthalmoplegia (HP:0000602), gastrointestinal dysmotility (HP:0002579), and demyelinating peripheral neuropathy (HP:0007108). These terms inputted together might be very specific for a particular gene (e.g. mitochondrial), and any variant found prioritised for analysis (even if not on the requested panel), and its classification potentially upgraded based on the information provided. Importantly, the term peripheral neuropathy (HP:0009830) provides no meaningful extra information if requesting the Hereditary Neuropathy panel. The absence of a clinical feature can also be recorded and may be relevant e.g. the absence of tremor in a syndrome of Parkinsonism. The clinical assessment by the neurologist can be most powerful tool for refining genetic variants and detailed and specific HPO terms are a way of quantifying this expertise.

COMPETING INTERESTS

None to declare.

ACKNOWLEDGEMENTS

CONTRIBUTORSHIP

FUNDING

ETHICS STATEMENT

All patients are recruited to our ethically approved research study 'Charcot-Marie-Tooth Disease and related disorders: A Natural History Study', reviewed by the London Queen Square Research Ethics Committee (REC No.: 09/H0716/61). Separate individual patient consent was obtained for use of photographs.

DATA AVAILIBILITY

The data that support the findings of this study are available from the corresponding author, upon reasonable request. The data are not publicly available since they contain information that could compromise the privacy of research participants.

Practical Neurology

REFERENCES

1. Jain V, Irving R, Williams A. Genomic testing in neurology. Pract Neurol. 2023 Oct 19;23(5):420–9.

2. Morris HR, Houlden H, Polke J. Whole-genome sequencing. Pract Neurol. 2021;21(4):322–6.

3. Laurá M, Pipis M, Rossor AM, Reilly MM. Charcot-Marie-Tooth disease and related disorders: An evolving landscape. Vol. 32, Current Opinion in Neurology. Lippincott Williams and Wilkins; 2019. p. 641–50.

4. Pipis M, Rossor AM, Laura M, Reilly MM. Next-generation sequencing in Charcot–Marie–Tooth disease: opportunities and challenges. Nat Rev Neurol. 2019 Nov 1;15(11):644–56.

5. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020 Dec 7;21(1):30.

6. Smith AC, Neveling K, Kanagal-Shamanna R. Optical genome mapping for structural variation analysis in hematologic malignancies. Am J Hematol. 2022;97(7):975–82.

7. Pervez MT, Hasnain MJU, Abbas SH, Moustafa MF, Aslam N, Shah SSM. A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. Biomed Res Int. 2022;2022.

8. Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020;583(7814):96–102.

9. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019;51(11):1560–5.

10. Cortese A, Zhu Y, Rebelo AP, Negri S, Courel S, Abreu L, et al. Biallelic mutations in SORD cause a common and potentially treatable hereditary neuropathy with implications for diabetes. Nat Genet. 2020;52(5):473–81.

11. Bhattacharya S, Khadilkar S V., Nalini A, Ganapathy A, Mannan AU, Majumder PP, et al. Mutation Spectrum of GNE Myopathy in the Indian Sub-Continent. J Neuromuscul Dis. 2018 Feb 21;5(1):85–92.

12. Sherman RM, Salzberg SL. Pan-genomics in the human genome era. Nat Rev Genet. 2020 Apr 7;21(4):243–54.

13.   Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617(7960):312–24.

14.   Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, Cipriani V, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. New England Journal of Medicine. 2021;385(20):1868–80.

15.   Robinson PN, Köhler S, Oellrich A, Genetics SM, Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014;24(2):340–8.

16.   Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. 2015 May 8;17(5):405–24.

17.   Ellard S, Baple EL, Callaway A, Berry I, Forrester N, Turnbull C, et al. ACGS best practice guidelines for variant classification in rare disease 2020. Association for Clinical Genomic Science. 2020;

18.   Moore AR, Yu J, Pei Y, Cheng EWY, Taylor Tavares AL, Walker WT, et al. Use of genome sequencing to hunt for cryptic second-hit variants: analysis of 31 cases recruited to the 100 000 Genomes Project. J Med Genet. 2023;1235–44.

19.   Pagnamenta AT, Camps C, Giacopuzzi E, Taylor JM, Hashim M, Calpena E, et al. Structural and non-coding variants increase the diagnostic yield of clinical whole genome sequencing for rare diseases. Genome Med. 2023;15(1):1–25.

20.   Ibañez K, Polke J, Hagelstrom RT, Dolzhenko E, Pasko D, Thomas ERA, et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. Lancet Neurol. 2022;21(3):234–45.

21.   Cortese A, Simone R, Sullivan R, Vandrovcova J, Tariq H, Yan YW, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nat Genet. 2019;51(4):649–58.

22.   Poole O V., Pizzamiglio C, Murphy D, Falabella M, Macken WL, Bugiardini E, et al. Mitochondrial DNA Analysis from Exome Sequencing Data Improves Diagnostic Yield in Neurological Diseases. Ann Neurol. 2021;89(6):1240–7.

23. Williamson SL, Rasanayagam CN, Glover KJ, Baptista J, Naik S, Satodia P, et al. Rapid exome sequencing : revolutionises the management of acutely unwell neonates. European Journal of Pediatricscs. 2021;180:3587–91.

24. Beijer D, Dohrn MF, Rebelo AP, Feely SME, Reilly MM, Scherer SS, et al. PNS 2022 Abstract Supplement. Journal of the Peripheral Nervous System. 2022;27(S3):352–63.

25. Schabhüttl M, Wieland T, Senderek J, Baets J, Timmerman V, De Jonghe P, et al. Whole-exome sequencing in patients with inherited neuropathies: Outcome and challenges. J Neurol. 2014;261(5):970–82.

26. Record CJ, Pipis M, Poh R, Polke JM, Reilly MM. Beware next-generation sequencing gene panels as the first-line genetic test in Charcot-Marie-Tooth disease. J Neurol Neurosurg Psychiatry. 2023 Apr 14;94(4):327–8.

27. Higuchi Y, Hashiguchi A, Yuan J, Yoshimura A, Mitsui J, Ishiura H, et al. Mutations in MME cause an autosomal-recessive Charcot-Marie-Tooth disease type 2. Ann Neurol. 2016;79(4):659–72.

28. Motley WW, Palaima P, Yum SW, Gonzalez MA, Tao F, Wanschitz J V., et al. De novo PMP2 mutations in families with type 1 Charcot-Marie-Tooth disease. Brain. 2016;139(6):1649–56.

29. Hong Y Bin, Joo J, Hyun YS, Kwak G, Choi YR, Yeo HK, et al. A Mutation in PMP2 Causes Dominant Demyelinating Charcot-Marie-Tooth Neuropathy. PLoS Genet. 2016;12(2):1–15.

FIGURE LEGEND

Figure 1 Should I send WGS? The stepwise process required for requesting WGS

Figure 2 Understanding the utility of WGS. CMT Charcot-Marie-Tooth disease

Figure 3 WGS workflow DNA library is prepared through fragmentation of the extracted DNA, followed by amplification on the flow cell. Sequencing then occurs through 'sequencing by synthesis' and the subsequent read data is aggregated in a .fastq file. The data is then processed through a bioinformatic pipeline; initially aligned to the reference genome (resulting in a .bam file) and then variants are identified ('variant calling') yielding a .vcf file; the common output file of WGS. Virtual panels can then be applied, and the resultant variants interpreted by clinical scientists, before a report is issued.

Figure 4 Bioinformatic methods of detecting variants A. Schematic of the normal process of paired-end reads of a single fragment, subsequent read alignment to the reference genome and the resultant read depth B. Visual representation of detection of a single nucleotide variant (SNV). One allele of the individual's DNA contains a SNV. When the aligned reads are analysed, 50% have a nucleotide that is different from the reference, and the variant is 'called'. C. structural variant detecting methods: 'Split-read' and 'read-pair' .i) A deletion in the fragment of DNA means that when the read-pairs are aligned to the reference they will appear too close together (red markers are closer than without the deletion). ii) An insertion means the aligned reads appear too far apart when aligned to the reference (red markers further apart than without the insertion). iii) A translocation results in one half of the read-pair aligning to a different part of the genome iv) A repeat expansion, particularly one longer than the read-pair, will result in one of the pairs sequencing only the repeat region. When alignment is attempted, it may not be able to align anywhere (the other read will be 'unmatched') or may align elsewhere in the genome. D. structural variant detecting method: 'Depth of coverage' – aligned reads of a deleted region identify a length of sequencing with a 50% drop in coverage suggesting a heterozygous deletion i.e. one normal allele and one allele containing a deletion.

Figure 5 Schematic based on the the American College of Medical Genetics and Genomics (ACMG) criteria. Every variant under review has evidence for pathogenicity scrutinised under each of the listed categories.

Figure 6 Clinical cases A-E. Distal limb muscle atrophy in proband of Case 1 F. Integrative Genomics Viewer (IGV) showing heterozygous variant in *ITPR3* (blue arrow) in proband and mother G. Electron microscopy of sural nerve biopsy of Case 1 (proband) showing reduced myelin thickness and onion bulb formation H. IGV showing 1.5x read depth of aligned reads in *PMP22* compared with a region on

chromosome 1 I. IGV shows loss of coverage in intron 2 of RFC1 (red circle) indicating a biallelic repeat expansion in that region. Reads highlighted in red (black arrow) are unmatched pairs J. IGV shows ~9kb 50% drop in coverage (read depth) in MME encompassing exons 15 and 16; corresponding to a heterozygous deletion K-L. Mild distal lower limb atrophy in proband of Case 5 M. IGV shows heterozygous variant in *PMP2* carried by proband and affected father

Lessons and pitfalls of whole genome sequencing

Christopher J Record, Mary M Reilly

Centre for Neuromuscular Diseases, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK

Lead corresponding author: Professor Mary M Reilly, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK. m.reilly@ucl.ac.uk

Secondary corresponding author: Christopher J Record, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK. Chris.record@ucl.ac.uk

4545/4400

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

ABSTRACT

Whole genome sequencing (WGS) has recently become the first line genetic investigation for many suspected genetic neurological disorders. Whilst its diagnostic capabilities are innumerable, as with any test, it has its limitations. Clinicians should be aware of where WGS is extremely reliable (detecting single nucleotide variants), where its reliability is much improved (detecting copy number variants and small repeat expansions) and where it may miss/misinterpret a variant (large repeat expansions, balanced structural variants or low heteroplasmy mitochondrial DNA variants). Bioinformatic technology and virtual gene panels are constantly evolving, and it is important to know what genes and type of variants are being tested; the current NHS Genomic Medicine Service WGS offers more than early iterations of the 100,000 Genomes Project analysis. Close communication between clinician and laboratory, ideally through a multidisciplinary team meeting, is encouraged where there is diagnostic uncertainty.

INTRODUCTION

Whole genome sequencing (WGS) and the use of genomic testing in neurology, including consent, indications and results, have recently been expertly reviewed in Practical Neurology.(1,2) The success of the Genomics England 100,000 Genomes Project (100KGP), sequencing patients with cancer and rare-diseases, has led to the introduction of WGS with virtual panels into routine clinical practice for many neurological diseases via the National Health Service Genomic Medicine Service (NHS-GMS, https://www.england.nhs.uk/genomics/nhs-genomic-med-service/). The theoretical benefits of WGS are clear; sequencing the entire genome (many orders of magnitude more DNA than previous routine testing, at comparable costs) wherein, provided the clinical diagnosis of a genetic disorder is correct, the molecular diagnosis should lie. However, as with every new technology, WGS has its limitations. This article aims to outline the diagnostic utility of WGS, but also where caution needs to be exerted. A critical step in any patient's diagnostic journey is the decision to request WGS. Where appropriate, especially in sporadic cases, acquired diseases should be excluded first. Key points to consider before requesting WGS are highlighted in Figure 1

IS CHARCOT-MARIE-TOOTH DISEASE A GOOD DISEASE PROTOTYPE FOR UNDERSTANDING WGS?

Charcot-Marie-Tooth (CMT) disease is an umbrella term for inherited neuropathies but is a clinically and genetically heterogenous group of diseases. The clinical sub-types of CMT include demyelinating sensory and motor neuropathy (CMT1), axonal sensory motor neuropathy (CMT2), sensory and motor neuropathy with intermediate conduction slowing (upper limb motor conduction velocity between 25 and 45 m/s, CMTi), hereditary sensory neuropathy (HSN), and hereditary motor neuropathy (HMN).(3,4)

The diagnostic utility of WGS for an individual lies in its ability to detect vast numbers, and theoretically different types, of genetic variant. Figure 2 illustrates the features of a disease group that make it suitable for considering WGS testing.

One down-side of CMT as a disease prototype, is that functional validation of novel variants/genes is challenging but this underpins how important WGS is in CMT clinical practice. Gold standard functional evidence would be ex-vivo human diseased tissue demonstrating absent, deficient or dysfunctional protein contributing to pathology. This is theoretically possible with peripheral nerve biopsies, but this is an invasive procedure requiring technical expertise. Alternatively, RNA sequencing can be used to demonstrate aberrant transcripts in appropriate tissues; Schwan cells are clearly easier to study than

dorsal root ganglia or anterior horn cells. Overall, we feel CMT is an excellent disease to demonstrate the lessons and pitfalls of WGS and will explore these herein.

