Genetic and functional characterisation of ataxias and neuropathies with a focus on biallelic expansions in *RFC1* and biallelic variants in *ARHGAP19*

By Natalia Katarzyna Dominik

Supervised by Doctor Andrea Cortese, Professor Henry Houlden and Doctor Stephanie Efthymiou

A thesis submitted to University College London for the degree of Doctor of Philosophy

Department of Neuromuscular disorders, UCL Queen Square Institute of Neurology, University College London

May 2024

Declaration

I, Natalia Katarzyna Dominik, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Some chapters are a slightly modified version of the work published by myself, and has been reproduced here with the permission of the copyright holder. Collaborative work is also indicated in this thesis.

Natalia Dominik

Abstract

The understanding of genetic and functional aspects of neuropathies and ataxias can provide evidence for therapeutic targets and translation into clinic. In my work, I employed genetic screening and sequencing techniques, optical genome mapping and disease modelling using *Drosophila melanogaster*.

In Chapter 2, I describe the work performed to characterise the RFC1 spectrum disorder. Screening a large cohort of patients for repeat expansions in the gene together with the phenotypic data can give a detailed picture of the disease progression. Here, we demonstrate that *RFC1* repeat size is a key predictor of disease onset, phenotype and severity therefore providing evidence for unmet need of sizing the *RFC1* repeat expansions in diagnostic settings.

In Chapter 3, I describe the genetic heterogeneity in RFC1 disease, and investigate novel pathogenic repeat expansion motifs in *RFC1*. Here, we address a need for additional genetic testing beyond PCR screening in patients presenting with typical CANVAS symptoms but negative screening for the most common biallelic pathogenic AAGGG expansion to correctly diagnose patients.

In chapter 4, I describe the work performed to identify and characterise a novel recessive Charcot-Marie-Tooth (CMT) gene, *ARHGAP19*. Here, we add another important gene to the growing list of CMT genes and we demonstrate loss of activity of the GTPase activating protein domain in functional and in-silico assays.

Finally, in chapter 5, I describe the utility of *Drosophila melanogaster* as a model organism in neurogenetic research. I use various tools to knock down gene expression in the fly to recapitulate the phenotype of the patients with RFC1 and ARHGAP19 diseases.

In conclusion, in my thesis, I present functional and genetic characterisation of two important genes – *RFC1* and *ARHGAP19* - that will enable me and other members of the neurogenetic field to further research the disease mechanisms and address the need of translation into diagnostic and potentially therapeutic avenues.

Impact Statement

The research detailed in this thesis has had a direct translational impact on implementation of diagnostic screening procedures for *RFC1* repeat expansion; implementation in Queens Square Institute of Neurology, and validation of novel repeat expansion sizing methodology using Bionano Optical Genome Mapping and discovery of additional genetic heterogeneity of RFC1 repeat expansion disorder as well as correlation of the size of the repeat expansion to the disease onset and progression which has a direct impact on the ability to better counsel patients affected with the disease. Moreover, a novel gene, *ARHGAP19*, biallelic mutations in which cause Charcot-Marie-Tooth disease, has been genetically and functionally characterised in this thesis and added to a growing list of genes associated with the neuropathy. Additionally, I successfully established *Drosophila melanogaster* loss-of-function models of *RFC1* and *ARHGAP19* gene orthologs which recapitulate the locomotion phenotype of the patients. These models may further help to elucidate the pathomechanisms of these diseases.

The work described in this thesis advances knowledge on the functional consequences of *RFC1* repeat expansions and the novel neuropathy gene *ARHGAP19* and may lead to future therapeutic avenues which could benefit the patients and communities worldwide.

Moreover, the recent disease discovery work presented in this thesis as part of the International Centre for Genomic Medicine in Neuromuscular Diseases (ICGNMD) and Synaptopathies and Paroxysmal Syndromes study groups (SYNaPS) collaborative projects has placed UCL lab at the epicentre of collaborative research. The work generated as part of this thesis but also the many collaborative projects with departments within UCL, other institutions in the UK as well as globally with neurologists, bioinformaticians and basic scientists, have generated a model for future research into complex rare disorders. This work highlights the importance of international collaborations and inclusion of large consortia, big datasets and diverse populations which are essential in genetic studies.

The research described in this thesis has been presented orally at 15th UK Neuromuscular Translational Research Conference in April 2022 in London UK; Postgraduate Research Conference at UCL London in May 2022; The Brain Conference March 2023 (online) and Peripheral Nerve Society (PNS) June 2023 in Copenhagen, Denmark. The work on the novel neuropathy gene *ARHGAP19* has been awarded Guarantors of Brain travel grant of £500 to present in PNS annual meeting, where I was awarded Richard and Mary Bunge award for best presentation.

My work was part of major projects in our laboratory that include contributions from clinicians and scientists, and I describe my contribution at the start of each results chapter. Moreover, parts of the research described in this thesis have been published in international journals which have positive impact on advancing the knowledge of the scientific community.

UCL Research Paper Declaration Forms

1

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

a) **What is the title of the manuscript?**

CANVAS: a late onset ataxia due to biallelic intronic AAGGG expansions

b) **Please include a link to or doi for the work**

10.1007/s00415-020-10183-0

c) **Where was the work published?**

Journal of Neurology

d) **Who published the work?** (e.g. OUP)

Springer

e) **When was the work published?**

10 September 2020

f) **List the manuscript's authors in the order they appear on the publication**

Natalia Dominik, Valentina Galassi Deforie, Andrea Cortese, Henry Houlden

g) **Was the work peer reviewed?**

yes

h) **Have you retained the copyright?**

yes

i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

no

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☒

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

   a) **What is the current title of the manuscript?**

   Click or tap here to enter text.

   b) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

   Click or tap here to enter text.

   c) **Where is the work intended to be published?** (e.g. journal names)

   Click or tap here to enter text.

   d) **List the manuscript's authors in the intended authorship order**

   Click or tap here to enter text.

   e) **Stage of publication** (e.g. in submission)

   Click or tap here to enter text.

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

   ND contributed to experiment design, figure creation, manuscript writing and editing. VGD contributed to experiment design, figure creation, manuscript writing and editing. AC contributed to overseeing the experiment design and reviewed manuscript writing and editing. HH contributed to manuscript editing.

4. **In which chapter(s) of your thesis can this material be found?**

   Chapter 2

5. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

   Candidate

Click or tap here to enter text.

   Date:

   22/05/2024

   Supervisor/ Senior Author (where appropriate)

Click or tap here to enter text.

   Date

   22/05/2024

2

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

6. For a research manuscript that has already been published (if not yet published, please skip to section 2)

j) **What is the title of the manuscript?**

Role of the repeat expansion size in predicting age of onset and severity in RFC1 disease

k) **Please include a link to or doi for the work**

10.1093/brain/awad436

l) **Where was the work published?**

Brain

m) **Who published the work?** (e.g. OUP)

Oxford University Press

n) **When was the work published?**

09/01/2024

o) **List the manuscript's authors in the order they appear on the publication**

Riccardo Currò, **Natalia Dominik**, Stefano Facchini, Elisa Vegezzi, Roisin Sullivan, Valentina Galassi Deforie, Gorka Fernández-Eulate, Andreas Traschütz, Salvatore Rossi, Matteo Garibaldi, Mariusz Kwarciany, Franco Taroni, Alfredo Brusco, Jean-Marc Good, Francesca Cavalcanti, Simon Hammans, Gianina Ravenscroft, Richard H Roxburgh, RFC1 repeat expansion study group , Ricardo Parolin Schnekenberg, Bianca Rugginini, Elena Abati, Arianna Manini, Ilaria Quartesan, Arianna Ghia, Adolfo Lòpez de Munaìn, Fiore Manganelli, Marina Kennerson, Filippo Maria Santorelli, Jon Infante, Wilson Marques, Manu Jokela, Sinéad M Murphy, Paola Mandich, Gian Maria Fabrizi, Chiara Briani, David Gosal, Davide Pareyson, Alberto Ferrari, Ferran Prados, Tarek Yousry, Vikram Khurana, Sheng-Han Kuo, James Miller, Claire Troakes, Zane Jaunmuktane, Paola Giunti, Annette Hartmann, Nazli Basak, Matthis Synofzik, Tanya Stojkovic, Marios Hadjivassiliou, Mary M Reilly, Henry Houlden, and Andrea Cortese

p) **Was the work peer reviewed?**

yes

q) **Have you retained the copyright?**

yes

r) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

no

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☒

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

7. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

   f) **What is the current title of the manuscript?**

   Click or tap here to enter text.

   g) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

   Click or tap here to enter text.

   h) **Where is the work intended to be published?** (e.g. journal names)

   Click or tap here to enter text.

   i) **List the manuscript's authors in the intended authorship order**

   Click or tap here to enter text.

   j) **Stage of publication** (e.g. in submission)

   Click or tap here to enter text.

8. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

   RC contributed to experiment design and experiments, figure creation, manuscript writing and editing; ND contributed to experiment design and experiments, figure creation, manuscript writing and editing SF contributed to bioinformatic analyses, RS and VGD contributed to experiments, the following authors contributed to patient recruitment, clinical investigations and manuscript editing: EV, GF, AT, SR, MG, MK, FT, AB, JMG, FC, SH, GR, RHR, RFC1 repeat expansion study group , RPS, BR, EA, AM, IQ, AG, ALM, FM, MK, FMS, JI, WM, MJ, SMM, PM, GMF, CB, DG, DP, AF, FP, TY, VK, S-HK, JM, CT, ZJ, PG, AH, NB, MS, TS, MH, MMR, HH, AC led the study and contributed to overseeing the experiment design and reviewed manuscript writing and editing

9. **In which chapter(s) of your thesis can this material be found?**

   Chapter 2

10. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

    Candidate

Click or tap here to enter text.

Date:

22/05/2024

Supervisor/ Senior Author (where appropriate)

Click or tap here to enter text.

Date

22/05/2024

3

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

11. For a research manuscript that has already been published (if not yet published, please skip to section 2)

s) **What is the title of the manuscript?**

Normal and pathogenic variation of RFC1 repeat expansions: implications for clinical diagnosis

t) **Please include a link to or doi for the work**

10.1093/brain/awad240

u) **Where was the work published?**

Brain

v) **Who published the work?** (e.g. OUP)

Oxford University Press

w) **When was the work published?**

14 July 2023

x) **List the manuscript's authors in the order they appear on the publication**

Natalia Dominik, Stefania Magri, Riccardo Currò, Elena Abati, Stefano Facchini, Marinella Corbetta, Hannah Macpherson, Daniela Di Bella, Elisa Sarto, Igor Stevanovski, Sanjog R Chintalaphani, Fulya Akcimen, Arianna Manini, Elisa Vegezzi, Ilaria Quartesan, Kylie-Ann Montgomery, Valentina Pirota, Emmanuele Crespan, Cecilia Perini, Glenda Paola Grupelli, Pedro J Tomaselli, Wilson Marques, Genomics England Research Consortium , Joseph Shaw, James Polke, Ettore Salsano, Silvia Fenu, Davide Pareyson, Chiara Pisciotta, George K Tofaris, Andrea H Nemeth, John Ealing, Aleksandar Radunovic, Seamus Kearney, Kishore R Kumar, Steve Vucic, Marina Kennerson, Mary M Reilly, Henry Houlden, Ira Deveson, Arianna Tucci, Franco Taroni, and Andrea Cortese

y) **Was the work peer reviewed?**

yes

z) **Have you retained the copyright?**

yes

aa) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

no

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☒

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

12. For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3)

k) **What is the current title of the manuscript?**

Click or tap here to enter text.

l) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

Click or tap here to enter text.

m) **Where is the work intended to be published?** (e.g. journal names)

Click or tap here to enter text.

n) **List the manuscript's authors in the intended authorship order**

Click or tap here to enter text.

o) **Stage of publication** (e.g. in submission)

Click or tap here to enter text.

13. For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4)

ND contributed to experiment design and experiments, figure creation, manuscript writing and editing, SM contributed to experiment design and experiments, figure creation, manuscript writing and editing, RC contributed to clinical data writing experiment design and experiments, EA contributed to clinical data writing and figure creation, SF contributed to bioinformatic analyses, the following authors contributed to sequencing and manuscript editing MC, HM, DB, ES, IS, SRC, FA, AM, EV, IQ, KAM, VP, EC, CP, GPG, the following authors contributed to patient recruitment and investigations as well as manuscript editing PJT, WM, Genomics England Research Consortium , JS, JP, ES, SF, DP, CP, GKT, AHN, JE, AR, SK, KRK, SV, MK, MMR, HH, ID, AT, FT, AC led the study and contributed to overseeing the experiment design and reviewed manuscript writing and editing

14. In which chapter(s) of your thesis can this material be found?

Chapter 3

15. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate

Click or tap here to enter text.

Date:

22/05/2024

Supervisor/ Senior Author (where appropriate)

Click or tap here to enter text.

Date

22/05/2024

4

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

16. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

bb) **What is the title of the manuscript?**

Optical Genome Mapping Enables Detection and Accurate Sizing of RFC1 Repeat Expansions

cc) **Please include a link to or doi for the work**

https://doi.org/10.3390/biom13101546

dd) **Where was the work published?**

Biomolecules

ee) **Who published the work?** (e.g. OUP)

MDPI

ff) **When was the work published?**

19/10/2023

gg) **List the manuscript's authors in the order they appear on the publication**

Stefano Facchini*, **Natalia Dominik***, Arianna Manini, Stephanie Efthymiou, Riccardo Currò, Bianca Rugginini, Elisa Vegezzi, Ilaria Quartesan, Benedetta Perrone, Shahedah Koya Kutty, Valentina Galassi Deforie, Ricardo P. Schnekenberg, Elena Abati, Anna Pichiecchio, Enza Maria Valente, Cristina Tassorelli, Mary M. Reilly, Henry Houlden, Enrico Bugiardini, Andrea Cortese

hh) **Was the work peer reviewed?**

yes

ii) **Have you retained the copyright?**

yes

jj) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

no

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☒

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

17. For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3)

p) **What is the current title of the manuscript?**

Click or tap here to enter text.

q) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

Click or tap here to enter text.

r) **Where is the work intended to be published?** (e.g. journal names)

Click or tap here to enter text.

s) **List the manuscript's authors in the intended authorship order**

Click or tap here to enter text.

t) **Stage of publication** (e.g. in submission)

Click or tap here to enter text.

18. For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4)

Conceptualization, S.F. and A.C.; methodology, S.F. and N.D.; software, S.F.; investigation, N.D., R.C., S.E., B.P., S.K.K., E.V., I.Q., B.P., R.P.S., E.A., B.R., A.P., E.M.V., C.T. and V.G.D.; writing—original draft preparation, S.F., N.D. and A.M.; writing—review and editing, A.C.; supervision, A.C.; project administration, E.B., M.M.R., A.C. and H.H.; funding acquisition, A.C.

19. In which chapter(s) of your thesis can this material be found?

Chapter 2

20. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate

Click or tap here to enter text.

Date:

22/05/2024

Supervisor/ Senior Author (where appropriate)

Click or tap here to enter text.

Date

22/05/2024

5

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

Please use this form to declare if parts of your thesis are already available in another format, e.g. if data, text, or figures:

- *have been uploaded to a preprint server*
- *are in submission to a peer-reviewed publication*
- *have been published in a peer-reviewed publication, e.g. journal, textbook.*

This form should be completed as many times as necessary. For instance, if you have seven thesis chapters, two of which containing material that has already been published, you would complete this form twice.

21. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

kk) **What is the title of the manuscript?**

Click or tap here to enter text.

ll) **Please include a link to or doi for the work**

Click or tap here to enter text.

mm) **Where was the work published?**

Click or tap here to enter text.

nn) **Who published the work?** (e.g. OUP)

Click or tap here to enter text.

oo) **When was the work published?**

Click or tap here to enter text.

pp) **List the manuscript's authors in the order they appear on the publication**

Click or tap here to enter text.

qq) **Was the work peer reviewed?**

Click or tap here to enter text.

rr) **Have you retained the copyright?**

Click or tap here to enter text.

ss) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

Click or tap here to enter text.

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☐

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

22. For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3)

u) **What is the current title of the manuscript?**

Biallelic variants in ARHGAP19 cause a motor-predominant neuropathy with asymmetry and conduction slowing

v) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

Yes https://doi.org/10.1101/2024.05.10.24306768

w) **Where is the work intended to be published?** (e.g. journal names)

Nature Genetics, Communications, Brain

x) **List the manuscript's authors in the intended authorship order**

Natalia Dominik, Stephanie Efthymiou, Christopher J. Record, Xinyu Miao, Renee Lin, Jevin M. Parmar, Annarita Scardamaglia, Reza Maroofian, Gabriel Aughey, Abigail D. Wilson, Simon Lowe, Riccardo Curro, Ricardo P. Schnekenberg , Shahryar Alavi, Leif Leclaire, Yi He, Kristina Zhelchenska, Yohanns Bellaiche, Isabelle Gaugué, Mariola Skorupinska, Liedewei Van de Vondel, Sahar I. Da'as, Valentina Turchetti, Serdal Güngör, Ehsan Ghayoor Karimiani, Yalda Jamshidi, Phillipa J. Lamont, Camila Armirola Ricaurte, Haluk Topaloglu, Albena Jordanova, Mashaya Zaman, Selina H. Banu, Wilson Marques, Pedro José Tomaselli, Busra Aynekin, Ali Cansu, Huseyin Per, Ayten Güleç, Javeria Raza Alvi, Tipu Sultan, Arif Khan, Giovanni Zifarelli, Shahnaz Ibrahim, Grazia M. S. Mancini, M. Mahdi Motazacker, Esther Brusse, Vincenzo Lupo, Teresa Sevilla, A. Nazlı Başak, Seyma Tekgul, Robin J. Palvadeau, Jonathan Baets, Yesim Parman, Arman Çakar, Rita Horvath, Tobias B. Haack, Jan-Hendrik Stahl, Kathrin Grundmann-Hauser, Joohyun Park, Stephan Züchner, Nigel G. Laing, Lindsay Wilson, Alexander M. Rossor, James Polke, Fernanda Barbosa Figueiredo, André Luiz Pessoa, Fernando Kok, Antônio Rodrigues Coimbra-Neto, Marcondes C. França Jr, Gianina Ravenscroft, Sherifa Ahmed Hamed, Wendy K. Chung, Daniel P. Osborn, Michael Hanna, Andrea Cortese, Mary M. Reilly, James E. C. Jepson, Nathalie Lamarche-Vane, Henry Houlden

y) **Stage of publication** (e.g. in submission)

In preparation for submission

23. For multi-authored work, please give a statement of contribution covering all authors (if single-author, please skip to section 4)

ND and SE contributed to manuscript writing, experimental design and experimental work, CJR contributed to clinical examination revision and writing, XM contributed GAP activity assays, RL contributed zebrafish work, JMP contributed in silico studies, AS contributes fibroblast work, RM contributed sequencing analysis, the following contributed help with drosophila work: GA, ADW, SL, the following contributed MN work: RC, RPS, SA contributed in silico RNAseq and haplotype, the following contributed GAP activity assays LL, YH, KZ contributed fibroblast work, the following contributed fly stocks: YB, IG, the following contributed patient examinations and data collections MS, LVV, SID, VT, SG, EGK, YJ, PJL, CAR, HT, AJ, MZ, SHB, WM, PJT, BA, AC, HP, AG, JRA, TS, AK, GZ, SI, GMSM, MMM, EB, VL, TS, ANB, ST, RJP, JB, YP, AÇ, RH, TBH, JHS, KGH, JP, SZ, NGL, LW, AMR, JP, FBF, ALP, FK, ARCN, MC F, GR, SAH, WKC, DPO, MH, AC, the following authors contributed to funding the study and contributed to overseeing the experiment design and reviewed manuscript writing and editing MMR, JECJ, NLV, HH

24. In which chapter(s) of your thesis can this material be found?

Chapter 4 and 5

25. e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate

Date:

22/05/2024

Supervisor/ Senior Author (where appropriate)

Click or tap here to enter text.

Date

22/05/2024

Acknowledgements

First and foremost, I would like to thank my supervisors Dr Andrea Cortese, Prof Henry Houlden and Dr Stephanie Efthymiou.

To Prof Houlden, thank you for accepting me into your lab in 2019. Thank you for providing me with the exciting research opportunities, mentoring and guiding me and allowing me to shape and explore my research interests. I will always remember and appreciate the positive and collaborative atmosphere you have fostered.

To Dr Cortese, I am deeply grateful to you for accepting me into your group. Thank you for always making time for me, teaching me many valuable lab and research skills and for your confidence in my abilities. Your mentorship, patience, and the excellent leadership have allowed me to thrive as a young researcher.

To Dr Efthymiou, thank you for your guidance and the invaluable lab and collaboration skills that you have taught me over the years. Thank you for encouraging me, providing constant support and guidance which have been crucial to my development. And importantly, thank you for all the times when we worked hand in hand, and made a great team.

Secondly, I would like to express my gratitude to Prof James Jepson for his collaboration and supervision on all the *Drosophila* projects. Prof Jepson, thank you for accepting me in your lab, for your kindness and encouragement and always making me feel welcome. I have enjoyed learning about *Drosophila* as a model organism and being able to use various tools in our research.

I would also like to take an opportunity to thank the ICGNMD project, particularly Prof Michael Hanna for having me as part of the Team, Dr Lindsay Wilson for all her invaluable help, kindness and encouragement and Prof Mary Reilly for all her support, guidance and positivity. Thank you to Dr Enrico Bugiardini for having me as part of the Bionano project and entrusting me with the precious samples and the whole pipeline. I would also like to thank all the Colleagues and Collaborators from ICGNMD and SYNAPS projects as well as RFC1 and ARHGAP19 studies.