WHOLE GENOME SEQUENCING TECHNOLOGIES

A basic understanding of the molecular techniques involved in WGS is important to appreciate its potential pitfalls. First, WGS when used in common medical parlance, refers to 'short-read' WGS (srWGS). Some useful terminology is highlighted in Table 1. Other forms of genomic sequencing exist, and although currently used mostly in the research setting, their use is increasing in diagnostic genetic laboratories worldwide. Long-read WGS (lrWGS), as suggested in the description, continuously sequences long molecules of DNA, typically tens of kbp in length, but up to many hundreds of kbp depending on the sequencing technology used. The major benefit of lrWGS is the ability to accurately detect and size repeat expansions, and detect complex, balanced structural variants (SVs). The drawbacks include the cost and longer sequencing time, and its error rate on an individual nucleotide level which, when combined with low read depth, affects its ability to reliably detect single nucleotide variants (SNVs) or insertion-deletions (indels).(5) Optical genome mapping (OGM) is another form of genomic interrogation, and more appropriately termed 'genome imaging'. Its uses have been compared to those previously investigated with karyotyping (i.e. large SVs) but with the benefit of up to 20,000-fold higher resolution. DNA molecules are enzymatically labelled, and the resultant ligated DNA then 'imaged' for its pattern of periodically spaced fluorescent signals. Its ability to detected large SVs (0.5 – 1 Mbp) is superior to srWGS and lrWGS, and it is less costly to get higher coverage. As with srWGS and lrWGS it cannot detect aneuploidy (an abnormal number of chromosomes), although this is less relevant in the setting of non-developmental disorders. Another potential drawback of OGM is the requirement for DNA extraction from a fresh blood sample.(6) As neither lrWGS nor OGM are used in standard NHS testing, from this point forward we will not discuss them further, and we will refer to srWGS simply as WGS.

TABLE 1 USEFUL TERMINOLOGY

Alignment	Process by which the sequenced reads of the individual are matched to the appropriate region in the reference genome
Bioinformatic pipeline	Series of computational steps processing raw WGS data resulting in variants for analysis; includes alignment, filtering, variant calling
Copy number variant	A type of structural variant, where there is an increase or decrease in dosage across a region of the genome e.g. deletion or duplication
Insertion-deletion (indel)	A small variant, usually <50bp, where a series of consecutive nucleotides are altered/inserted/deleted

Long-read	A single strand of DNA is sequenced continuously, depending on the technology, up to hundreds of kbp in length
Paired-end reads (read-pair)	A fragment of DNA is sequenced from both ends to create paired-end reads, or a read pair. Information can be gleaned when the pairs are aligned to the reference if they are unexpectedly too close, too far, on separate chromosomes or if the one half of the pair is 'unmatched'
Read	The basic molecular output of next generation sequencing; a read is a single consecutively sequenced strand of DNA, before alignment to the reference genome
Read depth	The number of cumulative reads aligned at a particular genomic locus i.e. how many times a particular nucleotide been sequenced
Read length	The number of nucleotides in a single read
Repeat expansion	Repeating nucleotide motif (e.g. CAG_n – the common polyglutamate expansion, or $AAGGG_n$, the most common configuration in *RFC1* CANVAS) where n is the number of repeats. The number of repeats that is considered pathogenic varies widely between diseases.
Short-read	A single strand of DNA is consecutively sequenced between 75-300 bp in length
Single nucleotide variant (SNV)	The alteration/insertion/deletion of a single nucleotide
Structural variant (SV)	Medium to large (typically 100s to Mbp in length) variants including duplications, deletions, insertions, balanced translocations and more complex rearrangements
Variant calling	The process by which alterations (variants) in the individual's genome are identified by comparing to the reference genome

NEXT-GENERATION SEQUENCING TECHNOLOGY

WGS uses next-generation sequencing (NGS) technology, also known as high-throughput or massively-parallel sequencing. NGS has been used for many years in clinical diagnostic laboratories for the sequencing of disease-specific gene panels and whole exome sequencing. There are a number of sequencing platforms,(7) but the dominant provider worldwide is Illumina, which is also used by NHS-GMS, and the process described hereafter. A flow diagram of the process involved is shown in Figure 3.(8)

The first step is **library preparation** (Figure 3A); the genomic DNA library is a series of short fragments ready for sequencing. The DNA (typically extracted from leukocytes in blood; purple EDTA tube) is fragmented and then each fragment amplified. Fragments are then **sequenced** in a process called 'sequencing by synthesis', whereby fluorescently tagged nucleotides are added to a linear single strand of DNA complementary to the fragment; the resultant fluorescent DNA strand is known as a 'read' and can be sequenced by its characteristic spectral emission (one wavelength for each of the four nucleotides, Figure 3B). The fragment is sequenced from both ends forming 'paired-end reads',

allowing additional information to be gleaned when the reads are aligned. Data is then fed into the **bioinformatic pipeline** (Figure 3C). The millions of reads are **aligned** to the reference genome, which when visually represented, form piles of overlapping reads. The overall coverage of the WGS describes what proportion of the reference genome is sequenced to a satisfactory read depth. Figure 4A shows in detail how an unmutated fragment is sequenced and aligned to the reference.

Variant calling is the process of identifying variants i.e. variation in an individual's genome when compared with the reference. The basic output of a WGS bioinformatic pipeline is the identification of small variants; alteration/insertion/deletion of single nucleotides (SNVs, Figure 4B) or a small number of consecutive nucleotides (indels). The universal final output for the millions of variants generated is a .vcf file. Other types of genetic variant can also be detected including SVs (both copy number variants, and balanced rearrangements; the latter where there is no change in dosage at a particular locus), repeat expansions and mitochondrial DNA (mtDNA) variants, but their detection and calling is variable (Figure 4 and see *When WGS might not be the correct test*).

After variant calling, the variants are **filtered** according to specified criteria (see *Filtering and prioritisation*). Application of a **virtual panel**(s) may yield possible candidate variants, which are **interpreted** by clinical scientists (Figure 3D). If there is ambiguity or uncertainty, results are ideally discussed at a multidisciplinary team (MDT) meeting, following which a genetic report can be issued.

VIRTUAL PANELS

Although WGS theoretically allows analysis of variants from an individual's entire genome, this is neither desirable (incidental unwanted findings) nor practical (a human genome contains approximately five million SNVs) therefore virtual panels are essential to refine the search. In the NHS, clinicians are required to select virtual gene panel(s) when requesting WGS. The NHS-GMS PanelApp (https://nhsgms-panelapp.genomicsengland.co.uk/panels) is a publicly available resource that utilises genetic expertise through crowdsourcing to curate disease-specific gene panels.(9) For a gene to be included it needs to be approved as 'green' by a number of verified experts; a green gene is broadly one in which plausible disease-causing variants have been found in three or more unrelated individuals/families. However, the panels can only be as correct and up-to-date as their reviewers and the current available evidence. For example, *SORD* was discovered as a common, and potentially treatable, cause of CMT in 2020,(10) but was not approved as a green gene until November 2022. Panels are periodically updated, and previous iterations can be found on PanelApp. Genes that cause a complex phenotype which include the disease group of interest e.g. *ABHD12* causing polyneuropathy, hearing loss, ataxia, retinitis pigmentosa, and cataract (PHARC) syndrome, are often not included if the panel specifies an isolated phenotype; it is not a green gene on the current

'Hereditary Neuropathy or pain disorder' panel (R78 version 3.24). Similarly, novel, rare genes may not meet green inclusion criteria. It is therefore important to have an understanding of which genes are tested in a specific panel, and if there is a particular gene of interest in a clinical case, this should be discussed with the genetic laboratory. It is currently recommended that broad rather than narrow use of panels is applied to maximise chances of identifying causative variants.

FILTERING AND PRIORITISATION

Refining the vast number of variants detected through WGS requires filtering strategies. The two most powerful tools are the allele frequency of the variant in reference databases (the most commonly used is gnomAD; https://gnomad.broadinstitute.org/, Box 1) and in family studies, the inheritance pattern, as defined by relative disease status.

Population allele frequency

Historically the upper limit for the population allele frequencies was set at < 1 in 100 for autosomal recessive, and < 1 in 1000 for autosomal dominant (AD) disease, however we know that for most rare diseases these thresholds are far too high. A useful online calculator for the estimation of a disease-specific population allele count and frequency is found at https://cardiodb.org/allelefrequencyapp/. It is important to remember that if a variant seemingly occurs at too high an allele frequency, it will be filtered by the bioinformatic pipeline, and not considered for interpretation. The most common variant c.757delG in *SORD*-related CMT is present in a highly homologous non-functioning pseudogene *SORD2P* in 95% of controls; the two variants can be challenging to delineate bioinformatically and therefore the *SORD* variant is potentially inappropriately filtered.(10) This problem with this particular variant has been overcome but was a barrier to its discovery.

One must also be wary of regional 'hotspots' for particular variants. The *GNE* variant p.Val696Met (previously p.Val727Met) causing the rare recessive hereditary inclusion body myopathy/Nonaka myopathy is exceedingly common in the South Asian population where the majority of the disease is seen.(11) The overall quoted allele frequency appears too high for the prevalence of the disease in the UK, and may result in the variant being discounted. Only when the regional breakdown is examined, can it be appreciated that the variant is very rare in European populations, in keeping with disease prevalence.

Reference genome

The current human reference genome, denoted GRCh38, originates from the genomes of 20 anonymous volunteers from the USA. It has been shown that two-thirds is comprised of the genome

of a single individual of mixed European and African descent.(12) It is widely recognised that the current reference genome has significant limitations; it contains some gaps (~5%), has regions of unreliable coverage (e.g. around the centromere), and reflects a very narrow ancestry. The Human Pangenome Reference consortium have set out to rectify the flaws in the current reference by creating a new reference built from 350 human genomes, and have recently published a draft from 47 individuals from diverse backgrounds.(13) Until the 'Pangenome' comes into routine clinical practice, clinicians must we aware that patients from certain ethnic backgrounds (e.g. the Indian subcontinent) may have variants missed because the reference does not reflect their ancestry.

Family studies and relative disease status

Variant segregation through family studies (WGS in more than one family member that are subsequently analysed together) enhances diagnostic success.(14) At recruitment, participants are assigned as affected, unaffected or unknown. Downstream in the process, if a dominant variant is detected in the affected proband and a reportedly unaffected parent, it will be disregarded or deprioritised. Therefore, caution should be exerted, when the disease has an adult-onset or a variable presentation, that relatives' disease status is appropriately assigned.

Human phenotype ontology terms

As part of the process of requesting WGS through the NHS-GMS, the clinician is required to include Human Phenotype Ontology (HPO) terms (https://hpo.jax.org/app/, Box 2). These phenotypic descriptors can be used to prioritise variants using Exomiser, a programme utilised by NHS-GMS.(15) For example, a male patient with a demyelinating neuropathy and upper motor neurone signs underwent WGS in the 100KGP with the Hereditary Neuropathy virtual panel applied. There were no candidate variants from the panel, but because the HPO terms included 'demyelinating neuropathy' and 'Babinski sign', a variant in *ABCD1*, known to cause X-linked myeloneuropathy, was identified. Subsequent discussion at our MDT, and further clinical and laboratory assessments, confirmed this to be the causative gene. This gene is not present in the current Hereditary Neuropathy panel.

VARIANT INTERPRETATION AND REPORTING

Every candidate variant is classified according to established criteria. UK laboratories use the American College of Medical Genetics and Genomics (ACMG) and Association for Clinical Genomic Service (ACGS) guidelines.(16,17) Any given variant, with no supporting data, starts as a 'variant of uncertain significance' (VUS). Evidence is combined, from different categories (including data on allele frequency, functional studies, segregation and prior literature reports) to upgrade the variant as likely

pathogenic or pathogenic, or downgrade to likely benign or benign (Figure 5). As with gene panels, variant interpretation is reliant upon the available evidence, and its application, and therefore variant classification may differ between laboratories. Ideally, clinicians will have access to an MDT (with clinical scientists) to discuss WGS results of unsolved cases, cases with unexpected pathogenic variants, or those with a very typical phenotype for a particular gene, in which no variants have been reported. There is a criterion within the ACMG/ACGS (PP4) criteria that uses phenotype specificity to upgrade variants e.g. absence of dystrophin in a muscle biopsy in a male patient with muscular dystrophy phenotype, when considering a variant in *DMD*. Without the communication of clinical information from clinician to laboratory, the variant might remain a VUS.

Historically, relevant VUSs were listed as an addendum to genetic reports. However, NHS-GMS have adopted guidance from the ACGS that VUSs should not be reported unless exceptional circumstances apply, after a discussion at an MDT meeting. This change is critical for practising clinicians to be aware of. The rationale is that reporting a VUS may lead to confusion on the part of referring clinician or patient, misinterpretation and potentially misdiagnosis. Even when a VUS is likely to be causative, family screening for the variant would still need careful discussion and counselling, and preimplantation genetic testing or entry into a clinical drug trial would only be considered in exceptional circumstances.

However, we have experience that transparent reporting of VUSs to clinicians with genetic expertise, has been vital in clinching a genetic diagnosis with the passage of time. A 'warm' VUS may be upgraded to pathogenic following, for example, a new publication implicating the gene/variant in disease. Without information about VUSs made available on a genetic report, such cases may remain unsolved.

Another example of the need for careful reporting is the presence of a single pathogenic variant in a recessive gene, a so called 'single hit'. If reported, it should be made clear that the diagnosis is not confirmed, but a single pathogenic variant has been detected. With a suggestive phenotype, a 'single hit' will often trigger a discussion with the laboratory to look on the other allele for deep intronic variants (that might affect splicing or create pseudoexons), or SVs (i.e. deletion of a portion of the gene), or explorative analysis of the genome in a research setting.(18)

<mark>WHEN WGS MIGHT NOT BE THE CORRECT TEST</mark>

<mark>The essential first step for genetic testing is ensuring the right test is sent. Jain *et al*. have previously discussed this in detail.(1) In the UK clinicians must consult the NHS Genomic Test Directory https://www.england.nhs.uk/publication/national-genomic-test-directories/.</mark> Many neurological diseases, including some that are treatable, have their molecular basis in non-SNV genetic variation.