I thank my thesis committee, Prof Robert Pitceathly and Dr Jana Vandrovcova for helping me along on this journey and providing invaluable insights and comments on my work. I also would like to thank my viva examiners, Dr Alice Davidson and Dr Alessandra Bolino for taking the time out of their busy schedules to join me in the discussion of my work.

I am grateful to Medical Research Council (MRC) and Guarantors of Brain for the financial support behind my projects and the opportunity to travel to conferences.

Very importantly, I would like to thank all the patients and their families, without them this research would not be possible.

To all my closest Colleagues and Teammates, thank you all for being my rocks and making work even more fun. Special thanks to Dr Riccardo Curro for always listening to me, never laughing at my silly questions, giving fantastic clinical advice and processing hundreds of samples with me. To Dr Christopher Record, thank you for all your CMT expertise, your hard work and help with my presentations but most importantly thank you for finding the time to help me, and being so kind and encouraging. To Dr Valentina Galassi Deforie, thank you for putting in the long hours with me, having a laugh when things went wrong and all those Southern blots. To Dr Abigail D. Wilson, thank you for your guidance on all things fly and work but also thank you for being a wonderful lab run partner. To Dr Gabriel Aughey, thank you for constructively criticising my work and helping to come up with way too many experiments but also for laughing at my jokes. To Dr Simon Lowe, thank you for all the dissections and confocal work but also for your patience when I started the fly work and probably asked too many questions. To Dr Stefano Facchini, thank you for all your help with the data analysis and all the hard work you put into our long-read sequencing and Bionano data but also for being kind when I needed refreshers on the same question I had already asked twice.

Thank you to all the past and present members of the Houlden and Cortese labs, Dr Reza Maroofian, Dr Roisin Sullivan (special thanks for teaching me Southern botting), Valentina Turchetti, Dr Viorica Chelban, Dr Rauan Kairzhanov, Hau Ying Yip (Janice), Dr Menelaos Pipis, Marilena Christoforou, Dr Ben O'Callaghan, Dr Elena Abati, Dr Riccardo Ronco, Dr Roberto Simone, Dr Arianna Manini, Dr Arianna Ghia, Dr Ricardo P Schnekenberg, Dr Sara Nagy, Dr Elisa Vegezzi, Alice Gennari, Benedetta Perrone, Clarissa Rocca, Elisa Cali, Rahema Mohammad, Dr Emil Gustavsson, Malgorzata Murray, Dr Francesca Magrinelli, Dr Zhongbo Chen, David Murphy, Yee Goh, Tracy Lau, Kristina Zhelcheska, Ellie Self, Dr Shahryar Alavi, Renee Lin, Irem Karagoz, Ambreen Tariq, you have all helped me a lot during the last years. Special thank you to Annarita Scardamaglia for all TC and logistics help, Dr James Polke for the diagnostic lab perspectives, Hannah Macpherson, Jasmaine Lee and Kylie-Ann

Montgomery for all the long-read sequencing help and tutoring and Mark Gaskin for all the lab management and sample requests.

A huge thank you to the IoN Administration teams, especially Kully Sunner and Debbie Hadley and all the IT Support Teams. Thank you also to the IoN Security Team. You have all been extremely helpful.

On a person note, I would like to thank my partner, Wojciech, for his sacrifices and ongoing support to make this journey possible from the start to finish. I appreciate it has not been easy, but I am so grateful I always have you by my side.

Thank you to my mum, for teaching me to persevere in the hardest of times, to my dad for teaching me the discipline and to my brother for always being keen to help. And to my good friend here in London, Elisa Corsiero. Ely thanks for listening to my problems and giving me your perspectives, finding time to squeeze in a coffee chat or a much needed bike ride.

And finally, and most importantly, I thank my son, Henry. Thank you for making mama happy, teaching me what taking a break really means and sharing your toys when I needed encouragement. It has been fun being able to join many Teams meetings and have you in them with me since you were tiny. I hope one day you will read this and be very proud of mummy but also you will be inspired to follow the career path that makes you happy.

Abbreviations

Abbreviation	Full description
AAO	Age at symptom onset
ACMG	American College of Medical Genetics and Genomics
ARCAs	Autosomal recessive cerebellar ataxias
ARSACS	Autosomal recessive spastic ataxia of Charlevoix-Saguenay
ATP	Adenosine triphosphate
AVV	Adenosine Associated Virus
BCA	Bicinchoninic acid
BCR	Breakpoint cluster region protein
BSA	Bovine serum albumin
CANVAS	Cerebellar ataxia, neuropathy and vestibular areflexia syndrome
CCS	Circular consensus sequencing
CI	Confidence interval
CIDP	Chronic inflammatory demyelinating polyneuropathy
CMT	Charcot-Marie-Tooth
CMTi	Intermediate CMT
CRIMIC	CRISPR Mediated Integration Cassette
CRISPants	CRISPR-induced mutants
CRISPR/Cas9	Clustered Regularly Interspaced Short Palindromic Repeats/ CRISPR-associated endonuclease 9
DAM	Drosophila Activity Monitor
dHMN	Distal hereditary motor neuropathy
DIG	Digoxigenin
DLS	Direct Label and Stain
DM1	Myotonic dystrophy type 1
dpf	Days post-fertilization
EB	Embryoid bodies
EHDN	ExpansionHunterDeNovo
EMG	Electromyography
ExAC	Exome Aggregation Consortium
FBS	Fetal bovine serum
FRDA	Friedreich's ataxia
G4s	G-quadruplexes
GAP	GTPase activating protein
GC	Guanine – Cytosine
GDF	GDI displacement factor
GDIs	Guanine Nucleotide Dissociation Inhibitors
GDP	Guanine nucleotide diphosphate
GEFs	Guanine exchange factors
GEL	Genomics England
GERP	Genetic evolutionary rate profiling
GFP	Green fluorescent protein

GJB1	Gap junction protein beta 1
gnomAD	Genome Aggregation Database
GOF	Gain-of-function
GST	Glutathione-S-transferase
GTEx	Genotype-Tissue Expression
GTP	Guanosine triphosphate
H2Av	Histone 2A variant
HMSN	Hereditary motor and sensory neuropathy
HPO	Human Phenotype Ontology
ICC	Immunocytochemistry
ICGNMD	International Centre for Genomic Medicine in Neuromuscular Diseases
Indels	Insertions or deletions
iPSC	Induced pluripotent stem cells
IPTG	Isopropyl β- d-1-thiogalactopyranoside
IRB	Institutional review board
LBB	Lysis and Binding Buffer
LOF	Loss-of-function
LRS	Long read sequencing
MFN2	Mitofusin 2
MLPA	Multiplex ligation-dependent probe amplification
MO	Morpholino
MPZ	Myelin protein zero
MRCA	Most recent common ancestor
MSA	Multiple System Atrophy
NCS	Nerve conduction studies
NGS	Next generation sequencing
NHNN	National Hospital for Neurology and Neurosurgery
NHS	UK National Health Service
Nrf2	Erythroid 2-related factor 2
OGM	Optical genome mapping
OMIM	Online Mendelian Inheritance in Man
ONT	Oxford Nanopore Technologies
PacBio	Pacific Bioscience
PBS	Phosphate Buffered Saline
PCA	Principal component analysis
PCNA	Proliferating cell nuclear antigen
PMP22	Peripheral myelin protein 22
QGRS	Quadruplex forming G-Rich Sequences
qPCR	Quantitative PCR
RAN	Repeat associated non-AUG
RE	Repeat expansion
RFC	Replication factor C complex
RISC	RNAi silencing complex
RNAi	RNA interference

RNP	Ribonucleoprotein
ROCK	Rho-associated protein kinase
RP-PCR	Repeat primed PCR
SB	Southern Blotting
SCAs	Autosomal dominant spinocerebellar ataxias
SD	Standard deviation
SEM	Standard error of the mean
shRNA	Short harpin RNA
SMRT	Single molecule real time sequencing
SNPs	Single nucleotide polymorphisms
SORD	Sorbitol dehydrogenase
spMNs	Spinal motor neurons
SSC	Saline-sodium citrate
STR	Short tandem repeats
SV	Structural variations
SYD1	Synapse defective protein 1
SYNaPS	Synaptopathies and Paroxysmal Syndromes study groups
SYNE1	Spectrin repeat-containing nuclear envelope 1
TAE	Tris-Acetate-EDTA
UAS	Upstream activating sequence
UCL	University College London
UIC	Uninjected Control
VCF	Variant calling format
VNC	Ventral Nerve Cord
VUS	Variant of uncertain significance
WBC	White blood cells
WES	Whole exome sequencing
WGS	Whole genome sequencing
WISH	Whole mount in situ hybridization
WT	Wild type
ZT	Zeitgeber Time

Table of Contents

List of publications

Part of the work described in this thesis has been published in the following articles:

DOMINIK, N., EFTHYMIOU, S., RECORD, C. J., MIAO, X., LIN, R., PARMAR, J., SCARDAMAGLIA, A., MAROOFIAN, R., AUGHEY, G., WILSON, A., LOWE, S., CURRO, R., SCHNEKENBERG, R. P., ALAVI, S., LECLAIRE, L., HE, Y., ZHELCHENSKA, K., BELLAICHE, Y., GAUGUE, I., SKORUPINSKA, M., VAN DE VONDEL, L., DA'AS, S., TURCHETTI, V., GUNGOR, S., KARIMIANI, E. G., ARMIROLA RICAURTE, C., TOPALOGLU, H., JORDANOVA, A., ZAMAN, M., BANU, S. H., MARQUES, W., TOMASELLI, P. J., AYNEKIN, B., CANSU, A., PER, H., GULEC, A., ALVI, J. R., SULTAN, T., KHAN, A., ZIFARELLI, G., IBRAHIM, S., MANCINI, G. M. S., MOTAZACKER, M. M., BRUSSE, E., LUPO, V., SEVILLA, T., TEKGUL, S., PALVADEAU, R., BASAK, A. N., BAETS, J., PARMAN, Y., CAKAR, A., HORVATH, R., HAACK, T. B., STAHL, J.-H., GRUNDMANN-HAUSER, K., PARK, J., ZUCHNER, S., LAING, N. G., WILSON, L., ROSSOR, A. M., POLKE, J., FIGUEIREDO, F. B., PESSOA, A. L., KOK, F., COIMBRA-NETO, A. R., FRANCA, M. C., JAMSHIDI, Y., RAVENSCROFT, G., AHMED HAMED, S., CHUNG, W. K., OSBORN, D. P., HANNA, M., CORTESE, A., JEPSON, J. E. C., REILLY, M. M., LAMARCHE-VANE, N. & HOULDEN, H. 2024. Biallelic variants in ARHGAP19 cause a motor-predominant neuropathy with asymmetry and conduction slowing. *medRxiv*, 2024.05.10.24306768.

DOMINIK, N., GALASSI DEFORIE, V., CORTESE, A. & HOULDEN, H. 2021. CANVAS: a late onset ataxia due to biallelic intronic AAGGG expansions. *J Neurol,* 268**,** 1119-1126.

DOMINIK, N., MAGRI, S., CURRO, R., ABATI, E., FACCHINI, S., CORBETTA, M., MACPHERSON, H., DI BELLA, D., SARTO, E., STEVANOVSKI, I., CHINTALAPHANI, S. R., AKCIMEN, F., MANINI, A., VEGEZZI, E., QUARTESAN, I., MONTGOMERY, K. A., PIROTA, V., CRESPAN, E., PERINI, C., GRUPELLI, G. P., TOMASELLI, P. J., MARQUES, W., GENOMICS ENGLAND RESEARCH, C., SHAW, J., POLKE, J., SALSANO, E., FENU, S., PAREYSON, D., PISCIOTTA, C., TOFARIS, G. K., NEMETH, A. H., EALING, J., RADUNOVIC, A., KEARNEY, S., KUMAR, K. R., VUCIC, S., KENNERSON, M., REILLY, M. M., HOULDEN, H., DEVESON, I., TUCCI, A., TARONI, F. & CORTESE, A. 2023. Normal and pathogenic variation of RFC1 repeat expansions: implications for clinical diagnosis. *Brain,* 146**,** 5060-5069.

CURRO, R., **DOMINIK, N**., FACCHINI, S., VEGEZZI, E., SULLIVAN, R., GALASSI DEFORIE, V., FERNANDEZ-EULATE, G., TRASCHUTZ, A., ROSSI, S., GARIBALDI, M., KWARCIANY, M., TARONI, F., BRUSCO, A., GOOD, J. M., CAVALCANTI, F., HAMMANS, S., RAVENSCROFT, G., ROXBURGH, R. H., GROUP, R. F. C. R. E. S., PAROLIN SCHNEKENBERG, R., RUGGININI, B., ABATI, E., MANINI, A., QUARTESAN, I., GHIA, A., LOPEZ DE MUNAIN, A., MANGANELLI, F., KENNERSON, M., SANTORELLI, F. M., INFANTE, J., MARQUES, W., JOKELA, M., MURPHY, S. M., MANDICH, P., FABRIZI, G. M., BRIANI, C., GOSAL, D., PAREYSON, D., FERRARI, A., PRADOS, F., YOUSRY, T., KHURANA, V., KUO, S. H., MILLER, J., TROAKES, C., JAUNMUKTANE, Z., GIUNTI, P., HARTMANN, A., BASAK, N., SYNOFZIK, M., STOJKOVIC, T., HADJIVASSILIOU, M., REILLY, M. M., HOULDEN, H. & CORTESE, A. 2024.

Role of the repeat expansion size in predicting age of onset and severity in RFC1 disease. *Brain,* 147**,** 1887-1898.

FACCHINI, S., **DOMINIK, N**., MANINI, A., EFTHYMIOU, S., CURRO, R., RUGGININI, B., VEGEZZI, E., QUARTESAN, I., PERRONE, B., KUTTY, S. K., GALASSI DEFORIE, V., SCHNEKENBERG, R. P., ABATI, E., PICHIECCHIO, A., VALENTE, E. M., TASSORELLI, C., REILLY, M. M., HOULDEN, H., BUGIARDINI, E. & CORTESE, A. 2023. Optical Genome Mapping Enables Detection and Accurate Sizing of RFC1 Repeat Expansions. *Biomolecules,* 13.

Other publications coauthored during this thesis:

CURRO, R., SALVALAGGIO, A., TOZZA, S., GEMELLI, C., **DOMINIK, N**., GALASSI DEFORIE, V., MAGRINELLI, F., CASTELLANI, F., VEGEZZI, E., BUSINARO, P., CALLEGARI, I., PICHIECCHIO, A., COSENTINO, G., ALFONSI, E., MARCHIONI, E., COLNAGHI, S., GANA, S., VALENTE, E. M., TASSORELLI, C., EFTHYMIOU, S., FACCHINI, S., CARR, A., LAURA, M., ROSSOR, A. M., MANJI, H., LUNN, M. P., PEGORARO, E., SANTORO, L., GRANDIS, M., BELLONE, E., BEAUCHAMP, N. J., HADJIVASSILIOU, M., KASKI, D., BRONSTEIN, A. M., HOULDEN, H., REILLY, M. M., MANDICH, P., SCHENONE, A., MANGANELLI, F., BRIANI, C. & CORTESE, A. 2021. RFC1 expansions are a common cause of idiopathic sensory neuropathy. *Brain,* 144**,** 1542-1550.

DWORSCHAK, G. C., PUNETHA, J., KALANITHY, J. C., MINGARDO, E., ERDEM, H. B., AKDEMIR, Z. C., KARACA, E., MITANI, T., MARAFI, D., FATIH, J. M., JHANGIANI, S. N., HUNTER, J. V., DAKAL, T. C., DHABHAI, B., DABBAGH, O., ALSAIF, H. S., ALKURAYA, F. S., MAROOFIAN, R., HOULDEN, H., EFTHYMIOU, S., **DOMINIK, N**., SALPIETRO, V., SULTAN, T., HAIDER, S., BIBI, F., THIELE, H., HOEFELE, J., RIEDHAMMER, K. M., WAGNER, M., GUELLA, I., DEMOS, M., KEREN, B., BURATTI, J., CHARLES, P., NAVA, C., HERON, D., HEIDE, S., VALKANAS, E., WADDELL, L. B., JONES, K. J., OATES, E. C., COOPER, S. T., MACARTHUR, D., SYRBE, S., ZIEGLER, A., PLATZER, K., OKUR, V., CHUNG, W. K., O'SHEA, S. A., ALCALAY, R., FAHN, S., MARK, P. R., GUERRINI, R., VETRO, A., HUDSON, B., SCHNUR, R. E., HOGANSON, G. E., BURTON, J. E., MCENTAGART, M., LINDENBERG, T., YILMAZ, O., ODERMATT, B., PEHLIVAN, D., POSEY, J. E., LUPSKI, J. R. & REUTTER, H. 2021. Biallelic and monoallelic variants in PLXNA1 are implicated in a novel neurodevelopmental disorder with variable cerebral and eye anomalies. *Genet Med,* 23**,** 1715-1725.

EFTHYMIOU, S., LEMMERS, R., VISHNU, V. Y., **DOMINIK, N**., PERRONE, B., FACCHINI, S., VEGEZZI, E., RAVAGLIA, S., WILSON, L., VAN DER VLIET, P. J., MISHRA, R., REYAZ, A., AHMAD, T., BHATIA, R., POLKE, J. M., SRIVASTAVA, M. P., CORTESE, A., HOULDEN, H., VAN DER MAAREL, S. M., HANNA, M. G. & BUGIARDINI, E. 2023. Optical Genome Mapping for the Molecular Diagnosis of Facioscapulohumeral Muscular Dystrophy: Advancement and Challenges. *Biomolecules,* 13.

HADJIVASSILIOU, M., CURRO, R., BEAUCHAMP, N., **DOMINIK, N**., GRUNEWALD, R. A., SHANMUGARAJAH, P., ZIS, P., HOGGARD, N. & CORTESE, A. 2024. Can CANVAS due to RFC1 biallelic expansions present with pure ataxia? *J Neurol Neurosurg Psychiatry,* 95**,** 171-174.

HUSAIN, R. A., JIAO, X., HENNINGS, J. C., GIESECKE, J., PALSULE, G., BECK-WODL, S., OSMANOVIC, D., BJORGO, K., MIR, A., ILYAS, M., ABBASI, S. M., EFTHYMIOU, S., **DOMINIK, N**., MAROOFIAN, R., HOULDEN, H., RANKIN, J., PAGNAMENTA, A. T., NASHABAT, M., ALTWAIJRI, W., ALFADHEL, M., UMAIR, M., KHOUJ, E., REARDON, W., EL-HATTAB, A. W., MEKKI, M., HOUGE, G., BEETZ, C., BAUER, P., PUTOUX, A., LESCA, G., SANLAVILLE, D., ALKURAYA, F. S., TAYLOR, R. W., MENTZEL, H. J., HUBNER, C. A., HUPPKE, P., HART, R. P., HAACK, T. B., KILEDJIAN, M. & RUBIO, I. 2024. Biallelic NUDT2 variants defective in mRNA decapping cause a neurodevelopmental disease. *Brain,* 147**,** 1197-1205.

KAIYRZHANOV, R., ZAKI, M. S., MAROOFIAN, R., **DOMINIK, N**., RAD, A., VONA, B. & HOULDEN, H. 2021. A Novel Homozygous ADCY5 Variant is Associated with a Neurodevelopmental Disorder and Movement Abnormalities. *Mov Disord Clin Pract,* 8**,** 1140-1143.

LUO, H., GUSTAVSSON, E. K., MACPHERSON, H., **DOMINIK, N**., ZHELCHESKA, K., MONTGOMERY, K., ANDERSON, C., YAU, W. Y., EFTHYMIOU, S., TURNER, C., DETURE, M., DICKSON, D. W., JOSEPHS, K. A., REVESZ, T., LASHLEY, T., HALLIDAY, G., ROWE, D. B., MCCANN, E., BLAIR, I., LEES, A. J., TIENARI, P. J., SUOMALAINEN, A., MOLINA-PORCEL, L., KOVACS, G. G., GELPI, E., HARDY, J., HALTIA, M. J., TUCCI, A., JAUNMUKTANE, Z., RYTEN, M., HOULDEN, H. & CHEN, Z. 2024. Letter to the editor on: Hornerin deposits in neuronal intranuclear inclusion disease: direct identification of proteins with compositionally biased regions in inclusions by Park *et al*. (2022). *Acta Neuropathol Commun,* 12**,** 2.

PAGNAMENTA, A. T., KAIYRZHANOV, R., ZOU, Y., DA'AS, S. I., MAROOFIAN, R., DONKERVOORT, S., **DOMINIK, N**., LAUFFER, M., FERLA, M. P., ORIOLI, A., GIESS, A., TUCCI, A., BEETZ, C., SEDGHI, M., ANSARI, B., BARRESI, R., BASIRI, K., CORTESE, A., ELGAR, G., FERNANDEZ-GARCIA, M. A., YIP, J., FOLEY, A. R., GUTOWSKI, N., JUNGBLUTH, H., LASSCHE, S., LAVIN, T., MARCELIS, C., MARKS, P., MARINI-BETTOLO, C., MEDNE, L., MOSLEMI, A. R., SARKOZY, A., REILLY, M. M., MUNTONI, F., MILLAN, F., MURARESKU, C. C., NEED, A. C., NEMETH, A. H., NEUHAUS, S. B., NORWOOD, F., O'DONNELL, M., O'DRISCOLL, M., RANKIN, J., YUM, S. W., ZOLKIPLI-CUNNINGHAM, Z., BRUSIUS, I., WUNDERLICH, G., GENOMICS ENGLAND RESEARCH, C., KARAKAYA, M., WIRTH, B., FAKHRO, K. A., TAJSHARGHI, H., BONNEMANN, C. G., TAYLOR, J. C. & HOULDEN, H. 2021. An ancestral 10-bp repeat expansion in VWA1 causes recessive hereditary motor neuropathy. *Brain,* 144**,** 584-600.