Huntington's disease (CAG trinucleotide repeat expansion in *HTT*), genetic motor neurone disease/frontotemporal dementia (GGGGCC hexanucleotide repeat expansion in *C9ORF72*), spinal muscular atrophy (biallelic deletion of exon 7 +/- 8 in *SMN1*), fragile X syndrome (CCG trinucleotide repeat expansion in *FMR1*) and Duchenne muscular dystrophy, (~60% caused by exon-level deletions in the X-linked *DMD*) are all caused by either repeat expansion or SVs. More than 50% of CMT is caused by a duplication of *PMP22*, and the remainder, a mixture of genetic variant types.

The limitation of WGS to accurately detect SVs and repeat expansions lies in the read length. Put simply, it is difficult to quantify a variant with genomic size potentially orders of magnitude larger than the unit of measurement. Figure 4 details the use of paired-end reads in the sequencing and alignment process, and how they can be used to detect non-SNV variants. When the DNA fragment is sequenced from both ends, the two paired-end reads contain markers that identify them as a pair. If, when the reads are aligned to the reference genome, they align too far apart or too close together, this can be bioinformatically detected. Similarly, if a read aligns without a 'mate' (the other part if the pair cannot satisfactorily align to the reference), this can also be flagged. This approach for detecting non-SNV variants is shown in Figure 4Ci and ii and is known as a 'paired-end' (or 'read-pair') approach to detecting SVs. Similarly, the 'split-read' approach uses information that a single read is disrupted, or split, by a SV. The read depth or 'depth of coverage' approach replies upon algorithms detecting regions where there is a significant increase or decrease in coverage (Figure 4D). All of these computational approaches have their limitations for different SVs, and the best algorithms combine more than one approach.(19) Structural variation on a chromosomal level e.g. aneuploidy or ringed chromosomes, will not be detected by WGS and karyotyping should be requested separately.

Repeat expansions, where the number of repeats is critical to the diagnosis, can be challenging to size through WGS; large repeat expansions will be longer than the read, or read-pair (Figure 4Civ). ExpansionHunter is a tool that estimates the repeat size at the loci of known expansions, which when paired with visual inspection, was sensitive and specific for correctly sizing expansions in the 100KGP when the expansion size was less than the read length.(20)

However, there are three important caveats to the above. First, as with virtual panels, if the gene and specifically the expansion (if that is the diagnostic question) is not on the virtual panel, non-SNVs will not be tested. Second, when the expansion is larger than the read length (as seen in *FMR1*, *C9orf72*, *DMPK* (myotonic dystrophy type 1) and *FXN* (Friedreich's Ataxia), although an expansion could be identified, it was often significantly underestimated by ExpansionHunter (Figure 4Civ). Although *RFC1*, the gene recently identified as causing cerebellar ataxia, neuropathy and vestibular areflexia syndrome (CANVAS) through biallelic pentanucleotide repeat expansions, was not examined by Ibañez

et al., the same would apply; the expansion is typically >1000 repeats (>5000 nucleotides).(21) In NHS laboratories *RFC1* is currently tested using non-WGS methods. Third, early iterations of the 100KGP pipeline did not routinely analyse for any non-SNVs, and many were missed and not reported.

MtDNA sequencing is currently requested as a separate test to sequencing of the nuclear genome. Studies have shown that with a satisfactory read depth, WGS can detect mtDNA variants at a heteroplasmy level down to 10%,(22) but if there is a significant suspicion for mitochondrial disease, mtDNA sequencing should be requested separately. Other types of genetic mechanism including epigenetic factors such as DNA methylation or imprinting will not be detected using WGS and should have separate testing requested. Lastly, in the NHS, if a rapid result is critical to guide management, the R14 'Acutely unwell children with a likely monogenic disorder' WGS can be requested for critically ill children and adults, with a turnaround time of 2-3 weeks.(23)

EXAMPLES FROM THE CLINIC

Key to our diagnostic success in the 100KGP was access to the data in the research environment, and regular review of cases at a dedicated clinical-research-genetic MDT. We illustrate with clinical cases practical examples of potential pitfalls discussed above.

Case 1

A woman in her late teens presented with a subacute history of sensory changes in her hands, a few weeks following a viral illness. She developed progressive weakness and wasting of intrinsic hand muscles. At initial assessment she also had mild sub-clinical distal lower limb weakness (Figure 6A-E). There was no family history of neuromuscular illness and parents were non-consanguineous. Initial neurophysiology showed a patchy, widespread, conduction slowing neuropathy. She was treated in her local hospital with intravenous immunoglobulin for presumed chronic inflammatory demyelinating neuropathy. Subsequent CSF examination showed normal constituents, nerve roots were markedly thickened and pathologically enhancing on MRI, and nerve biopsy demonstrated a chronic demyelinating neuropathy without inflammation (Figure 6G). She progressed slowly despite treatment; initial genetic testing, including CMT1A with multiplex ligation-dependent probe amplification (MLPA), and a 14 gene panel in 2015, was negative. She was enrolled into the 100KGP with her parents, with no primary findings. Through a research collaboration we identified the variant c.4271C>T p.(Thr1424Met) in *ITPR3,* a gene only reported in three families and not included in the virtual panel.(24,25) Additionally, the variant was maternally inherited (Figure 6F). Clinically the mother had no symptoms and a completely normal neurological examination, but neurophysiology showed a clear conduction slowing neuropathy. The diagnosis is CMT, with remarkable variability in

severity, due to an *ITPR3* variant. This case highlighted the importance of the assigned affected status; segregation was confirmed but only by neurophysiology. Similarly, research access to the 100KGP data was essential to identify a gene not on the virtual panel but in the literature.

Case 2

A man in his late 60s was referred for a diagnostic opinion. He had a progressive sensory and motor neuropathy since his 20s. Neurophysiology was clearly demyelinating with a median nerve motor conduction velocity of 22 m/s. The family history was of AD disease. His 100KGP primary findings report was negative. We examined the aligned sequence data and discovered 1.5x the read depth in the region of *PMP22* compared with the rest of the genome (Figure 6H). MLPA confirmed the 17p.22 duplication; the diagnosis was CMT1A. The bioinformatic pipeline did not call this common copy number variant seen in CMT. We have now seen 3 cases of CMT1A referred for a diagnostic opinion where the chromosome 17 duplication was either missed or not looked for as clinicians were not aware that next-generation sequencing gene panels and WGS in the 100KGP did not reliably detect the duplication.(26) Despite the panel name 'Hereditary Neuropathy NOT PMP22 copy number', the current WGS panel does now include the *PMP22* duplication, but the first line test in conduction slowing neuropathies should still be 'R77 Hereditary Neuropathy – PMP22 copy number' (MLPA).

Case 3

A man presented in his early 50s with a four-year history of progressive unsteadiness, particularly in the dark, and reduced sensation in his distal limbs. He had a longstanding cough. Examination revealed a sensory ataxia and large and small fibre sensory loss, without weakness. Neurophysiology showed a severe pure sensory axonal neur(on)opathy. Extensive investigations including antibody testing, neuroaxis imaging, positron emission spectroscopy scan, nerve and lip biopsy excluded inflammatory, nutritional, and malignant causes. A 56-gene CMT panel, *FXN* and *POLG* sequencing and 100KGP testing was negative. We examined the aligned WGS sequence data of *RFC1* in the research environment and found a complete drop of read depth within intron 2 (Figure 6I). Subsequent repeat-primed polymerase chain reaction confirmed biallelic AAGGG repeat expansions in *RFC1*, and a diagnosis of CANVAS. This case highlights a missed large intronic repeat expansion, still not reliably called on WGS. Currently *RFC1* testing must be requested separately.

Case 4

A man in his early forties presented with a ten-year history of progressive walking difficulties due to distal lower limb weakness. There was no family history. Examination revealed a length-dependent motor neuropathy; this was confirmed on neurophysiology and there was no slowing or conduction

block. Lead and hexosaminidase A levels were normal. Testing for *AR* expansion, 32-gene CMT2/distal HMN panel and 100KGP were negative. Review in the research environment identified a heterozygous variant in *MME* (c.202C>T p.(Arg68Ter)), a gene known to cause adult onset recessive, motor predominant CMT.(27) The single variant is classed as pathogenic when in *trans* with a second pathogenic variant; this was a single hit in a recessive disease. We then examined the aligned sequence data and identified a 9kbp drop in read depth in *MME*, consistent with a deletion including exons 15 and 16, predicted to be pathogenic (Figure 6J). Both variants were confirmed in the diagnostic laboratory. The diagnosis was distal HMN due to compound heterozygous variants in *MME*; one that was missed because a single recessive variant was not reported, and the SV was not identified by the analysis pipelines.

Case 5

A man in his late teens was assessed as he transitioned to the adult neuropathy clinic. He had a normal birth but began walking with in-turning feet aged four. His feet then began to slap as he developed slowly progressive weakness. His father had mild symptoms compatible with CMT. Examination of the proband revealed relatively mild, length-dependent motor deficits (Figure 6K-L). His neurophysiology showed a sensory and motor demyelinating neuropathy; a clinical diagnosis of CMT1 was made. A 56-gene CMT panel was negative and the 100KGP project had no primary findings. Review of genes not included in the virtual panel used by the 100KGP in the research environment revealed a paternally inherited, previously reported pathogenic variant in the myelin protein gene *PMP2*, confirming the genetic diagnosis (Figure 6M).(28) Despite *PMP2* being established as a cause of CMT in 2016, the gene was not included in the 100KGP panel.(29)

CONCLUSIONS

The diagnostic opportunities through WGS are clear and are reflected in the introduction of WGS into routine NHS diagnostic testing. However, caution must be taken when reading a 'negative' report. WGS has its technical limitations; it very reliably detects SNVs and small indels, and although bioinformatic algorithms are now confidently detecting copy number variants, this wasn't always the case, and detecting balanced SVs and sizing large repeat expansions remains unreliable. Variants are prioritised according to the information provided by the requesting clinician; a detailed phenotypic description and, if applicable, broad use of virtual panels, increases the chances of a correct genetic diagnosis. Family studies increase the diagnostic yield but rely upon correct assignment of disease status of relatives. If a negative report is received but there is high diagnostic suspicion, we encourage discussion with the genetic laboratory and/or an MDT meeting to consider further focused analysis.

Provided the diagnosis of a genetic disorder is correct (excluding mtDNA disorders), although the answer should in theory like within 'whole' genome sequencing, WGS is not always the correct test to request. Lastly, all the cases in this review were diagnosed through research access to 100KGP data; there will always be unsolved and novel causes for neurological disease and the authors feel strongly that clinical genomic researchers should, where their patient has consented, have access to their data to ensure we continue to increase genetic diagnoses for individuals and their families, and advance the field as a whole. Access to research data is not universal, and if after discussion with the local genetic laboratory there is no diagnosis, clinicians should consider referring to a specialist centre.

KEY POINTS

- WGS is the first line test for many, but not all, suspected genetic neurological disorders. Before requesting WGS, clinicians should first ensure relevant initial single genetic tests are negative (e.g. *PMP22* duplication in CMT).

- Gene panels are constantly evolving, and it is important to check which genes and/or type of genetic variant is offered, particularly if there is a specific genetic diagnosis in mind.

- Accurate phenotype information, via HPO terms, and correct assignment of relative affected status, are critical to maximise diagnostic yield. Relative testing is desirable, and sometimes essential.

- Discussion, ideally in an MDT setting, with the genetics laboratory is recommended for selected unsolved cases and where there are unexpected or uncertain results. Where VUSs remain unreported, communication of specific phenotype data may be the key to their reclassification to pathogenic.

FURTHER READING

- 100,000 Genomes Project Pilot Investigators; Smedley D, Smith KR, *et al*. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care – Preliminary Report. N Engl J Med. 2021 Nov 11;385(20):1868-1880.

- Ibañez K, Polke J, Hagelstrom RT, *et al*. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. Lancet Neurol. 2022 Mar;21(3):234-245.

- Moore AR, Yu J, Pei Y, Cheng EWY, *et al*. Use of genome sequencing to hunt for cryptic second-hit variants: analysis of 31 cases recruited to the 100 000 Genomes Project. J Med Genet. 2023 Nov 27;60(12):1235-1244.

- Pipis M, Rossor AM, Laura M, Reilly MM. Next-generation sequencing in Charcot-Marie-Tooth disease: opportunities and challenges. Nat Rev Neurol. 2019 Nov;15(11):644-656.

Box 1 GnomAD

The Genome Aggregation Database (gnomAD, pronounced nō̧mad) is the most widely used population database of genomic variation. Launched in 2014 as the Exome Aggregation Consortium (ExAC), it is now in its fourth iteration (gnomAD v4, released in November 2023). The open access online database contains genomic data from around 730,000 exomes and 76,000 genomes (up to 1.6 million alleles), derived from more than 100 studies in more than 25 countries. The major output is variant frequency data i.e. how many times has a particular variant been observed in this dataset – 'the population'? The genomic data is broadly derived from a mixture of case-control studies, and large biobanks, including more than 400,000 individuals from the UK Biobank; this is not a healthy control database and will contain affected individuals, with a frequency probably no higher than the disease prevalence.

Box 2 Human Phenotype Ontology (HPO) terms

The concept of HPO is straightforward; to standardise the description of a clinical phenotype. HPO terms can include symptoms, examination findings, syndromes, investigation results, disease severity and onset. The NHS-GMS WGS request form requires inputting of at least one, but ideally several, HPO terms for the patient in question. This can be time consuming and seem unnecessary, but detailed clinical information maximises the chances of WGS finding an answer for the patient. Consider the scenario of a patient deemed by the neurologist to have a unique phenotype of ophthalmoplegia (HP:0000602), gastrointestinal dysmotility (HP:0002579), and demyelinating peripheral neuropathy (HP:0007108). These terms inputted together might be very specific for a particular gene (e.g. mitochondrial), and any variant found prioritised for analysis (even if not on the requested panel), and its classification potentially upgraded based on the information provided. Importantly, the term peripheral neuropathy (HP:0009830) provides no meaningful extra information if requesting the Hereditary Neuropathy panel. The absence of a clinical feature can also be recorded and may be relevant e.g. the absence of tremor in a syndrome of Parkinsonism. The clinical assessment by the neurologist can be most powerful tool for refining genetic variants and detailed and specific HPO terms are a way of quantifying this expertise.