QUARTESAN, I., VEGEZZI, E., CURRO, R., HESLEGRAVE, A., PISCIOTTA, C., IRUZUBIETA, P., SALVALAGGIO, A., FERNANDEZ-EULATE, G., **DOMINIK, N.,** RUGGININI, B., MANINI, A., ABATI, E., FACCHINI, S., MANSO, K., ALBAJAR, I., LABAN, R., ROSSOR, A. M., PICHIECCHIO, A., COSENTINO, G., SAVERI, P., SALSANO, E., ANDREETTA, F., VALENTE, E. M., ZETTERBERG, H., GIUNTI, P., STOJKOVIC, T., BRIANI, C., LOPEZ DE MUNAIN, A., PAREYSON, D., REILLY, M. M., HOULDEN, H., TASSORELLI, C. & CORTESE, A. 2024. Serum Neurofilament Light Chain in Replication Factor Complex Subunit 1 CANVAS and Disease Spectrum. *Mov Disord,* 39**,** 209-214.

RONCO, R., PERINI, C., CURRO, R., **DOMINIK, N**., FACCHINI, S., GENNARI, A., SIMONE, R., STUART, S., NAGY, S., VEGEZZI, E., QUARTESAN, I., EL-SADDIG, A., LAVIN, T., TUCCI, A., SZYMURA, A., NOVIS DE FARIAS, L. E.,

GARY, A., DELFELD, M., KANDIKATLA, P., NIU, N., TAWDE, S., SHAW, J., POLKE, J., REILLY, M. M., WOOD, N. W., CRESPAN, E., GOMEZ, C., CHEN, J. Y. H., SCHMAHMANN, J. D., GOSAL, D., HOULDEN, H., DAS, S. & CORTESE, A. 2023. Truncating Variants in RFC1 in Cerebellar Ataxia, Neuropathy, and Vestibular Areflexia Syndrome. *Neurology,* 100**,** e543-e554.

SAIDA, K., MAROOFIAN, R., SENGOKU, T., MITANI, T., PAGNAMENTA, A. T., MARAFI, D., ZAKI, M. S., O'BRIEN, T. J., KARIMIANI, E. G., KAIYRZHANOV, R., TAKIZAWA, M., OHORI, S., LEONG, H. Y., AKAY, G., GALEHDARI, H., ZAMANI, M., ROMY, R., CARROLL, C. J., TOOSI, M. B., ASHRAFZADEH, F., IMANNEZHAD, S., MALEK, H., AHANGARI, N., TOMOUM, H., GOWDA, V. K., SRINIVASAN, V. M., MURPHY, D., **DOMINIK, N.,** ELBENDARY, H. M., RAFAT, K., YILMAZ, S., KANMAZ, S., SERIN, M., KRISHNAKUMAR, D., GARDHAM, A., MAW, A., RAO, T. S., ALSUBHI, S., SROUR, M., BUHAS, D., JEWETT, T., GOLDBERG, R. E., SHAMSELDIN, H., FRENGEN, E., MISCEO, D., STROMME, P., MAGLIOCCO CERONI, J. R., KIM, C. A., YESIL, G., SENGENC, E., GULER, S., HULL, M., PARNES, M., AKTAS, D., ANLAR, B., BAYRAM, Y., PEHLIVAN, D., POSEY, J. E., ALAVI, S., MADANI MANSHADI, S. A., ALZAIDAN, H., AL-OWAIN, M., ALABDI, L., ABDULWAHAB, F., SEKIGUCHI, F., HAMANAKA, K., FUJITA, A., UCHIYAMA, Y., MIZUGUCHI, T., MIYATAKE, S., MIYAKE, N., ELSHAFIE, R. M., SALAYEV, K., GULIYEVA, U., ALKURAYA, F. S., GLEESON, J. G., MONAGHAN, K. G., LANGLEY, K. G., YANG, H., MOTAVAF, M., SAFARI, S., ALIPOUR, M., OGATA, K., BROWN, A. E. X., LUPSKI, J. R., HOULDEN, H. & MATSUMOTO, N. 2023. Brain monoamine vesicular transport disease caused by homozygous SLC18A2 variants: A study in 42 affected individuals. *Genet Med,* 25**,** 90-102.

SCRIBA, C. K., BEECROFT, S. J., CLAYTON, J. S., CORTESE, A., SULLIVAN, R., YAU, W. Y., DOMINIK, N., RODRIGUES, M., WALKER, E., DYER, Z., WU, T. Y., DAVIS, M. R., CHANDLER, D. C., WEISBURD, B., HOULDEN, H., REILLY, M. M., LAING, N. G., LAMONT, P. J., ROXBURGH, R. H. & RAVENSCROFT, G. 2020. A novel RFC1 repeat motif (ACAGG) in two Asia-Pacific CANVAS families. *Brain,* 143**,** 2904-2910.

SULLIVAN, R., YAU, W. Y., CHELBAN, V., ROSSI, S., **DOMINIK, N**., O'CONNOR, E., HARDY, J., WOOD, N., CORTESE, A. & HOULDEN, H. 2021. RFC1-related ataxia is a mimic of early multiple system atrophy. *J Neurol Neurosurg Psychiatry,* 92**,** 444-6.

TOZZA, S., CORTESE, A., IOVINO, A., ESPOSITO, M., **DOMINIK, N**., IODICE, R. & MANGANELLI, F. 2021. Bedside Head Impulse Test: A Useful Tool for Patients With Sensory Ataxia. *Neurol Genet,* 7**,** e541.

TRASCHUTZ, A., CORTESE, A., REICH, S., **DOMINIK, N**., FABER, J., JACOBI, H., HARTMANN, A. M., RUJESCU, D., MONTAUT, S., ECHANIZ-LAGUNA, A., ERER, S., SCHUTZ, V. C., TARNUTZER, A. A., STURM, M., HAACK, T. B., VAUCAMPS-DIEDHIOU, N., PUCCIO, H., SCHOLS, L., KLOCKGETHER, T., VAN DE WARRENBURG, B. P., PAUCAR, M., TIMMANN, D., HILGERS, R. D., GAZULLA, J., STRUPP, M., MORIS, G., FILLA, A., HOULDEN, H., ANHEIM, M., INFANTE, J., BASAK, A. N., SYNOFZIK, M. & GROUP, R. F. C. S. 2021. Natural History, Phenotypic Spectrum, and Discriminative Features of Multisystemic RFC1 Disease. *Neurology,* 96**,** e1369-e1382.

ULLAH, I., WAQAS, M., ILYAS, M., HALIM, S. A., AHMAD, A., **DOMINIK, N**., ULLAH, W., ABBAS, M., AAMIR, M., GROUP, S. Y. S., QUEEN SQUARE, G., HOULDEN, H., EFTHYMIOU, S., KHAN, A. & AL-HARRASI, A. 2023. A novel

variant of GALC in a familial case of Krabbe disease: Insights from structural bioinformatics and molecular dynamics simulation. *Genes Dis,* 10**,** 2263-2266.

VAN HAUTE, L., O'CONNOR, E., DIAZ-MALDONADO, H., MUNRO, B., POLAVARAPU, K., HOCK, D. H., ARUNACHAL, G., ATHANASIOU-FRAGKOULI, A., BARDHAN, M., BARTH, M., BONNEAU, D., BRUNETTI-PIERRI, N., CAPPUCCIO, G., CARUANA, N. J., **DOMINIK, N**., GOEL, H., HELMAN, G., HOULDEN, H., LENAERS, G., MENTION, K., MURPHY, D., NANDEESH, B., OLIMPIO, C., POWELL, C. A., PREETHISH-KUMAR, V., PROCACCIO, V., RIUS, R., REBELO-GUIOMAR, P., SIMONS, C., VENGALIL, S., ZAKI, M. S., ZIEGLER, A., THORBURN, D. R., STROUD, D. A., MAROOFIAN, R., CHRISTODOULOU, J., GUSTAFSSON, C., NALINI, A., LOCHMULLER, H., MINCZUK, M. & HORVATH, R. 2023. TEFM variants impair mitochondrial transcription causing childhood-onset neurological disease. *Nat Commun,* 14**,** 1009.

VISHNU, V. Y., LEMMERS, R., REYAZ, A., MISHRA, R., AHMAD, T., VAN DER VLIET, P. J., KRETKIEWICZ, M. M., MACKEN, W. L., EFTHYMIOU, S., **DOMINIK, N**., MORROW, J. M., BHATIA, R., WILSON, L. A., HOULDEN, H., HANNA, M. G., BUGIARDINI, E., VAN DER MAAREL, S. M. & SRIVASTAVA, M. V. P. 2024. The first genetically confirmed cohort of Facioscapulohumeral Muscular Dystrophy from Northern India. *Eur J Hum Genet*.

WILSON, L. A., MACKEN, W. L., PERRY, L. D., RECORD, C. J., SCHON, K. R., FREZATTI, R. S. S., RAGA, S., NAIDU, K., KOKEN, O. Y., POLAT, I., KAPAPA, M. M., **DOMINIK, N**., EFTHYMIOU, S., MORSY, H., NEL, M., FASSAD, M. R., GAO, F., PATEL, K., SCHOONEN, M., BISSCHOFF, M., VORSTER, A., JONVIK, H., HUMAN, R., LUBBE, E., NONYANE, M., VENGALIL, S., NASHI, S., SRIVASTAVA, K., LEMMERS, R., REYAZ, A., MISHRA, R., TOPF, A., TRAINOR, C. I., STEYN, E. C., MAHUNGU, A. C., VAN DER VLIET, P. J., CEYLAN, A. C., HIZ, A. S., CAVDARLI, B., SEMERCI GUNDUZ, C. N., CEYLAN, G. G., NAGAPPA, M., TALLAPAKA, K. B., GOVINDARAJ, P., VAN DER MAAREL, S. M., NARAYANAPPA, G., NANDEESH, B. N., WA SOMWE, S., BEARDEN, D. R., KVALSUND, M. P., RAMDHARRY, G. M., OKTAY, Y., YIS, U., TOPALOGLU, H., SARKOZY, A., BUGIARDINI, E., HENNING, F., WILMSHURST, J. M., HECKMANN, J. M., MCFARLAND, R., TAYLOR, R. W., SMUTS, I., VAN DER WESTHUIZEN, F. H., SOBREIRA, C., TOMASELLI, P. J., MARQUES, W., JR., BHATIA, R., DALAL, A., SRIVASTAVA, M. V. P., YAREEDA, S., NALINI, A., VISHNU, V. Y., THANGARAJ, K., STRAUB, V., HORVATH, R., CHINNERY, P. F., PITCEATHLY, R. D. S., MUNTONI, F., HOULDEN, H., VANDROVCOVA, J., REILLY, M. M. & HANNA, M. G. 2023. Neuromuscular disease genetics in under-represented populations: increasing data diversity. *Brain,* 146**,** 5098-5109.