COMPETING INTERESTS

None to declare.

ACKNOWLEDGEMENTS

CONTRIBUTORSHIP

CJR analysed the data and wrote the manuscript. MMR conceptualised the study and provided senior critical review and revisions.

FUNDING

None to declare.

ETHICS STATEMENT

All patients are recruited to our ethically approved research study 'Charcot-Marie-Tooth Disease and related disorders: A Natural History Study', reviewed by the London Queen Square Research Ethics Committee (REC No.: 09/H0716/61). Separate individual patient consent was obtained for use of photographs.

DATA AVAILIBILITY

The data that support the findings of this study are available from the corresponding author, upon reasonable request. The data are not publicly available since they contain information that could compromise the privacy of research participants.

REFERENCES

1. Jain V, Irving R, Williams A. Genomic testing in neurology. Pract Neurol. 2023 Oct 19;23(5):420–9.

2. Morris HR, Houlden H, Polke J. Whole-genome sequencing. Pract Neurol. 2021;21(4):322–6.

3. Laurá M, Pipis M, Rossor AM, Reilly MM. Charcot-Marie-Tooth disease and related disorders: An evolving landscape. Vol. 32, Current Opinion in Neurology. Lippincott Williams and Wilkins; 2019. p. 641–50.

4. Pipis M, Rossor AM, Laura M, Reilly MM. Next-generation sequencing in Charcot–Marie–Tooth disease: opportunities and challenges. Nat Rev Neurol. 2019 Nov 1;15(11):644–56.

5. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020 Dec 7;21(1):30.

6. Smith AC, Neveling K, Kanagal-Shamanna R. Optical genome mapping for structural variation analysis in hematologic malignancies. Am J Hematol. 2022;97(7):975–82.

7. Pervez MT, Hasnain MJU, Abbas SH, Moustafa MF, Aslam N, Shah SSM. A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. Biomed Res Int. 2022;2022.

8. Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020;583(7814):96–102.

9. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019;51(11):1560–5.

10. Cortese A, Zhu Y, Rebelo AP, Negri S, Courel S, Abreu L, et al. Biallelic mutations in SORD cause a common and potentially treatable hereditary neuropathy with implications for diabetes. Nat Genet. 2020;52(5):473–81.

11. Bhattacharya S, Khadilkar S V., Nalini A, Ganapathy A, Mannan AU, Majumder PP, et al. Mutation Spectrum of GNE Myopathy in the Indian Sub-Continent. J Neuromuscul Dis. 2018 Feb 21;5(1):85–92.

12. Sherman RM, Salzberg SL. Pan-genomics in the human genome era. Nat Rev Genet. 2020 Apr 7;21(4):243–54.

13.   Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617(7960):312–24.

14.   Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, Cipriani V, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. New England Journal of Medicine. 2021;385(20):1868–80.

15.   Robinson PN, Köhler S, Oellrich A, Genetics SM, Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014;24(2):340–8.

16.   Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. 2015 May 8;17(5):405–24.

17.   Ellard S, Baple EL, Callaway A, Berry I, Forrester N, Turnbull C, et al. ACGS best practice guidelines for variant classification in rare disease 2020. Association for Clinical Genomic Science. 2020;

18.   Moore AR, Yu J, Pei Y, Cheng EWY, Taylor Tavares AL, Walker WT, et al. Use of genome sequencing to hunt for cryptic second-hit variants: analysis of 31 cases recruited to the 100 000 Genomes Project. J Med Genet. 2023;1235–44.

19.   Pagnamenta AT, Camps C, Giacopuzzi E, Taylor JM, Hashim M, Calpena E, et al. Structural and non-coding variants increase the diagnostic yield of clinical whole genome sequencing for rare diseases. Genome Med. 2023;15(1):1–25.

20.   Ibañez K, Polke J, Hagelstrom RT, Dolzhenko E, Pasko D, Thomas ERA, et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. Lancet Neurol. 2022;21(3):234–45.

21.   Cortese A, Simone R, Sullivan R, Vandrovcova J, Tariq H, Yan YW, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nat Genet. 2019;51(4):649–58.

22.   Poole O V., Pizzamiglio C, Murphy D, Falabella M, Macken WL, Bugiardini E, et al. Mitochondrial DNA Analysis from Exome Sequencing Data Improves Diagnostic Yield in Neurological Diseases. Ann Neurol. 2021;89(6):1240–7.

23. Williamson SL, Rasanayagam CN, Glover KJ, Baptista J, Naik S, Satodia P, et al. Rapid exome sequencing : revolutionises the management of acutely unwell neonates. European Journal of Pediatricscs. 2021;180:3587–91.

24. Beijer D, Dohrn MF, Rebelo AP, Feely SME, Reilly MM, Scherer SS, et al. PNS 2022 Abstract Supplement. Journal of the Peripheral Nervous System. 2022;27(S3):352–63.

25. Schabhüttl M, Wieland T, Senderek J, Baets J, Timmerman V, De Jonghe P, et al. Whole-exome sequencing in patients with inherited neuropathies: Outcome and challenges. J Neurol. 2014;261(5):970–82.

26. Record CJ, Pipis M, Poh R, Polke JM, Reilly MM. Beware next-generation sequencing gene panels as the first-line genetic test in Charcot-Marie-Tooth disease. J Neurol Neurosurg Psychiatry. 2023 Apr 14;94(4):327–8.

27. Higuchi Y, Hashiguchi A, Yuan J, Yoshimura A, Mitsui J, Ishiura H, et al. Mutations in MME cause an autosomal-recessive Charcot-Marie-Tooth disease type 2. Ann Neurol. 2016;79(4):659–72.

28. Motley WW, Palaima P, Yum SW, Gonzalez MA, Tao F, Wanschitz J V., et al. De novo PMP2 mutations in families with type 1 Charcot-Marie-Tooth disease. Brain. 2016;139(6):1649–56.

29. Hong Y Bin, Joo J, Hyun YS, Kwak G, Choi YR, Yeo HK, et al. A Mutation in PMP2 Causes Dominant Demyelinating Charcot-Marie-Tooth Neuropathy. PLoS Genet. 2016;12(2):1–15.

FIGURE LEGEND

Figure 1 Should I send WGS? The stepwise process required for requesting WGS

Figure 2 Understanding the utility of WGS. CMT Charcot-Marie-Tooth disease

Figure 3 WGS workflow DNA library is prepared through fragmentation of the extracted DNA, followed by amplification on the flow cell. Sequencing then occurs through 'sequencing by synthesis' and the subsequent read data is aggregated in a .fastq file. The data is then processed through a bioinformatic pipeline; initially aligned to the reference genome (resulting in a .bam file) and then variants are identified ('variant calling') yielding a .vcf file; the common output file of WGS. Virtual panels can then be applied, and the resultant variants interpreted by clinical scientists, before a report is issued.

Figure 4 Bioinformatic methods of detecting variants A. Schematic of the normal process of paired-end reads of a single fragment, subsequent read alignment to the reference genome and the resultant read depth B. Visual representation of detection of a single nucleotide variant (SNV). One allele of the individual's DNA contains a SNV. When the aligned reads are analysed, 50% have a nucleotide that is different from the reference, and the variant is 'called'. C. structural variant detecting methods: 'Split-read' and 'read-pair' .i) A deletion in the fragment of DNA means that when the read-pairs are aligned to the reference they will appear too close together (red markers are closer than without the deletion). ii) An insertion means the aligned reads appear too far apart when aligned to the reference (red markers further apart than without the insertion). iii) A translocation results in one half of the read-pair aligning to a different part of the genome iv) A repeat expansion, particularly one longer than the read-pair, will result in one of the pairs sequencing only the repeat region. When alignment is attempted, it may not be able to align anywhere (the other read will be 'unmatched') or may align elsewhere in the genome. D. structural variant detecting method: 'Depth of coverage' – aligned reads of a deleted region identify a length of sequencing with a 50% drop in coverage suggesting a heterozygous deletion i.e. one normal allele and one allele containing a deletion.

Figure 5 Schematic based on the the American College of Medical Genetics and Genomics (ACMG) criteria. Every variant under review has evidence for pathogenicity scrutinised under each of the listed categories.

Figure 6 Clinical cases A-E. Distal limb muscle atrophy in proband of Case 1 F. Integrative Genomics Viewer (IGV) showing heterozygous variant in *ITPR3* (blue arrow) in proband and mother G. Electron microscopy of sural nerve biopsy of Case 1 (proband) showing reduced myelin thickness and onion bulb formation H. IGV showing 1.5x read depth of aligned reads in *PMP22* compared with a region on

chromosome 1 I. IGV shows loss of coverage in intron 2 of RFC1 (red circle) indicating a biallelic repeat expansion in that region. Reads highlighted in red (black arrow) are unmatched pairs J. IGV shows ~9kb 50% drop in coverage (read depth) in MME encompassing exons 15 and 16; corresponding to a heterozygous deletion K-L. Mild distal lower limb atrophy in proband of Case 5 M. IGV shows heterozygous variant in *PMP2* carried by proband and affected father

Figure 1 Should I send WGS? The stepwise process required for requesting WGS

158x158mm (300 x 300 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

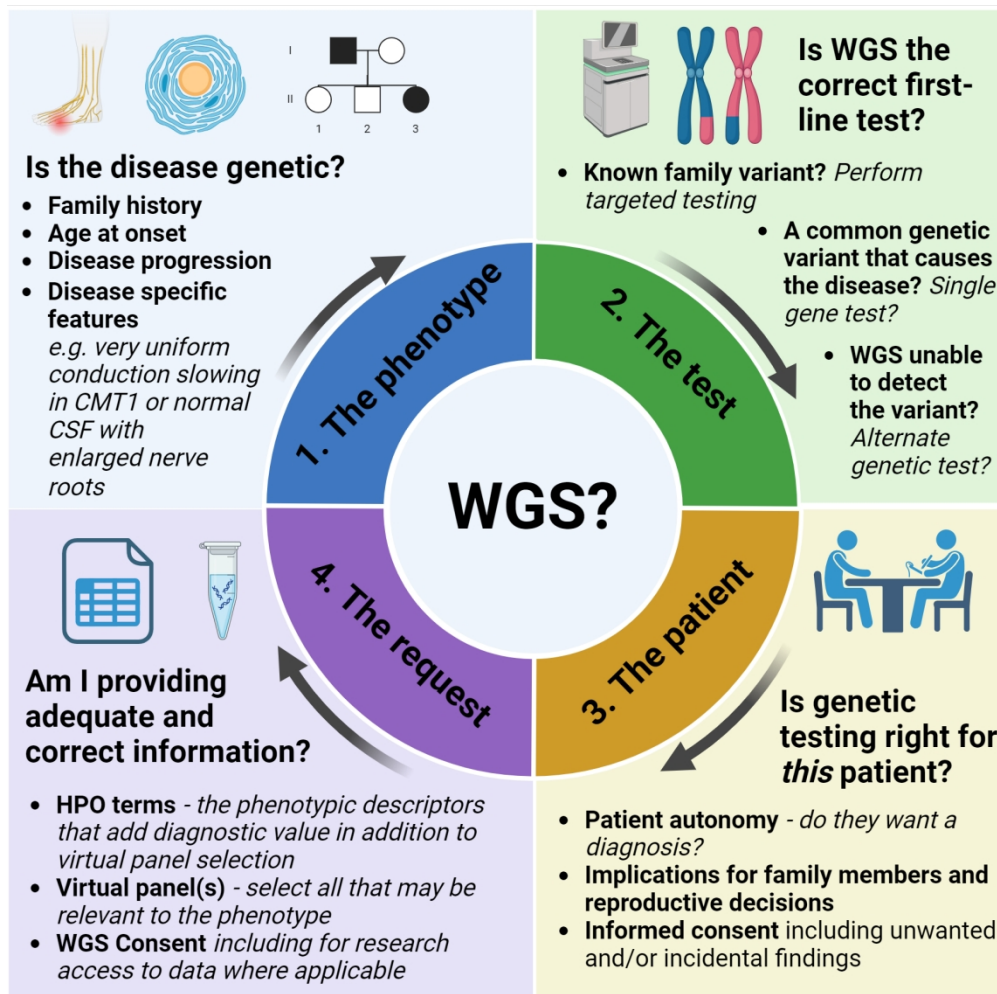# Is Charcot-Marie-Tooth disease a good disease prototype for understanding WGS?

Genetic (or locus) heterogeneity

Q: Are there a lot of genes that cause the disease? A: Yes.

- The current Genomics England R78 Hereditary Neuropathy panel contains 102 'green' genes (those genes established to cause CMT). A further 12 have been denoted green since the last release and await inclusion.
- 95 genes are listed as 'amber' - some evidence that they cause CMT but not meeting the threshold for inclusion and therefore not tested in the panel; PanelApp https://nhsgms-panelapp.genomicsengland.co.uk/panels; 5 February 2024)

Allelic heterogeneity within disease causing genes

Q: Are there multiple disease-causing variants for a given gene? A: Yes.

- For example, our recent study of 387 patients with CMTX1, the second most common form of CMT, identified 109 pathogenic or likely pathogenic variants (Record *et al.*, Brain, 2023).
- ClinVar, a freely accessible public archive of genetic variants, lists 727 submitted variants in *GJB1* of which 192 are deemed pathogenic or likely pathogenic (as of 7 December 2023; https://www.ncbi.nlm.nih.gov/clinvar/).

Molecular heterogeneity

Q: What is the most common type of genetic variant that causes the disease?

- Around 50% of solved CMT is caused by the 1.4Mbp duplication in the short arm of chromosome 17 (a copy number variant, a form of structural variant)
- The remaining ~50% is accounted for by:
  - Single nucleotide variants and insertion-deletions (small, typically up to 50 nucleotides) which are very reliably detected by WGS.
  - Smaller structural variants, mainly copy number variants, and some repeat expansions. Bioinformatic technology is becoming increasingly reliable at detecting these from WGS.

A Diagnostic gap

Q: Is there a reasonable proportion of unsolved cases? A: Yes.

- In a cohort of 1515 patients from our specialist CMT centre, we have a diagnostic rate of 76.9% (Record *et al.*, Brain *in press*) leaving 23.1% unsolved.

Penetrant disease

Q: Do carriers of pathogenic variants manifest the disease? A: Yes.

- Our experience is that CMT is usually fully penetrant (excluding female carriers in X-linked forms of the disease, and patients with low heteroplasmy mitochondrial DNA variants).
- One of the major challenges with WGS is deciphering the many 'variant(s) of uncertain significance'.
- Where a disease has incomplete penetrance, variant interpretation is complicated because an unaffected carrier of a variant doesn't automatically imply the variant is benign. Neurophysiology is an essential tool for phenotyping in CMT and provides objective evidence of affected status.

Figure 2 Understanding the utility of WGS. CMT Charcot-Marie-Tooth disease
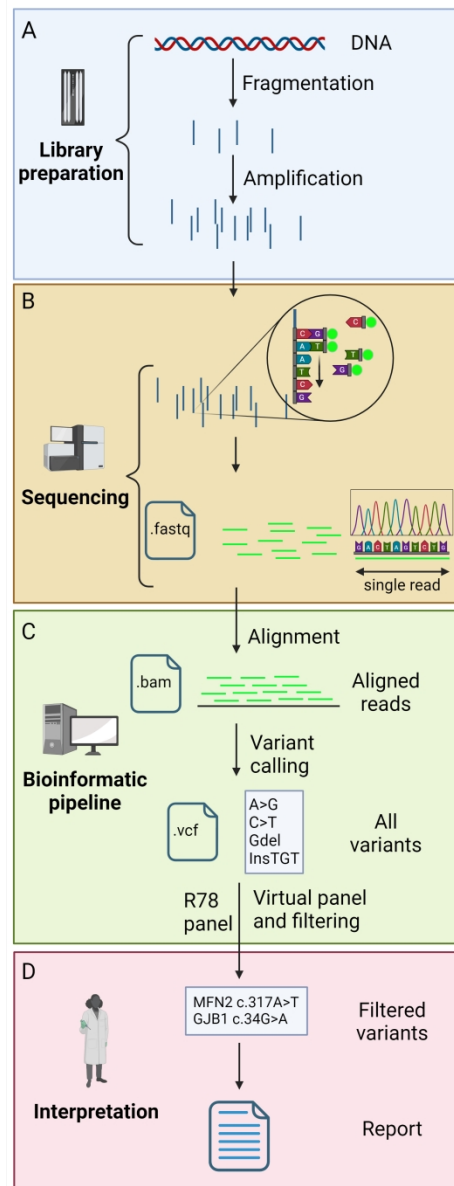
208x242mm (300 x 300 DPI)

Figure 3 WGS workflow DNA library is prepared through fragmentation of the extracted DNA, followed by amplification on the flow cell. Sequencing then occurs through 'sequencing by synthesis' and the subsequent read data is aggregated in a .fastq file. The data is then processed through a bioinformatic pipeline; initially aligned to the reference genome (resulting in a .bam file) and then variants are identified ('variant calling') yielding a .vcf file; the common output file of WGS. Virtual panels can then be applied, and the resultant variants interpreted by clinical scientists, before a report is issued.
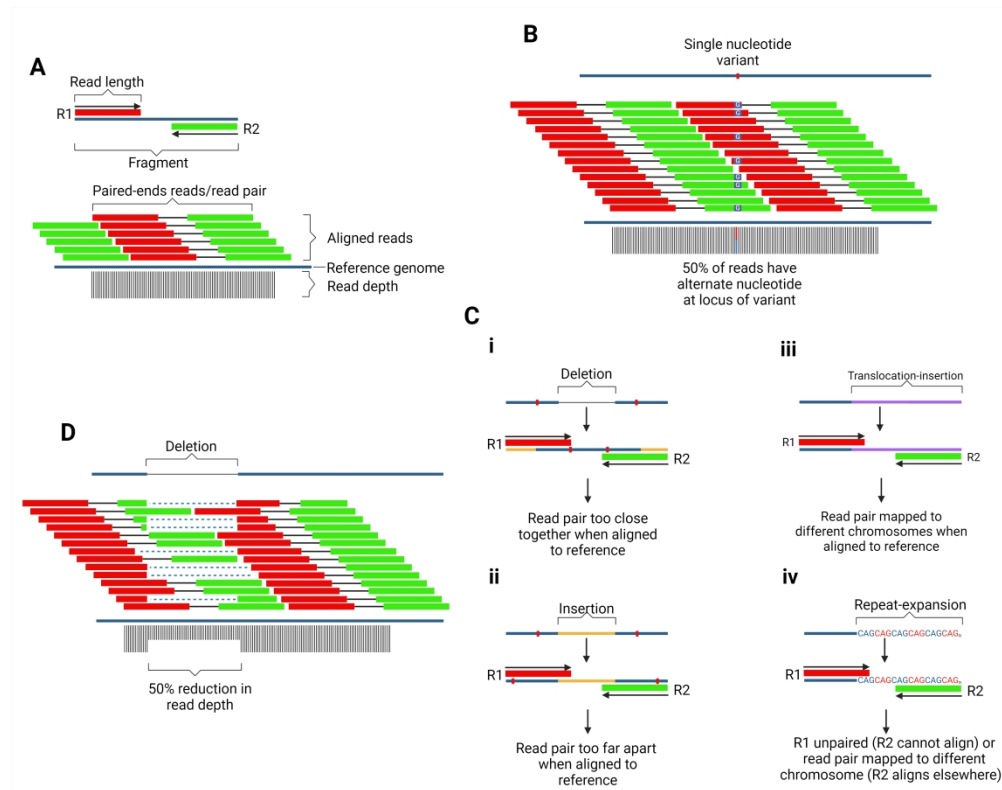
145x372mm (300 x 300 DPI)

Figure 4 Bioinformatic methods of detecting variants A. Schematic of the normal process of paired-end reads of a single fragment, subsequent read alignment to the reference genome and the resultant read depth B. Visual representation of detection of a single nucleotide variant (SNV). One allele of the individual's DNA contains a SNV. When the aligned reads are analysed, 50% have a nucleotide that is different from the reference, and the variant is 'called'. C. structural variant detecting methods: 'Split-read' and 'read-pair' .i) A deletion in the fragment of DNA means that when the read-pairs are aligned to the reference they will appear too close together (red markers are closer than without the deletion). ii) An insertion means the aligned reads appear too far apart when aligned to the reference (red markers further apart than without the insertion). iii) A translocation results in one half of the read-pair aligning to a different part of the genome iv) A repeat expansion, particularly one longer than the read-pair, will result in one of the pairs sequencing only the repeat region. When alignment is attempted, it may not be able to align anywhere (the other read will be 'unmatched') or may align elsewhere in the genome. D. structural variant detecting method: 'Depth of coverage' – aligned reads of a deleted region identify a length of sequencing with a 50% drop in coverage suggesting a heterozygous deletion i.e. one normal allele and one allele containing a deletion.
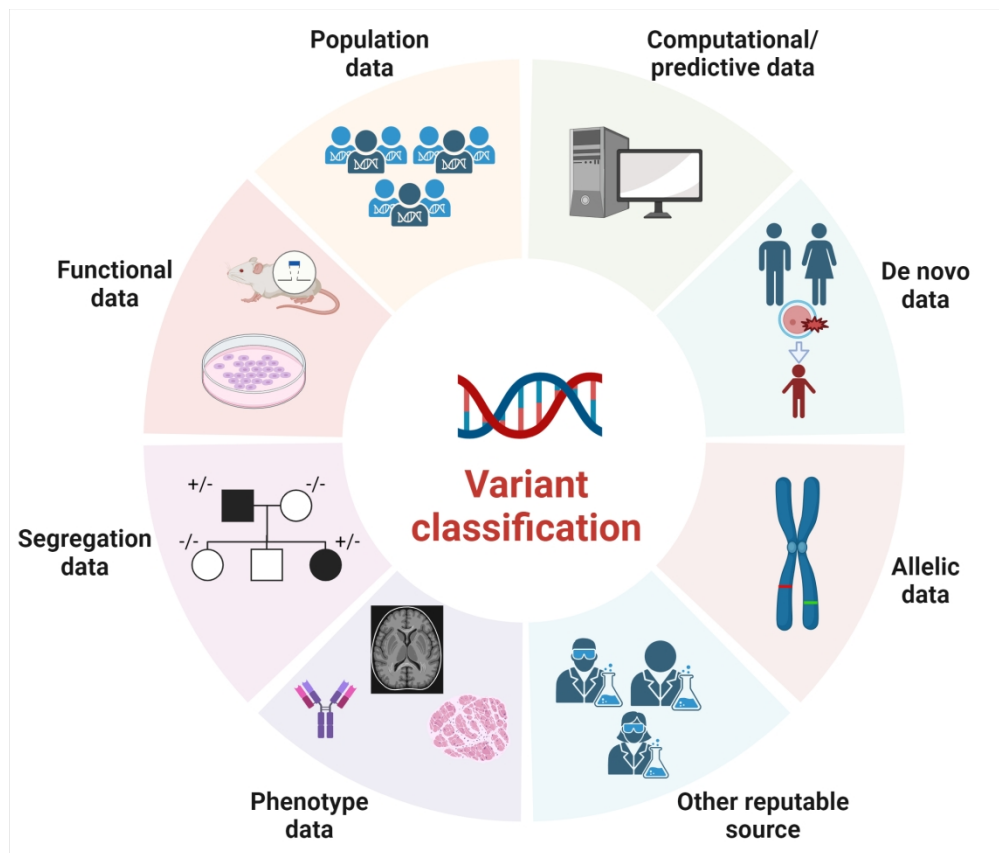
374x294mm (300 x 300 DPI)

Figure 5 Schematic based on the the American College of Medical Genetics and Genomics (ACMG) criteria. Every variant under review has evidence for pathogenicity scrutinised under each of the listed categories.
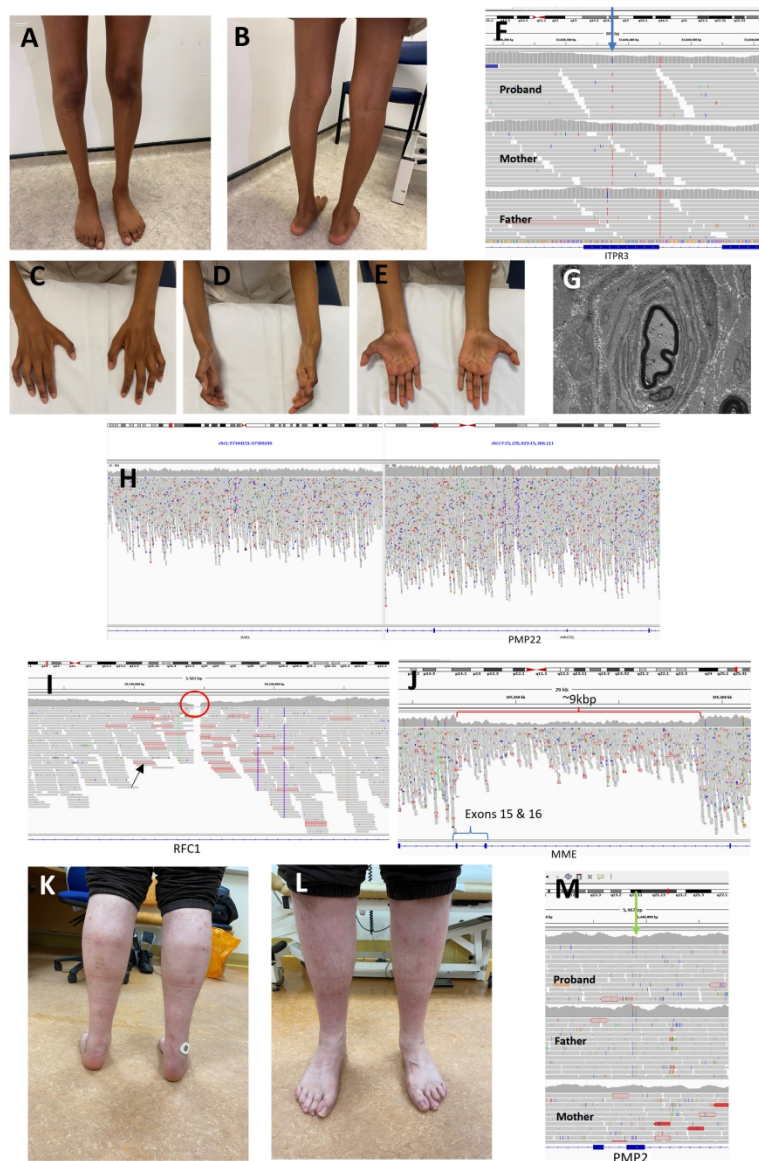
183x155mm (300 x 300 DPI)

Figure 6 Clinical cases A-E. Distal limb muscle atrophy in proband of Case 1 F. Integrative Genomics Viewer (IGV) showing heterozygous variant in ITPR3 (blue arrow) in proband and mother G. Electron microscopy of sural nerve biopsy of Case 1 (proband) showing reduced myelin thickness and onion bulb formation H. IGV showing 1.5x read depth of aligned reads in PMP22 compared with a region on chromosome 1 I. IGV shows loss of coverage in intron 2 of RFC1 (red circle) indicating a biallelic repeat expansion in that region. Reads highlighted in red (black arrow) are unmatched pairs J. IGV shows ~9kb 50% drop in coverage (read depth) in MME encompassing exons 15 and 16; corresponding to a heterozygous deletion K-L. Mild distal lower limb atrophy in proband of Case 5 M. IGV shows heterozygous variant in PMP2 carried by proband and affected father

190x281mm (300 x 300 DPI)

Lessons and pitfalls of whole genome sequencing

Christopher J Record, Mary M Reilly

Centre for Neuromuscular Diseases, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK

Lead corresponding author: Professor Mary M Reilly, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK. m.reilly@ucl.ac.uk

Secondary corresponding author: Christopher J Record, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, WC1N 3BG, UK. Chris.record@ucl.ac.uk

4545/4400

ABSTRACT

Whole genome sequencing (WGS) has recently become the first-line genetic investigation for many suspected genetic neurological disorders. Whilst its diagnostic capabilities are innumerable, as with any test, it has its limitations. Clinicians should be aware of where WGS is extremely reliable (detecting single nucleotide variants), where its reliability is much improved (detecting copy number variants and small repeat expansions) and where it may miss/misinterpret a variant (large repeat expansions, balanced structural variants or low heteroplasmy mitochondrial DNA variants). Bioinformatic technology and virtual gene panels are constantly evolving, and it is important to know what genes and what types of variant are being tested; the current NHS Genomic Medicine Service WGS offers more than early iterations of the 100,000 Genomes Project analysis. Close communication between clinician and laboratory, ideally through a multidisciplinary team meeting, is encouraged where there is diagnostic uncertainty.