CHAPTER 1. General introduction

1.1 Rare disorders

Rare diseases are disorders which affect a small proportion of a given population and can have debilitating effects on the quality of life of affected individuals and their families. The definition of a rare disorder varies by region, for example in Europe, a disease would be considered rare if there is one person affected in 2,000 people, while in the USA it would be fewer than one in 200,000 people (Ferreira, 2019). Although these diseases individually have a low prevalence, collectively they affect around 6% of the world population, and it is believed that as many as 80% may have a genetic aetiology (Wakap *et al*., 2020; Frederiksen *et al*., 2022).

Studying rare genetic diseases is important for several reasons. Thanks to advances in genetic technologies and screening initiatives, an increasing number of patients can nowadays receive a genetic diagnosis for their disease. Currently, the estimated number of genetic disorders is between 6 and 7 thousand according to the Online Mendelian Inheritance in Man (OMIM) catalogue with many genes becoming disease associated through research every year and up to 15,000 rare disease-casing genes reported thus far. For example, Charcot-Marie-Tooth (CMT) disease has now over 100 causative genes described in literature (Pisciotta and Shy, 2023).

Importantly, many of the rare diseases are characterised with significant disability, social and financial burden, increased risk for comorbidities and increased mortality (Gahl 2012). Therefore, understanding the genetic causes, could not only benefit patients by leading to diagnosis but also to understanding their disease prognosis and progression, aiding in decisions in family planning, and in some cases could lead to treatments guided by genetics. Indeed, many rare diseases can be now diagnosed at earlier stages which has potential for personalised medicine and slowing of disease progression. One such recent example is hereditary neuropathy caused by recessive mutations in the *SORD* gene, the patients have increased levels of blood sorbitol which can not only act as a biomarker for the disease, but also provide a target for therapeutics (Cortese *et al*., 2020).

Remarkably, at discovery, a disease may appear to be rare, but with research and advancements in medical understanding and testing availability, it may emerge that the disease is more

common and that it had been underdiagnosed. Indeed, this appears to be the case with *RFC1* repeat expansions (Cortese *et al*., 2019) that I will describe in chapters 2 and 3.

In addition, the study of rare diseases caused by a mutation in a single gene, known as monogenic disorder, can provide valuable insights into the normal function of the gene by revealing the consequences of absence of a functional gene or its malfunction. Furthermore, such research may potentially lead to the discovery of pathways involved, identification of new pathways, receptors and other key elements involved.

1.2 Mendelian genetics

Genetic disorders can have various modes of inheritance (Hernandez *et al*., 2016; Zschocke *et al*., 2023) (fig.1.1), and there are five main ones for monogenic disease: dominant – caused by one faulty allele inherited from a parent with 50% chance of inheriting the disease allele, with the disease being usually present in each generation; recessive – caused by inheriting two faulty alleles, one from each parent who are carriers and a chance of having an affected sibling is 25%. X-linked dominant and X-linked recessive diseases are inherited with mutated genes on X chromosomes, and mitochondrial inheritance is caused by mutations in mitochondrial DNA that are inherited from mothers but can affect both male and female offsprings.

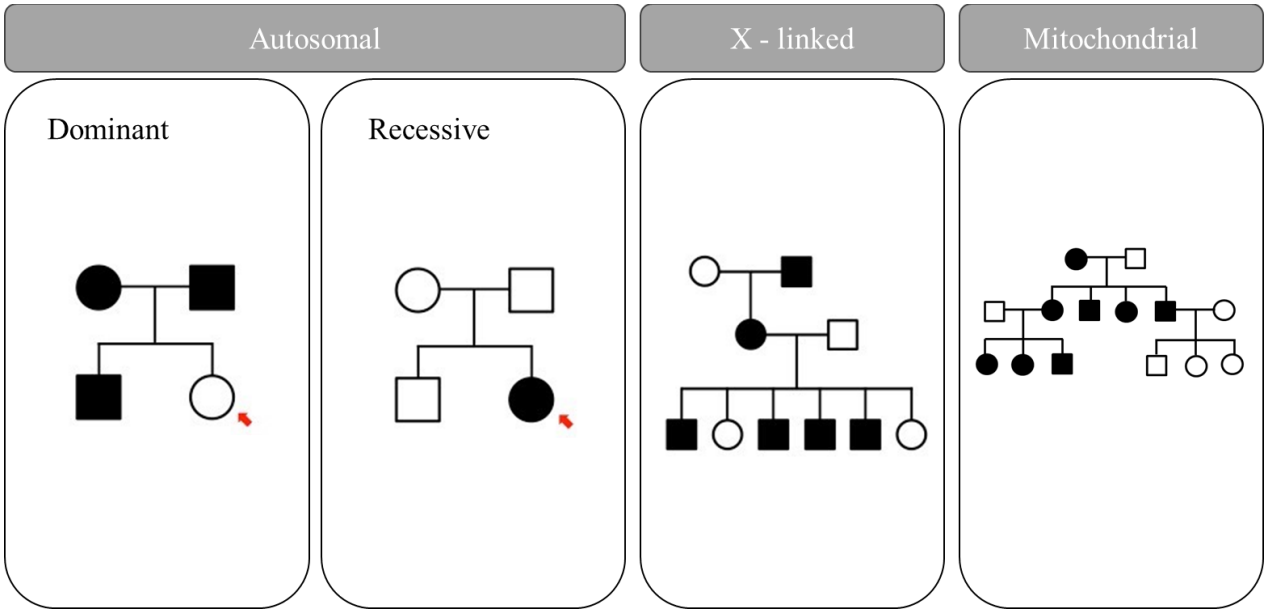


Figure 1.1 Five main modes of inheritance – autosomal dominant and recessive where mutation is passed on to child from parents on autosomes. In autosomal dominant inheritance pattern, a child will be unaffected (red arrow) only if inheriting two healthy alleles, one of each parent. In autosomal recessive inheritance, a child will be affected (red arrow) when inheriting two faulty copies of the gene, one from each parent. In X-linked inheritance, the faulty gene is passed on X chromosome. In mitochondrial inheritance all offspring of affected mothers are affected but no offspring of an affected father is affected.

My thesis encompasses the identification of recessive disease-causing genetic mutations which can be caused by inheritance of the same pathogenic variant on both alleles – homozygous mutation, or two different pathogenic heterozygous variants of the same gene inherited in trans on separate alleles – compound heterozygous.

Homozygous gene mutations aid characterization of the function of human genes because they can lead to disruption of both copies of a gene and can result in phenotypic changes in the affected individuals. Nonfunctional alleles caused by null mutations are very infrequent in general population (Cortese *et al*., 2020), however, the likelihood of finding a homozygous disease-causing mutation is considerably higher in offspring from consanguineous marriages. This is true based on mendelian mode of inheritance whereby two unaffected parents with heterozygous alleles may each pass the recessive disease-causing trait onto the offspring.

1.3 Genetic mutations

Mutations in DNA can either be silent or may result in various consequences (fig.1.2). Single nucleotide polymorphisms (SNPs) are a change of single nucleotide in the sequence which usually have no deleterious consequence on a person's phenotype. However, point mutations, which also involve changes in single nucleotides, can result in alterations leading to missense, nonsense/stop gain mutations. Nonsense mutations introduce premature stop codon leading to truncated versions of the functional protein becoming prone to nonsense mediated decay (Benslimane *et al*., 2024). Missense mutations lead to a change of one amino acid to a different one, depending on the properties of the new amino acid, which in turn can have consequences in protein folding or interactions with enzymes or other proteins (Shinsato *et al*., 2024).

Structural variations (SV) (fig.1.2) in DNA can include insertions or deletions (indels) which can lead to frameshift mutations and therefore different protein sequence downstream or truncated protein (Porubsky and Eichler, 2024). Large deletions can cause excisions of whole exons (Fortunato *et al*., 2023) or chromosome fragments (Mitchel *et al*., 1993).

Repeat expansions are another form of SVs and they arise when a specific nucleotide sequence is repeated beyond the pathogenic threshold (Leitao *et al*., 2024).

When assessing the recessive inheritance of a variant, it is important to be able to phase the variants – variants in "trans" lay on two alleles, whereas variants in "cis" are located on the same allele and therefore are unlikely to cause disease if the disease is recessive.

There are other forms of mutations not mentioned here, however, in my thesis, I investigate repeat expansion mutations in CANVAS and disease-causing point mutations in *ARHGAP19*.

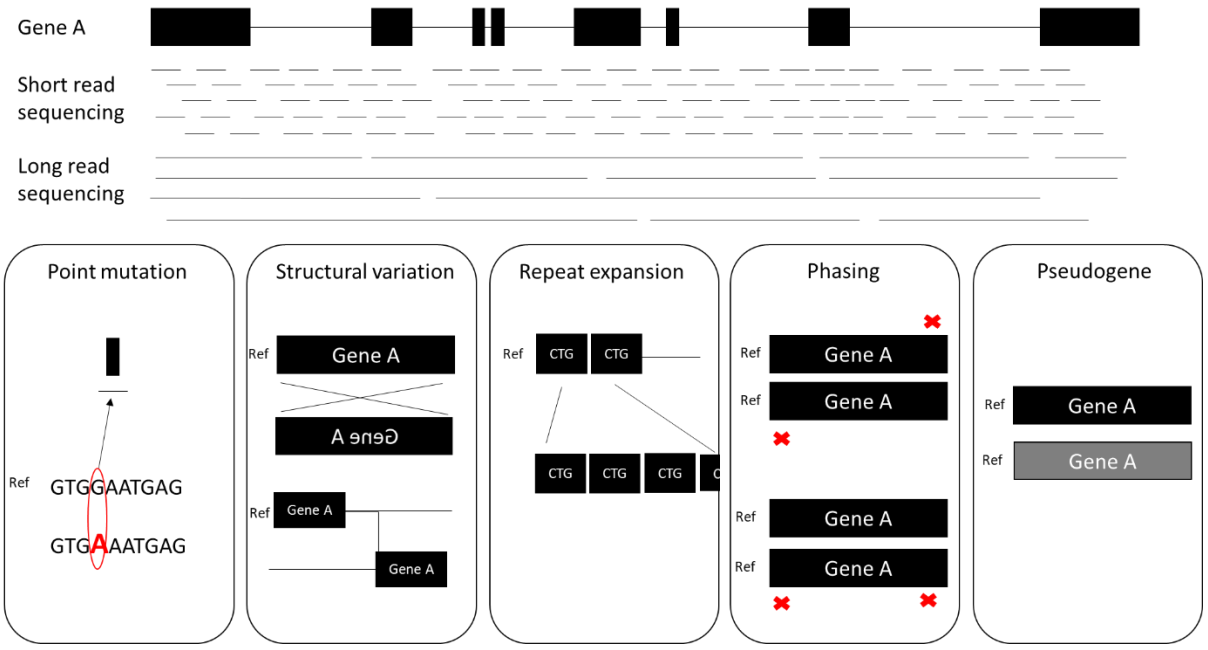


Figure 1.2 Schematic of a coverage of short read sequencing and long read sequencing and genetic variation concepts. Short read sequencing technologies allow for sequencing short fragments of DNA – 100 – 150 bps in length. Long read sequencing allow for sequencing long DNA fragments and therefore are superior in sequencing SVs. Point mutations refer to a change of a single nucleotide, structural variation can cause an inversion of read frame of a gene, gene translocation and other such as repeat expansions where repetitive sequences become expanded. Phasing refers to positioning of a mutation in respect to another allele – mutations in trans are located on two different alleles whereas mutations in cis are located on the same alle. A pseudogene is a DNA segment that resembles a coding gene but cannot code for a protein.