INTRODUCTION

Whole genome sequencing (WGS) and the use of genomic testing in neurology, including consent, indications and results, have recently been expertly reviewed in *Practical Neurology*.(1,2) The success of the Genomics England 100,000 Genomes Project (100KGP), sequencing patients with cancer and rare diseases, has led to the introduction of WGS with virtual panels into routine clinical practice for many neurological diseases via the UK National Health Service Genomic Medicine Service (NHS-GMS, https://www.england.nhs.uk/genomics/nhs-genomic-med-service/). The theoretical benefits of WGS are clear; sequencing the entire genome (many orders of magnitude more DNA than previous routine testing, at comparable costs) wherein the molecular diagnosis should lie, provided the clinical diagnosis of a genetic disorder is correct. However, as with every new technology, WGS has its limitations. This article aims to outline the diagnostic utility of WGS, but also to note where caution is needed. The decision to request WGS is a critical step in any patient's diagnostic journey. Where appropriate, especially in sporadic cases, acquired diseases should be excluded first. Figure 1 highlights key points to consider before requesting WGS.

IS CHARCOT-MARIE-TOOTH DISEASE A GOOD DISEASE PROTOTYPE FOR UNDERSTANDING WGS?

Charcot-Marie-Tooth (CMT) disease is an umbrella term for inherited neuropathies but is a clinically and genetically heterogenous group of diseases. The clinical sub-types of CMT include demyelinating sensory and motor neuropathy (CMT1), axonal sensory motor neuropathy (CMT2), sensory and motor neuropathy with intermediate conduction slowing (upper limb motor conduction velocity between 25 and 45 m/s, CMTi), hereditary sensory neuropathy (HSN), and hereditary motor neuropathy (HMN).(3,4)

The diagnostic utility of WGS for an individual lies in its ability to detect vast numbers, and in theory different types, of genetic variant. Figure 2 illustrates the features of a disease group that make it suitable for considering WGS testing.

One downside of CMT as a disease prototype is that functional validation of novel variants/genes is challenging but this underpins how important WGS is in CMT clinical practice. Gold standard functional evidence would be ex-vivo human diseased tissue demonstrating absent, deficient or dysfunctional protein contributing to pathology. This is possible in theory with peripheral nerve biopsies, but this is an invasive procedure requiring technical expertise. Alternatively, RNA sequencing can be used to demonstrate aberrant transcripts in appropriate tissues; Schwann cells

are clearly easier to study than dorsal root ganglia or anterior horn cells. Overall, we feel CMT is an excellent disease to demonstrate the lessons and pitfalls of WGS and will explore these herein.

WHOLE GENOME SEQUENCING TECHNOLOGIES

A basic understanding of the molecular techniques involved in WGS is important to appreciate its potential pitfalls. First, WGS when used in common medical parlance, refers to 'short-read' WGS (srWGS). Table 1 highlights some useful terminology. There are other forms of genomic sequencing, and although used currently mostly in the research setting, their use is increasing in diagnostic genetic laboratories worldwide. Long-read WGS (lrWGS), as suggested in the description, continuously sequences long molecules of DNA, typically tens of kbp in length, but up to many hundreds of kbp depending on the sequencing technology used. The major benefit of lrWGS is the ability to detect and size repeat expansions accurately, and to detect complex, balanced structural variants. The drawbacks include the cost and longer sequencing time, and its error rate on an individual nucleotide level which, when combined with low read depth, affects its ability reliably to detect single nucleotide variants (SNVs) or insertion-deletions (indels).(5) Optical genome mapping is another form of genomic interrogation, and more appropriately termed 'genome imaging'. Its uses have been compared to those previously investigated with karyotyping (i.e. large structural variants) but with the benefit of up to 20,000-fold higher resolution. DNA molecules are enzymatically labelled, and the resultant ligated DNA then 'imaged' for its pattern of periodically spaced fluorescent signals. Its ability to detected large structural variants (0.5 – 1 Mbp) is superior to srWGS and lrWGS, and it is less costly to get higher coverage. As with srWGS and lrWGS it cannot detect aneuploidy (an abnormal number of chromosomes), although this is less relevant in the setting of non-developmental disorders. Another potential drawback of optical genome mapping is the requirement for DNA extraction from a fresh blood sample.(6) As neither lrWGS nor optical genome mapping are used in standard NHS testing, from this point forward we will not discuss them further, and we will refer to srWGS simply as WGS.

TABLE 1 USEFUL TERMINOLOGY

Alignment	Process by which the sequenced reads of the individual are matched to the appropriate region in the reference genome
Bioinformatic pipeline	Series of computational steps processing raw WGS data resulting in variants for analysis; includes alignment, filtering, variant calling
Copy number variant	A type of structural variant, where there is an increase or decrease in dosage across a region of the genome e.g. deletion or duplication

Insertion-deletion (indel)	A small variant, usually <50bp, where a series of consecutive nucleotides are altered/inserted/deleted
Long-read	A single strand of DNA is sequenced continuously, depending on the technology, up to hundreds of kbp in length
Paired-end reads (read-pair)	A fragment of DNA is sequenced from both ends to create paired-end reads, or a read pair. Information can be gleaned when the pairs are aligned to the reference if they are unexpectedly too close, too far, on separate chromosomes or if the one half of the pair is 'unmatched'
Read	The basic molecular output of next generation sequencing; a read is a single consecutively sequenced strand of DNA, before alignment to the reference genome
Read depth	The number of cumulative reads aligned at a particular genomic locus i.e. how many times a particular nucleotide been sequenced
Read length	The number of nucleotides in a single read
Repeat expansion	Repeating nucleotide motif (e.g. CAG_n – the common polyglutamate expansion, or $AAGGG_n$, the most common configuration in *RFC1* CANVAS) where n is the number of repeats. The number of repeats that is considered pathogenic varies widely between diseases.
Short-read	A single strand of DNA is consecutively sequenced between 75-300 bp in length
Single nucleotide variant (SNV)	The alteration/insertion/deletion of a single nucleotide
Structural variant	Medium to large (typically 100s to Mbp in length) variants including duplications, deletions, insertions, balanced translocations and more complex rearrangements
Variant calling	The process by which alterations (variants) in the individual's genome are identified by comparing to the reference genome

NEXT-GENERATION SEQUENCING TECHNOLOGY

WGS uses next-generation sequencing (NGS) technology, also known as high-throughput or massively-parallel sequencing. NGS has been used for many years in clinical diagnostic laboratories for the sequencing of disease-specific gene panels and whole exome sequencing. There are several sequencing platforms,(7) but the dominant provider worldwide is Illumina, which is also used by NHS-GMS, and the process described hereafter. A flow diagram of the process involved is shown in Figure 3.(8)

The first step is **library preparation** (Figure 3A); the genomic DNA library is a series of short fragments ready for sequencing. The DNA (typically extracted from leukocytes in blood; purple EDTA tube) is fragmented and then each fragment amplified. Fragments are then **sequenced** in a process called 'sequencing by synthesis', whereby fluorescently tagged nucleotides are added to a linear single strand of DNA complementary to the fragment; the resultant fluorescent DNA strand is known

as a 'read' and can be sequenced by its characteristic spectral emission (one wavelength for each of the four nucleotides, Figure 3B). The fragment is sequenced from both ends forming 'paired-end reads', allowing additional information to be gleaned when the reads are aligned. Data are then fed into the **bioinformatic pipeline** (Figure 3C). The millions of reads are **aligned** to the reference genome, which when visually represented, form piles of overlapping reads. The overall coverage of the WGS describes what proportion of the reference genome is sequenced to a satisfactory read depth. Figure 4A shows in detail how an unmutated fragment is sequenced and aligned to the reference.

Variant calling is the process of identifying variants i.e. variation in an individual's genome when compared with the reference. The basic output of a WGS bioinformatic pipeline is the identification of small variants; alteration/insertion/deletion of single nucleotides (SNVs, Figure 4B) or a small number of consecutive nucleotides (indels). The universal final output for the millions of variants generated is a .vcf file. Other types of genetic variant can also be detected including structural variants (both copy number variants, and balanced rearrangements; the latter where there is no change in dose at a particular locus), repeat expansions and mitochondrial DNA (mtDNA) variants, but their detection and calling is variable (Figure 4 and see *When WGS might not be the correct test*).

After variant calling, the variants are **filtered** according to specified criteria (see *Filtering and prioritisation*). Application of a **virtual panel**(s) may yield possible candidate variants, which are **interpreted** by clinical scientists (Figure 3D). If there is ambiguity or uncertainty, results are ideally discussed at a multidisciplinary team (MDT) meeting, following which a genetic report can be issued.

VIRTUAL PANELS

Although WGS in theory allows analysis of variants from an individual's entire genome, this is neither desirable (incidental unwanted findings) nor practical (a human genome contains approximately five million SNVs) therefore virtual panels are essential to refine the search. In the NHS, clinicians are required to select virtual gene panel(s) when requesting WGS. The NHS-GMS PanelApp (https://nhsgms-panelapp.genomicsengland.co.uk/panels) is a publicly available resource that uses genetic expertise through crowdsourcing to curate disease-specific gene panels.(9) For a gene to be included it needs to be approved as 'green' by a number of verified experts; a green gene is broadly one in which plausible disease-causing variants have been found in three or more unrelated individuals/families. However, the panels can only be as correct and up-to-date as their reviewers and the current available evidence. For example, *SORD* was discovered as a common, and potentially treatable, cause of CMT in 2020,(10) but was not approved as a green gene until November 2022. Panels are periodically updated, and previous iterations can be found on PanelApp. Genes that cause

a complex phenotype which include the disease group of interest e.g. *ABHD12* causing polyneuropathy, hearing loss, ataxia, retinitis pigmentosa, and cataract (PHARC) syndrome, are often not included if the panel specifies an isolated phenotype; it is not a green gene on the current 'Hereditary Neuropathy or pain disorder' panel (R78 version 3.24). Similarly, novel, rare genes may not meet green inclusion criteria. It is therefore important to understand which genes are tested in a specific panel, and if there is a particular gene of interest in a clinical case, this should be discussed with the genetic laboratory. It is currently recommended that broad rather than narrow use of panels is applied to maximise chances of identifying causative variants.

FILTERING AND PRIORITISATION

Refining the vast number of variants detected through WGS requires filtering strategies. The two most powerful tools are the allele frequency of the variant in reference databases (the most commonly used is gnomAD; https://gnomad.broadinstitute.org/, Box 1) and in family studies, the inheritance pattern, as defined by relative disease status.

Population allele frequency

Historically the upper limit for the population allele frequencies was set at < 1 in 100 for autosomal recessive, and < 1 in 1000 for autosomal dominant disease, however we know that for most rare diseases these thresholds are far too high. A useful online calculator for the estimation of a disease-specific population allele count and frequency is found at https://cardiodb.org/allelefrequencyapp/. It is important to remember that if a variant seemingly occurs at too high an allele frequency, it will be filtered by the bioinformatic pipeline, and not considered for interpretation. The most common variant c.757delG in *SORD*-related CMT is present in a highly homologous non-functioning pseudogene *SORD2P* in 95% of controls; the two variants can be challenging to delineate bioinformatically and therefore the *SORD* variant is potentially inappropriately filtered.(10) This problem with this particular variant has been overcome but was a barrier to its discovery.

One must also be wary of regional 'hotspots' for particular variants. The *GNE* variant p.Val696Met (previously p.Val727Met) causing the rare recessive hereditary inclusion body myopathy/Nonaka myopathy is exceedingly common in the South Asian population where the majority of the disease is seen.(11) The overall quoted allele frequency appears too high for the prevalence of the disease in the UK, and may result in the variant being discounted. Only when the regional breakdown is examined, can it be appreciated that the variant is very rare in European populations, in keeping with disease prevalence.

Reference genome

The current human reference genome, denoted GRCh38, originates from the genomes of 20 anonymous volunteers from the USA. It has been shown that two-thirds comprises the genome of a single individual of mixed European and African descent.(12) It is widely recognised that the current reference genome has significant limitations; it contains some gaps (~5%), has regions of unreliable coverage (e.g. around the centromere), and reflects a very narrow ancestry. The Human Pangenome Reference consortium have set out to rectify the flaws in the current reference by creating a new reference built from 350 human genomes, and have recently published a draft from 47 individuals from diverse backgrounds.(13) Until the 'Pangenome' comes into routine clinical practice, clinicians must we aware that patients from certain ethnic backgrounds (e.g. the Indian subcontinent) may have variants missed because the reference does not reflect their ancestry.

Family studies and relative disease status

Variant segregation through family studies (WGS in more than one family member that are subsequently analysed together) enhances diagnostic success.(14) At recruitment, participants are assigned as affected, unaffected or unknown. Downstream in the process, if a dominant variant is detected in the affected proband and a reportedly unaffected parent, it will be disregarded or deprioritised. Therefore, caution is needed when the disease has an adult onset or a variable presentation, so that relatives' disease status is appropriately assigned.

Human phenotype ontology terms

As part of the process of requesting WGS through the NHS-GMS, the clinician is required to include Human Phenotype Ontology (HPO) terms (https://hpo.jax.org/app/, Box 2). These phenotypic descriptors can be used to prioritise variants using Exomiser, a programme utilised by NHS-GMS.(15) For example, a man with a demyelinating neuropathy and upper motor neurone signs underwent WGS in the 100KGP with the Hereditary Neuropathy virtual panel applied. There were no candidate variants from the panel, but because the Human Phenotype Ontology terms included 'demyelinating neuropathy' and 'Babinski sign', a variant in *ABCD1*, known to cause X-linked myeloneuropathy, was identified. Subsequent discussion at our MDT, and further clinical and laboratory assessments, confirmed this to be the causative gene. This gene is not present in the current Hereditary Neuropathy panel.

VARIANT INTERPRETATION AND REPORTING

Every candidate variant is classified according to established criteria. UK laboratories use the American College of Medical Genetics and Genomics (ACMG) and Association for Clinical Genomic Service (ACGS) guidelines.(16,17) Any given variant, with no supporting data, starts as a 'variant of uncertain significance' (VUS). Evidence is combined from different categories (including data on allele frequency, functional studies, segregation and prior literature reports) to upgrade the variant as likely pathogenic or pathogenic, or downgrade to likely benign or benign (Figure 5). As with gene panels, variant interpretation relies upon the available evidence, and its application, and therefore variant classification may differ between laboratories. Ideally, clinicians will have access to an MDT (with clinical scientists) to discuss WGS results of unsolved cases, cases with unexpected pathogenic variants, or those with a very typical phenotype for a particular gene, in which no variants have been reported. There is a criterion within the ACMG/ACGS (PP4) criteria that uses phenotype specificity to upgrade variants e.g. absence of dystrophin in a muscle biopsy in a male patient with muscular dystrophy phenotype, when considering a variant in *DMD*. Without the communication of clinical information from clinician to laboratory, the variant might remain a VUS.

Historically, relevant VUSs were listed as an addendum to genetic reports. However, NHS-GMS have adopted guidance from the ACGS that VUSs should not be reported unless exceptional circumstances apply, after a discussion at an MDT meeting. This change is critical for practising clinicians to be aware of. The rationale is that reporting a VUS may lead to confusion on the part of referring clinician or patient, misinterpretation and potentially misdiagnosis. Even when a VUS is likely to be causative, family screening for the variant would still need careful discussion and counselling, and preimplantation genetic testing or entry into a clinical drug trial would only be considered in exceptional circumstances.

However, we have experience that transparent reporting of VUSs to clinicians with genetic expertise, has been vital in clinching a genetic diagnosis with the passage of time. A 'warm' VUS may be upgraded to pathogenic following, for example, a new publication implicating the gene/variant in disease. Without information about VUSs made available on a genetic report, such cases may remain unsolved.

Another example of the need for careful reporting is the presence of a single pathogenic variant in a recessive gene, a so called 'single hit'. If reported, it should be made clear that the diagnosis is not confirmed, but a single pathogenic variant has been detected. With a suggestive phenotype, a 'single hit' will often trigger a discussion with the laboratory to look on the other allele for deep intronic

variants (that might affect splicing or create pseudoexons), or structural variants (i.e. deletion of a portion of the gene), or explorative analysis of the genome in a research setting.(18)

WHEN WGS MIGHT NOT BE THE CORRECT TEST

The essential first step for genetic testing is ensuring the right test is sent. Jain *et al*. have previously discussed this in detail.(1) In the UK clinicians must consult the NHS Genomic Test Directory https://www.england.nhs.uk/publication/national-genomic-test-directories/. Many neurological diseases, including some that are treatable, have their molecular basis in non-SNV genetic variation. Huntington's disease (CAG trinucleotide repeat expansion in *HTT*), genetic motor neurone disease/frontotemporal dementia (GGGGCC hexanucleotide repeat expansion in *C9ORF72*), spinal muscular atrophy (biallelic deletion of exon 7 +/- 8 in *SMN1*), fragile X syndrome (CCG trinucleotide repeat expansion in *FMR1*) and Duchenne muscular dystrophy, (~60% caused by exon-level deletions in the X-linked *DMD*) are all caused by either repeat expansion or structural variants. More than 50% of CMT is caused by a duplication of *PMP22*, and the remainder, a mixture of genetic variant types.

The limitation of WGS to accurately detect structural variants and repeat expansions lies in the read length. Put simply, it is difficult to quantify a variant with genomic size potentially orders of magnitude larger than the unit of measurement. Figure 4 details the use of paired-end reads in the sequencing and alignment process, and how they can be used to detect non-SNV variants. When the DNA fragment is sequenced from both ends, the two paired-end reads contain markers that identify them as a pair. If, when the reads are aligned to the reference genome, they align too far apart or too close together, this can be bioinformatically detected. Similarly, if a read aligns without a 'mate' (the other part if the pair cannot satisfactorily align to the reference), this can also be flagged. This approach for detecting non-SNV variants is shown in Figure 4Ci and ii and is known as a 'paired-end' (or 'read-pair') approach to detecting structural variants. Similarly, the 'split-read' approach uses information that a single read is disrupted, or split, by a structural variant. The read depth or 'depth of coverage' approach replies upon algorithms detecting regions where there is a significant increase or decrease in coverage (Figure 4D). All these computational approaches have their limitations for different structural variants, and the best algorithms combine more than one approach.(19) Structural variation on a chromosomal level e.g. aneuploidy or ringed chromosomes, will not be detected by WGS and karyotyping should be requested separately.

Repeat expansions, where the number of repeats is critical to the diagnosis, can be challenging to size through WGS; large repeat expansions will be longer than the read, or read-pair (Figure 4Civ). ExpansionHunter is a tool that estimates the repeat size at the loci of known expansions, which

when paired with visual inspection, was sensitive and specific for correctly sizing expansions in the 100KGP when the expansion size was less than the read length.(20)

However, there are three important caveats to the above. First, as with virtual panels, if the gene and specifically the expansion (if that is the diagnostic question) is not on the virtual panel, non-SNVs will not be tested. Second, when the expansion is larger than the read length (as seen in *FMR1*, *C9orf72*, *DMPK* (myotonic dystrophy type 1) and *FXN* (Friedreich's Ataxia), although an expansion could be identified, it was often significantly underestimated by ExpansionHunter (Figure 4Civ). Although *RFC1*, the gene recently identified as causing cerebellar ataxia, neuropathy and vestibular areflexia syndrome (CANVAS) through biallelic pentanucleotide repeat expansions, was not examined by Ibañez *et al.*, the same would apply; the expansion is typically >1000 repeats (>5000 nucleotides).(21) In NHS laboratories *RFC1* is currently tested using non-WGS methods. Third, early iterations of the 100KGP pipeline did not routinely analyse for any non-SNVs, and many were missed and not reported.

MtDNA sequencing is currently requested as a separate test to sequencing of the nuclear genome. Studies have shown that with a satisfactory read depth, WGS can detect mtDNA variants at a heteroplasmy level down to 10%,(22) but if there is a significant suspicion for mitochondrial disease, mtDNA sequencing should be requested separately. Other types of genetic mechanism including epigenetic factors such as DNA methylation or imprinting will not be detected using WGS and should have separate testing requested. Lastly, in the NHS, if a rapid result is critical to guide management, the R14 'Acutely unwell children with a likely monogenic disorder' WGS can be requested for critically ill children and adults, with a turnaround time of 2–3 weeks.(23)

EXAMPLES FROM THE CLINIC

Key to our diagnostic success in the 100KGP was access to the data in the research environment, and regular review of cases at a dedicated clinical-research-genetic MDT. We illustrate with clinical cases practical examples of potential pitfalls discussed above.

Case 1

A woman in her late teens presented with a subacute history of sensory changes in her hands, a few weeks following a viral illness. She developed progressive weakness and wasting of intrinsic hand muscles. At initial assessment she also had mild sub-clinical distal lower limb weakness (Figure 6A-E). There was no family history of neuromuscular illness and parents were non-consanguineous. Initial neurophysiology showed a patchy, widespread, conduction slowing neuropathy. She was treated in her local hospital with intravenous immunoglobulin for presumed chronic inflammatory

demyelinating neuropathy. Subsequent CSF examination showed normal constituents, nerve roots were markedly thickened and pathologically enhancing on MRI, and nerve biopsy showed a chronic demyelinating neuropathy without inflammation (Figure 6G). She progressed slowly despite treatment; initial genetic testing, including CMT1A with multiplex ligation-dependent probe amplification (MLPA), and a 14 gene panel in 2015, was negative. She was enrolled into the 100KGP with her parents, with no primary findings. Through a research collaboration we identified the variant c.4271C>T p.(Thr1424Met) in *ITPR3,* a gene only reported in three families and not included in the virtual panel.(24,25) Additionally, the variant was maternally inherited (Figure 6F). Clinically the mother had no symptoms and a completely normal neurological examination, but neurophysiology showed a clear conduction slowing neuropathy. The diagnosis is CMT, with remarkable variability in severity, due to an *ITPR3* variant. This case highlighted the importance of the assigned affected status; segregation was confirmed but only by neurophysiology. Similarly, research access to the 100KGP data was essential to identify a gene not on the virtual panel but in the literature.

Case 2

A man in his late 60s was referred for a diagnostic opinion. He had a progressive sensory and motor neuropathy since his 20s. Neurophysiology was clearly demyelinating with a median nerve motor conduction velocity of 22 m/s. The family history was of autosomal dominant disease. His 100KGP primary findings report was negative. We examined the aligned sequence data and discovered 1.5x the read depth in the region of *PMP22* compared with the rest of the genome (Figure 6H). MLPA confirmed the 17p.22 duplication; the diagnosis was CMT1A. The bioinformatic pipeline did not call this common copy number variant seen in CMT. We have now seen three cases of CMT1A referred for a diagnostic opinion where the chromosome 17 duplication was either missed or not looked for as clinicians were not aware that next-generation sequencing gene panels and WGS in the 100KGP did not reliably detect the duplication.(26) Despite the panel name 'Hereditary Neuropathy NOT PMP22 copy number', the current WGS panel does now include the *PMP22* duplication, but the first line test in conduction slowing neuropathies should still be 'R77 Hereditary Neuropathy – PMP22 copy number' (MLPA).

Case 3

A man his early 50s had a 4-year history of progressive unsteadiness, particularly in the dark, and reduced sensation in his distal limbs. He had a longstanding cough. Examination identified a sensory ataxia and large and small fibre sensory loss, without weakness. Neurophysiology showed a severe pure sensory axonal neur(on)opathy. Extensive investigations including antibody testing, neuroaxis

imaging, positron emission spectroscopy scan, nerve and lip biopsy excluded inflammatory, nutritional, and malignant causes. A 56-gene CMT panel, *FXN* and *POLG* sequencing and 100KGP testing was negative. We examined the aligned WGS sequence data of *RFC1* in the research environment and found a complete drop of read depth within intron 2 (Figure 6I). Subsequent repeat-primed polymerase chain reaction confirmed biallelic AAGGG repeat expansions in *RFC1*, and a diagnosis of CANVAS. This case highlights a missed large intronic repeat expansion, still not reliably called on WGS. Currently *RFC1* testing must be requested separately.

Case 4

A man in his early forties presented with a 10-year history of progressive walking difficulties due to distal lower limb weakness. There was no family history. Examination showed a length-dependent motor neuropathy; this was confirmed on neurophysiology and there was no slowing or conduction block. Lead and hexosaminidase A levels were normal. Testing for *AR* expansion, 32-gene CMT2/distal HMN panel and 100KGP were negative. Review in the research environment identified a heterozygous variant in *MME* (c.202C>T p.(Arg68Ter)), a gene known to cause adult onset recessive, motor predominant CMT.(27) The single variant is classed as pathogenic when in *trans* with a second pathogenic variant; this was a single hit in a recessive disease. We then examined the aligned sequence data and identified a 9kbp drop in read depth in *MME*, consistent with a deletion including exons 15 and 16, predicted to be pathogenic (Figure 6J). Both variants were confirmed in the diagnostic laboratory. The diagnosis was distal HMN due to compound heterozygous variants in *MME*; one that was missed because a single recessive variant was not reported, and the structural variant was not identified by the analysis pipelines.