1.4 Genetic methods

A variety of tools are available for detecting mutations in the human genome. Perhaps the most important one that leads to confirmation of many genetic discoveries is Sanger sequencing. Sanger sequencing was developed in late 1970s, and it used gel electrophoresis for detection of chain terminated amplified DNA fragments (Sanger *et al*., 1977). This technique is still used nowadays, we often use it to confirm mutations found in next generation sequencing (NGS) and segregate the identified variant in the proband and its family members.

NGS includes whole genome sequencing (WGS), whole exome sequencing (WES) and gene panels. Both coding and non-coding genome regions are sequenced with WGS, however, this is an expensive technique that demands large bioinformatics skill and processing of data (Efthymiou *et al*., 2016).

WES sequences exons, the coding gene regions only and can detect point mutations in genes and is considerably cheaper than WGS, however, it will not detect structural or intronic variants. Gene panels target specific groups of genes which are associated with specific phenotype or diagnosis.

NGS has many advantages, such as high throughput, where millions of DNA fragments can be simultaneously sequenced, producing large amounts of genomic data. In addition, NGS is considerably faster than traditional sequencing methods therefore it facilitates variant discovery in genomic studies. Perhaps the biggest disadvantage of NGS is that it allows for sequencing of short fragments of DNA, using Illumina HiSeq4000 or Hiseq X platforms, the read length is 100 -150 base pairs depending on paired end protocol used (Kim *et al*., 2018, Hernandez *et al*., 2014). Read length limitations make it challenging to assemble complex genomes or particularly repetitive sequences or large structural variants (Dominik *et al*., 2023; Dolzhenko *et al*., 2024).

Long read sequencing (LRS) is the next step in genome sequencing, the technology used is relatively new and it works to overcome the limitations of NGS, namely, it allows to sequence longer stretches (>Mega base pairs) of DNA, ranging from thousands to mega base pairs in length. The main technologies used in LRS are nanopore sequencing from Oxford Nanopore and single molecule real-time sequencing from Pacific Bioscience (PacBio) long read. Thanks to these technologies, it is now increasingly possible to sequence DNA fragments that span repeat regions, making it possible to infer a repeat expansion motif in repeat expansion disorders, identify larger structural variants in the genome such as insertions and deletions or chromosomal rearrangements or even identify full length transcripts including different isoforms and alternatively spliced regions (Leitao *et al*., 2024). Quite importantly, LRS makes it possible to phase variants where a traditional approach of trio testing – proband and the parents - is not possible to discern which allele was inherited from which parent. Such instances may be true in late onset diseases where the parents are deceased, in non-paternity or adoption cases.

While LRS has many advantages, high-error rates in sequencing and artefacts remain a challenge. In addition, these technologies can be far more expensive than NGS (Mitsuhashi and Matsumoto 2020).

Another new technology is BioNano Genomics optical genome mapping (OGM) which is a non-sequencing platform superior in reading structural variants and large repeat expansions. It

relies on fluorescent labelling of ultra-high molecular length DNA which can then be compared to the reference genome (Facchini*, Dominik* *et al*., 2023).

Those technologies are improving in accuracy and affordability; however, they are still not perfect, they are prone to errors, and artefacts often make analysis difficult. This may potentially lead to misinterpretation of the data. Therefore, a combination of the methods may be used in some cases to achieve most correct read outputs and Sanger sequencing is often used for validation of WES and WGS outputs.

1.5 Interpretation of genetic variants

Interpretation of NGS-derived variants is challenging due to the high volume of returned variants and the phenotypic and genetic heterogeneity of most neurogenetic conditions (Pipis *et al*., 2019). It is important however that variants are classified correctly as they may further contribute to improving clinical management of patients and aid identification of biological mechanisms, functions of genes and possibly targets for treatments.

A standardized system for classifying genetic variants was developed by the American College of Medical Genetics and Genomics (ACMG) (fig.1.3) (Richards *et al*., 2015) and is commonly followed by clinicians and scientist all over the globe to interpret sequencing data.

The first and broadest terms in the ACMG terms are benign and pathogenic and give an indication whether a variant might be associated with human disease or is not disease causing.

The evidence that must be used in this framework to classify variants include population data – how frequent the variant is in the population and its rarity in the control group where the frequency would be less than that of the observed disease; computational predictions with CADD (https://cadd.gs.washington.ed), SIFT (sift.bii.astar.edu.sg), Polyphen-2 (genetics.bwh.harvard.edu/pph2) and Mutation Taster (www.mutationtaster.org) scores are commonly used. Predictions of pathogenicity for missense variants take into account species conservation of amino acids as well as biochemical changes.

For example, when filtering for possible disease-causing variants, any common polymorphisms and synonymous amino acid changes can be initially removed/deprioritised. In addition, it is important to confirm the pathogenicity of variants found through sequencing analysis and carry out functional studies. Functional studies can aid finding out the gene function and mechanism of pathogenicity of mutations. Another important step in the framework is how the disease segregates within family, for which, often Sanger sequencing is used in research laboratories

or trios can be exome or genome sequenced. Importantly, clinical presentation of the patient, the phenotype, must be carefully studied and compared to information that is already available on relevant gene or its pathways.

	Benign		Pathogenic			
	Strong	**Supporting**	**Supporting**	**Moderate**	**Strong**	**Very Strong**
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*			Absent in population databases *PM2*	Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computational evidence suggest no impact on gene /gene product *BP4* Missense in gene where only truncating cause disease *BP1* Silent variant with non predicted splice impact *BP7*	Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3*	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5* Protein length changing variant *PM4*	Same amino acid change as an established pathogenic variant *PS1*	Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1*
Functional Data	Well-established functional studies show no deleterious effect *BS3*		Missense in gene with low rate of benign missense variants and path. missenses common *PP2*	Mutational hot spot or well-studied functional domain without benign variation *PM1*	Well-established functional studies show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		Co-segregation with disease in multiple affected family members *PP1*	Increased segregation data →		
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed in *trans* with a dominant variant *BP2* Observed in *cis* with a pathogenic variant *BP2*		For recessive disorders, detected in *trans* with a pathogenic variant *PM3*		
Other Database		Reputable source w/out shared data = benign *BP6*	Reputable source = pathogenic *PP5*			
Other Data		Found in case with an alternate cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Figure 1.3 American College of Medical Genetics and Genomics variant classification guidelines (Richards *et al*., 2015).

There are however limitations to the ACMG variant classification which can include difficulty in characterizing low penetrance variants, copy number variants or variants in non-coding parts of the gene. A variant of uncertain significance (VUS) is therefore a variant that is difficult to characterise according to the criteria and more investigations are needed. Such a variant may be ultra rare and only present in one patient – private mutation; concern a gene that has not previously been associated with disease or difficult to discern due to limited population diversity.

1.6 Genetic diversity

Over 85% of genetic studies published to date are based on populations with predominantly European ancestry and other populations are under-represented (Wilson *et al*., 2023). Therefore, there is limited knowledge of genetic diversity outside of European populations which undoubtedly results in patients from those populations missing out on genetic diagnosis. Moreover, incidence of consanguineous families is significantly higher in lower to middle income countries and offsprings from those families have higher risk of inheriting recessive, disease causing mutations.

Apart from gene discovery that can be aided by non-European populations, it is an important consideration, that genetic risk factors are poorly transferable between European and non-European populations and genome wide association studies in non-European populations are needed to assess the risk factors associated with neurologic disease (Kamiza *et al*., 2022; El-Boraie el al., 2021).

Additionally, the over-representation of European populations in genetic studies can potentially lead to misunderstanding of genetic variants, susceptibility of genetic disease especially in the context of environmental factors and variable responses to medicines which could have enormous consequence on health of the affected individuals (Pereira *et al*., 2021; D'Angelo 2020)

This highlights that much remains to be learned not only from European populations with rare diseases but also and importantly from populations from non-European ancestries.

1.7 Partnerships and patient recruitment

Synaptopathies and Paroxysmal Syndromes study groups (SYNaPS) was established in 2016 by a Welcome Trust Strategic Award in Houlden lab and other labs at Institute of Neurology (https://www.neurogenetics.co.uk/synaptopathies-synaps-project). Patients referred to National Hospital for Neurology between 2000-2015 that had been diagnosed with paroxysmal neurological disorders with previous consent to blood donation to Neurogenetics Unit-Biobank at the University College London (UCL) Institute of Neurology were also included. Importantly, where possible, families of the probands were also recruited and included affected and unaffected individuals. Researchers and clinicians from around the globe are involved in this initiative and it now includes more than 50 clinical collaborators from more than 30 worldwide countries (fig.1.4).

To date, around 30,000 individuals with neurological and neurodevelopmental disorders have been collected through SYNaPS. Clinical information including medical history with investigations such as nerve conduction studies (NCS), EEG, MRI and family history, phenotype and neurological information of each patient has been collected and deposited in UCL secure drives. All individuals in the study consented to be involved in research (ethics number UCL 07/Q0512/26), and their specimens were collected for genomic DNA extraction – mainly blood and/or saliva and skin biopsies.

International Centre for Genomic Medicine in Neuromuscular Diseases (ICGNMD) was established in 2019 by MRC strategic award (https://www.ucl.ac.uk/centre-for-neuromuscular-diseases/research/international-centre-genomic-medicine-neuromuscular-diseases).

ICGNMD has partners in Brazil, South Africa, Zambia, India, Turkey, Netherlands and UK (fig.1.4) and the initiative aims to build a diverse worldwide cohort of patients with neuromuscular diseases and over 5000 probands with families have been recruited in full compliance with local ethics and legislations. The patients were deeply phenotyped and all information stored in RedCAP in standardised form including positive and negative Human Phenotype Ontology (HPO) terms, sex, disease affection, age at onset and diagnosis, diagnostic category, clinical assessment scales used by all clinical fellows on the project and summarized genetic data and other relevant information (Wilson *et al*., 2023).

Figure 1.4 A world map showing regions where SYNAPS and ICGNMD partners are located. Drawn using Biorender.

These collaborations and diverse cohorts of participants allowed us for discovery of *ARHGAP19* which I will describe in chapter 4.