Case 5

A man in his late teens was assessed as he transitioned to the adult neuropathy clinic. He had a normal birth but began walking with in-turning feet aged four. His feet then began to slap as he developed slowly progressive weakness. His father had mild symptoms compatible with CMT. Examination of the proband revealed relatively mild, length-dependent motor deficits (Figure 6K-L). His neurophysiology showed a sensory and motor demyelinating neuropathy; a clinical diagnosis of CMT1 was made. A 56-gene CMT panel was negative and the 100KGP project had no primary findings. Review of genes not included in the virtual panel used by the 100KGP in the research environment revealed a paternally inherited, previously reported pathogenic variant in the myelin protein gene *PMP2*, confirming the genetic diagnosis (Figure 6M).(28) Despite *PMP2* being established as a cause of CMT in 2016, the gene was not included in the 100KGP panel.(29)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

CONCLUSIONS

The diagnostic opportunities through WGS are clear and are reflected in the introduction of WGS into routine NHS diagnostic testing. However, caution must be taken when reading a 'negative' report. WGS has its technical limitations; it very reliably detects SNVs and small indels, and although bioinformatic algorithms are now confidently detecting copy number variants, this wasn't always the case, and detecting balanced structural variants and sizing large repeat expansions remains unreliable. Variants are prioritised according to the information provided by the requesting clinician; a detailed phenotypic description and, if applicable, broad use of virtual panels, increases the chances of a correct genetic diagnosis. Family studies increase the diagnostic yield but rely upon correct assignment of disease status of relatives. If a negative report is received but there is high diagnostic suspicion, we encourage discussion with the genetic laboratory and/or an MDT meeting to consider further focused analysis. Provided the diagnosis of a genetic disorder is correct (excluding mtDNA disorders), although the answer should in theory like within 'whole' genome sequencing, WGS is not always the correct test to request. Lastly, all the cases in this review were diagnosed through research access to 100KGP data; there will always be unsolved and novel causes for neurological disease and the authors feel strongly that clinical genomic researchers should, where their patient has consented, have access to their data to ensure we continue to increase genetic diagnoses for individuals and their families, and advance the field as a whole. Access to research data is not universal, and if after discussion with the local genetic laboratory there is no diagnosis, clinicians should consider referring to a specialist centre.

KEY POINTS

- WGS is the first line test for many, but not all, suspected genetic neurological disorders. Before requesting WGS, clinicians should first ensure relevant initial single genetic tests are negative (e.g. *PMP22* duplication in CMT).
- Gene panels are constantly evolving, and it is important to check which genes and/or type of genetic variant is offered, particularly if there is a specific genetic diagnosis in mind.
- Accurate phenotype information, via Human Phenotype Ontology terms, and correct assignment of relative affected status, are critical to maximise diagnostic yield. Relative testing is desirable, and sometimes essential.
- Discussion, ideally in an MDT setting, with the genetics laboratory is recommended for selected unsolved cases and where there are unexpected or uncertain results. Where variants of uncertain significance remain unreported, communication of specific phenotype data may be the key to their reclassification to pathogenic.

FURTHER READING

- 100,000 Genomes Project Pilot Investigators; Smedley D, Smith KR, *et al*. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care – Preliminary Report. N Engl J Med. 2021 Nov 11;385(20):1868-1880.

- Ibañez K, Polke J, Hagelstrom RT, *et al*. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. Lancet Neurol. 2022 Mar;21(3):234-245.

- Moore AR, Yu J, Pei Y, Cheng EWY, *et al*. Use of genome sequencing to hunt for cryptic second-hit variants: analysis of 31 cases recruited to the 100 000 Genomes Project. J Med Genet. 2023 Nov 27;60(12):1235-1244.

- Pipis M, Rossor AM, Laura M, Reilly MM. Next-generation sequencing in Charcot-Marie-Tooth disease: opportunities and challenges. Nat Rev Neurol. 2019 Nov;15(11):644-656.

Box 1 GnomAD

The Genome Aggregation Database (gnomAD, pronounced nō̩mad) is the most widely used population database of genomic variation. Launched in 2014 as the Exome Aggregation Consortium (ExAC), it is now in its fourth iteration (gnomAD v4, released in November 2023). The open access online database contains genomic data from around 730,000 exomes and 76,000 genomes (up to 1.6 million alleles), derived from more than 100 studies in more than 25 countries. The major output is variant frequency data i.e. how many times has a particular variant been observed in this dataset – 'the population'? The genomic data is broadly derived from a mixture of case-control studies, and large biobanks, including more than 400,000 individuals from the UK Biobank; this is not a healthy control database and will contain affected individuals, with a frequency probably no higher than the disease prevalence.

Box 2 Human Phenotype Ontology (HPO) terms

The concept of HPO is straightforward; to standardise the description of a clinical phenotype. HPO terms can include symptoms, examination findings, syndromes, investigation results, disease severity and onset. The NHS-GMS WGS request form requires inputting of at least one, but ideally several, HPO terms for the patient in question. This can be time consuming and seem unnecessary, but detailed clinical information maximises the chances of WGS finding an answer for the patient. Consider the scenario of a patient deemed by the neurologist to have a unique phenotype of ophthalmoplegia (HP:0000602), gastrointestinal dysmotility (HP:0002579), and demyelinating peripheral neuropathy (HP:0007108). These terms inputted together might be very specific for a

particular gene (e.g. mitochondrial), and any variant found prioritised for analysis (even if not on the requested panel), and its classification potentially upgraded based on the information provided. Importantly, the term peripheral neuropathy (HP:0009830) provides no meaningful extra information if requesting the Hereditary Neuropathy panel. The absence of a clinical feature can also be recorded and may be relevant e.g. the absence of tremor in a syndrome of Parkinsonism. The clinical assessment by the neurologist can be most powerful tool for refining genetic variants and detailed and specific HPO terms are a way of quantifying this expertise.

COMPETING INTERESTS

None to declare.

ACKNOWLEDGEMENTS

CONTRIBUTORSHIP

CJR analysed the data and wrote the manuscript. MMR conceptualised the study and provided senior critical review and revisions.

FUNDING

None to declare.

ETHICS STATEMENT

All patients are recruited to our ethically approved research study 'Charcot-Marie-Tooth Disease and related disorders: A Natural History Study', reviewed by the London Queen Square Research Ethics Committee (REC No.: 09/H0716/61). Separate individual patient consent was obtained for use of photographs.

DATA AVAILIBILITY

The data that support the findings of this study are available from the corresponding author, upon reasonable request. The data are not publicly available since they contain information that could compromise the privacy of research participants.

REFERENCES

1.   Jain V, Irving R, Williams A. Genomic testing in neurology. Pract Neurol. 2023 Oct 19;23(5):420–9.

2.   Morris HR, Houlden H, Polke J. Whole-genome sequencing. Pract Neurol. 2021;21(4):322–6.

3.   Laurá M, Pipis M, Rossor AM, Reilly MM. Charcot-Marie-Tooth disease and related disorders: An evolving landscape. Vol. 32, Current Opinion in Neurology. Lippincott Williams and Wilkins; 2019. p. 641–50.

4.   Pipis M, Rossor AM, Laura M, Reilly MM. Next-generation sequencing in Charcot–Marie–Tooth disease: opportunities and challenges. Nat Rev Neurol. 2019 Nov 1;15(11):644–56.

5.   Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020 Dec 7;21(1):30.

6.   Smith AC, Neveling K, Kanagal-Shamanna R. Optical genome mapping for structural variation analysis in hematologic malignancies. Am J Hematol. 2022;97(7):975–82.

7.   Pervez MT, Hasnain MJU, Abbas SH, Moustafa MF, Aslam N, Shah SSM. A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. Biomed Res Int. 2022;2022.

8.   Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020;583(7814):96–102.

9.   Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019;51(11):1560–5.

10.   Cortese A, Zhu Y, Rebelo AP, Negri S, Courel S, Abreu L, et al. Biallelic mutations in SORD cause a common and potentially treatable hereditary neuropathy with implications for diabetes. Nat Genet. 2020;52(5):473–81.

11.   Bhattacharya S, Khadilkar S V., Nalini A, Ganapathy A, Mannan AU, Majumder PP, et al. Mutation Spectrum of GNE Myopathy in the Indian Sub-Continent. J Neuromuscul Dis. 2018 Feb 21;5(1):85–92.

12.   Sherman RM, Salzberg SL. Pan-genomics in the human genome era. Nat Rev Genet. 2020 Apr 7;21(4):243–54.

13. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617(7960):312–24.

14. Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, Cipriani V, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. New England Journal of Medicine. 2021;385(20):1868–80.

15. Robinson PN, Köhler S, Oellrich A, Genetics SM, Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014;24(2):340–8.

16. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. 2015 May 8;17(5):405–24.

17. Ellard S, Baple EL, Callaway A, Berry I, Forrester N, Turnbull C, et al. ACGS best practice guidelines for variant classification in rare disease 2020. Association for Clinical Genomic Science. 2020;

18. Moore AR, Yu J, Pei Y, Cheng EWY, Taylor Tavares AL, Walker WT, et al. Use of genome sequencing to hunt for cryptic second-hit variants: analysis of 31 cases recruited to the 100 000 Genomes Project. J Med Genet. 2023;1235–44.

19. Pagnamenta AT, Camps C, Giacopuzzi E, Taylor JM, Hashim M, Calpena E, et al. Structural and non-coding variants increase the diagnostic yield of clinical whole genome sequencing for rare diseases. Genome Med. 2023;15(1):1–25.

20. Ibañez K, Polke J, Hagelstrom RT, Dolzhenko E, Pasko D, Thomas ERA, et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. Lancet Neurol. 2022;21(3):234–45.

21. Cortese A, Simone R, Sullivan R, Vandrovcova J, Tariq H, Yan YW, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nat Genet. 2019;51(4):649–58.

22. Poole O V., Pizzamiglio C, Murphy D, Falabella M, Macken WL, Bugiardini E, et al. Mitochondrial DNA Analysis from Exome Sequencing Data Improves Diagnostic Yield in Neurological Diseases. Ann Neurol. 2021;89(6):1240–7.

23. Williamson SL, Rasanayagam CN, Glover KJ, Baptista J, Naik S, Satodia P, et al. Rapid exome sequencing : revolutionises the management of acutely unwell neonates. European Journal of Pediatricscs. 2021;180:3587–91.

24. Beijer D, Dohrn MF, Rebelo AP, Feely SME, Reilly MM, Scherer SS, et al. PNS 2022 Abstract Supplement. Journal of the Peripheral Nervous System. 2022;27(S3):352–63.

25. Schabhüttl M, Wieland T, Senderek J, Baets J, Timmerman V, De Jonghe P, et al. Whole-exome sequencing in patients with inherited neuropathies: Outcome and challenges. J Neurol. 2014;261(5):970–82.

26. Record CJ, Pipis M, Poh R, Polke JM, Reilly MM. Beware next-generation sequencing gene panels as the first-line genetic test in Charcot-Marie-Tooth disease. J Neurol Neurosurg Psychiatry. 2023 Apr 14;94(4):327–8.

27. Higuchi Y, Hashiguchi A, Yuan J, Yoshimura A, Mitsui J, Ishiura H, et al. Mutations in MME cause an autosomal-recessive Charcot-Marie-Tooth disease type 2. Ann Neurol. 2016;79(4):659–72.

28. Motley WW, Palaima P, Yum SW, Gonzalez MA, Tao F, Wanschitz J V., et al. De novo PMP2 mutations in families with type 1 Charcot-Marie-Tooth disease. Brain. 2016;139(6):1649–56.

29. Hong Y Bin, Joo J, Hyun YS, Kwak G, Choi YR, Yeo HK, et al. A Mutation in PMP2 Causes Dominant Demyelinating Charcot-Marie-Tooth Neuropathy. PLoS Genet. 2016;12(2):1–15.

FIGURE LEGEND

Figure 1 Should I send WGS? The stepwise process required for requesting WGS

Figure 2 Understanding the utility of WGS. CMT Charcot-Marie-Tooth disease

Figure 3 WGS workflow DNA library is prepared through fragmentation of the extracted DNA, followed by amplification on the flow cell. Sequencing then occurs through 'sequencing by synthesis' and the subsequent read data is aggregated in a .fastq file. The data is then processed through a bioinformatic pipeline; initially aligned to the reference genome (resulting in a .bam file) and then variants are identified ('variant calling') yielding a .vcf file; the common output file of WGS. Virtual panels can then be applied, and the resultant variants interpreted by clinical scientists, before a report is issued.

Figure 4 Bioinformatic methods of detecting variants A. Schematic of the normal process of paired-end reads of a single fragment, subsequent read alignment to the reference genome and the resultant read depth B. Visual representation of detection of a single nucleotide variant (SNV). One allele of the individual's DNA contains a SNV. When the aligned reads are analysed, 50% have a nucleotide that is different from the reference, and the variant is 'called'. C. structural variant detecting methods: 'Split-read' and 'read-pair' .i) A deletion in the fragment of DNA means that when the read-pairs are aligned to the reference they will appear too close together (red markers are closer than without the deletion). ii) An insertion means the aligned reads appear too far apart when aligned to the reference (red markers further apart than without the insertion). iii) A translocation results in one half of the read-pair aligning to a different part of the genome iv) A repeat expansion, particularly one longer than the read-pair, will result in one of the pairs sequencing only the repeat region. When alignment is attempted, it may not be able to align anywhere (the other read will be 'unmatched') or may align elsewhere in the genome. D. structural variant detecting method: 'Depth of coverage' – aligned reads of a deleted region identify a length of sequencing with a 50% drop in coverage suggesting a heterozygous deletion i.e. one normal allele and one allele containing a deletion.

Figure 5 Schematic based on the the American College of Medical Genetics and Genomics (ACMG) criteria. Every variant under review has evidence for pathogenicity scrutinised under each of the listed categories.

Figure 6 Clinical cases A-E. Distal limb muscle atrophy in proband of Case 1 F. Integrative Genomics Viewer (IGV) showing heterozygous variant in *ITPR3* (blue arrow) in proband and mother G. Electron microscopy of sural nerve biopsy of Case 1 (proband) showing reduced myelin thickness and onion

bulb formation H. IGV showing 1.5x read depth of aligned reads in *PMP22* compared with a region on chromosome 1 I. IGV shows loss of coverage in intron 2 of RFC1 (red circle) indicating a biallelic repeat expansion in that region. Reads highlighted in red (black arrow) are unmatched pairs J. IGV shows ~9kb 50% drop in coverage (read depth) in MME encompassing exons 15 and 16; corresponding to a heterozygous deletion K-L. Mild distal lower limb atrophy in proband of Case 5 M. IGV shows heterozygous variant in *PMP2* carried by proband and affected father.