In addition to SYNAPS and ICGNMD, patients with clinical diagnosis of sensory neuropathy (Sensory neuropathy was diagnosed according to clinical and neurophysiological criteria), ataxia or suspected CANVAS have been collected at National Hospital for Neurology and Neurosurgery (NHNN), at Pavia University in Italy and samples were also sent from external collaborators for screening of the repeat expansions in *RFC1* gene. Patients were consented for research at the participating institutions and genomic DNA, or blood specimens were collected.

1.8 Raise of genetic consortiums and gene depositories

Owing to advances in technology, falling cost of sequencing and the pioneering work of Human Genome Project that was completed in 2003 and fully sequenced about 92% of human genome for the first time (Green and Donohue, 2018), many consortia and gene depositories were established. In my thesis, apart from invaluable collaborations mentioned in sections above, I benefited from use of Genomics England's 100,000 Genome Project and Gene Matcher.

1.8.1 100,000 Genome Project

100,000 Genome Project was announced in 2012. Led by Genomics England, it aimed to recruit 100 thousand patients with rare diseases and cancer and fully sequenced their genomes using short read WGS to "make genomics part of routine healthcare, enhance genomic healthcare research and uncover answers for participants" (https://www.genomicsengland.co.uk/initiatives/100000-genomes-project). Up to 2018, 25% of affected probands received a genetic diagnosis. Importantly, the genetic data is available to researchers, and in my thesis the project is used for both CANVAS and ARHGAP19 disease.

1.8.2 Gene Matcher

Gene Matcher is a freely available gene depository, and it enables researchers and clinicians to connect about the gene(s) they are interested in (https://genematcher.org/; Sobreira *et al*., 2015). This is a truly invaluable tool, especially in the field of rare diseases. It allows for finding particular genes and their mutations all over the world and connecting with persons who submitted the gene, therefore aiding building cohort studies and gathering evidence on the role of specific genes in human diseases.

1.9 Neuropathies and ataxias

My thesis encompasses two disease groups – neuropathies and ataxias. These are both very heterogeneous disorders, and it is not uncommon for them to appear concurrently in a patient as a part of more complex syndrome.

1.9.1 Neuropathies

Neuropathies are disorders where peripheral nerves become damaged which can affect movement, sensation and even organ function. Neuropathies may affect single nerve, many nerves in the same area or nerves in different areas. Neuropathies are genetically heterogeneous (Nam *et al*., 2016).

Charcot-Marie-Tooth (CMT) disease also called hereditary motor and sensory neuropathy (HMSN) is the most prevalent mendelian inherited neuropathy (Record *et al*., 2024) and indeed, inherited peripheral neuropathies are amongst most commonly inherited neurologic diseases. The prevalence of CMT varies amongst populations but is estimated at around 1 in 2,500 individuals (Pisciotta and Shy, 2018).

Patients with CMT can range from mildly affected to severely disabled and the disease presents with progressive weakening and atrophy of muscles, especially in distal limbs. Often foot

abnormalities such as pes cavus or hammer toes may be associated with the disease. It is not uncommon that the patients do not suffer from any pain or sensory symptoms. Family history, nerve conduction studies and thorough clinical evaluation can aid differential diagnosis (Klein, 2020).

Historically, Charcot-Marie-Tooth and Dejerine-Sottas described neuropathies according to disease phenotype (fig.1.5). Advances in technology in early nineties, allowed for discovery of major genes causing CMT. CMT is classified according to nerve conduction studies in the upper limbs, mode of inheritance and phenotype into dominantly inherited CMT type 1, a demyelinating neuropathy and CMT type 2, an axonal neuropathy. CMTX has an X linked pattern of inheritance and CMT4 is recessive (Morena *et al*., 2019). An additional group includes intermediate CMT (CMTi) with nerve conduction velocities in between values of demyelinating and axonal (Matilde *et al*., 2019).

Sporadic, de novo variants have also been identified in genes causing neuropathies, for example *HSP27* in CMT2 (Houlden *et al*., 2008), *EGR2* in Dejerine-Sottas Neuropathy (Grosz *et al*., 2019) or *SLC12A6* in early onset sensorimotor neuropathy (Grosz *et al*., 2019). Moreover, it was estimated that de novo variants may account for 10% of CMT1 cases (Blair *et al*., 1996). Not only can CMT have various inheritance patterns, but also, with advances of sequencing technologies over the past two decades, a multitude of genes have been associated with the disease which highlights it genetic heterogeneity. I will overview the major genes associated with the disease and further highlight the gene discovery advances in CMT.

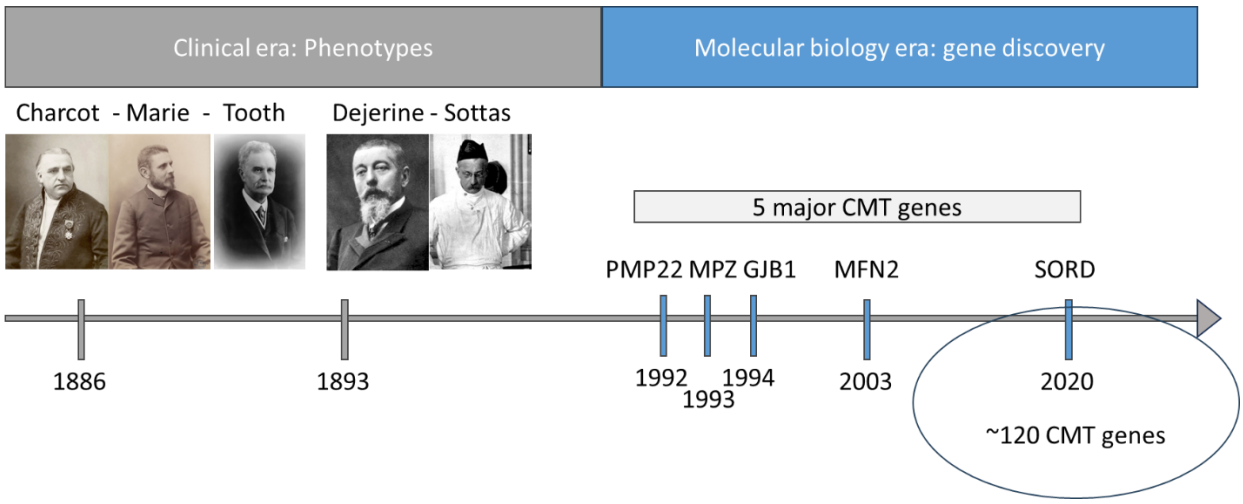


Figure 1.5 A timeline of CMT disease – from phenotypic descriptions of late 19th century to discovery of 5 major CMT associated genes and further discovery of over 100 CMT genes in recent years.

1.9.1.1 CMT1A

CMT1A is a demyelinating form of CMT, and it accounts for about 80% of all CMT cases therefore being the most common CMT. It is a dominantly inherited disease caused by duplication in chromosome 17 at location which contains peripheral myelin protein 22 (*PMP22*) gene. *PMP22* has an important role in synthesis and maintenance of myelin and is expressed in myelinating Schwann cells (Pisciotta and Shy 2018).

Patients with CMT1A usually have walking difficulties, weakness in distal limbs and wasting, foot deformities and sensory loss. The age of onset of their disease is typically in the first decades of life and the disease has a slow progression (Pisciotta and Shy 2023).

Genetic testing for *PMP22* is usually the first genetic investigation for patients with demyelinating CMT. Whilst there is no cure for CMT1A yet, since the disease is caused by duplication of *PMP22*, therapeutic strategies aiming at reducing PMP22 expression are being investigated (Pipis *et al*., 2019).

1.9.1.2 CMT1B

CMT1B is a demyelinating form of CMT, and it accounts for about 5% of all CMT cases (Pisciotta and Shy 2023). It is a dominantly inherited disease caused by mutations in myelin protein zero (*MPZ*) gene which is involved in the formation and maintaining stability and homeostasis of myelin in peripheral nerve (Shy *et al*., 2004). There are over 200 different mutations identified in MPZ and the arising phenotypes are characterised by different nerve pathology (axonal vs demyelinating) and age of onset (early vs late onset) (Pisciotta and Shy 2023).

There is no cure for CMT1B and due to the genetic heterogeneity of CMT1B mutations the therapeutic approaches may need to address the protein function rather than gene dosage as in the case of CMT1A.

1.9.1.3 CMTX1

CMTX1 is an X-linked CMT and is caused by mutations in gap junction protein beta 1 (GJB1) which forms gap junctions between myelin sheaths of Schwann cells. In this disease, males usually present with more severe phenotype and the first neurological symptoms usually occur in childhood. Females have a milder neuropathy, because of variability in X chromosome inactivation (Tomaselli *et al*., 2017; Panosyan *et al*., 2017).

1.9.1.4 CMT2A

CMT2 is an axonal form of CMT, and CMT2A contributes to 3.9-4.0% of genetically confirmed CMT2 (Pisciotta and Shy 2023; Cortese *et al*., 2020; Record *et al*., 2024). It is caused by mutations in mitofusin 2 (*MFN2*), which is a mitochondrial transmembrane GTPase protein that plays an important role in fusion and fission of mitochondria.

Patients with CMT2A typically have an early age of onset in infancy or early childhood and a severe progressive phenotype. The predominantly motor involvement results in patients requiring walking aids in early childhood and being wheelchair dependant by the age of 20 (Feely *et al*., 2011).

1.9.1.5 SORD

Mutations in sorbitol dehydrogenase (*SORD*) have recently been identified to cause recessive axonal distal hereditary motor neuropathy (dHMN) and CMT2 and they account for up to 10% of axonal cases. These mutations cause reduced level of SORD enzyme which results in accumulation of sorbitol in blood and tissues, and it may cause toxicity to peripheral nerves (Cortese *et al*., 2020).

Patients with SORD mutations suffer with slowly progressive length-dependent axonal neuropathy and it is not uncommon for them to suffer from sensory symptoms although these are milder than motor symptoms.

This is a potentially treatable neuropathy, and a clinical trial is currently ongoing to decrease the levels of blood sorbitol in patients.

1.9.2 Rise of next generation sequencing

Recently, in the diagnostic settings, the CMT diagnostic approach have been to exclude chromosome 17p duplication by Multiplex ligation-dependent probe amplification (MLPA) and follow with targeted NGS CMT gene panels or more recently virtual panels on short read WGS (Pipis *et al*., 2019). In phenotype-specific panels, which screen only for genes relevant to the patient's phenotype, the interpretation process is streamlined as irrelevant variants are excluded. However, this approach may not account for genetic heterogeneity of CMT and related genetic disorders which may have a significant phenotypic overlap; or where neuropathy is a part of more complex disorder.

In research settings, whole exome and genome sequencing can be used in patients where targeted NGS panels have not yielded a result and apart from improved diagnostic rate, WES

and WGS are imperative for identification of novel disease-causing genes. Owing to advances in next generation sequencing, over 100 CMT causative genes have been described in the past decade and undoubtedly, many more will be discovered (fig.1.6). Indeed, as part of my thesis and a large international collaboration, we add *ARHGAP19* to the rapidly expanding list, and I will talk about the gene in chapter 4.

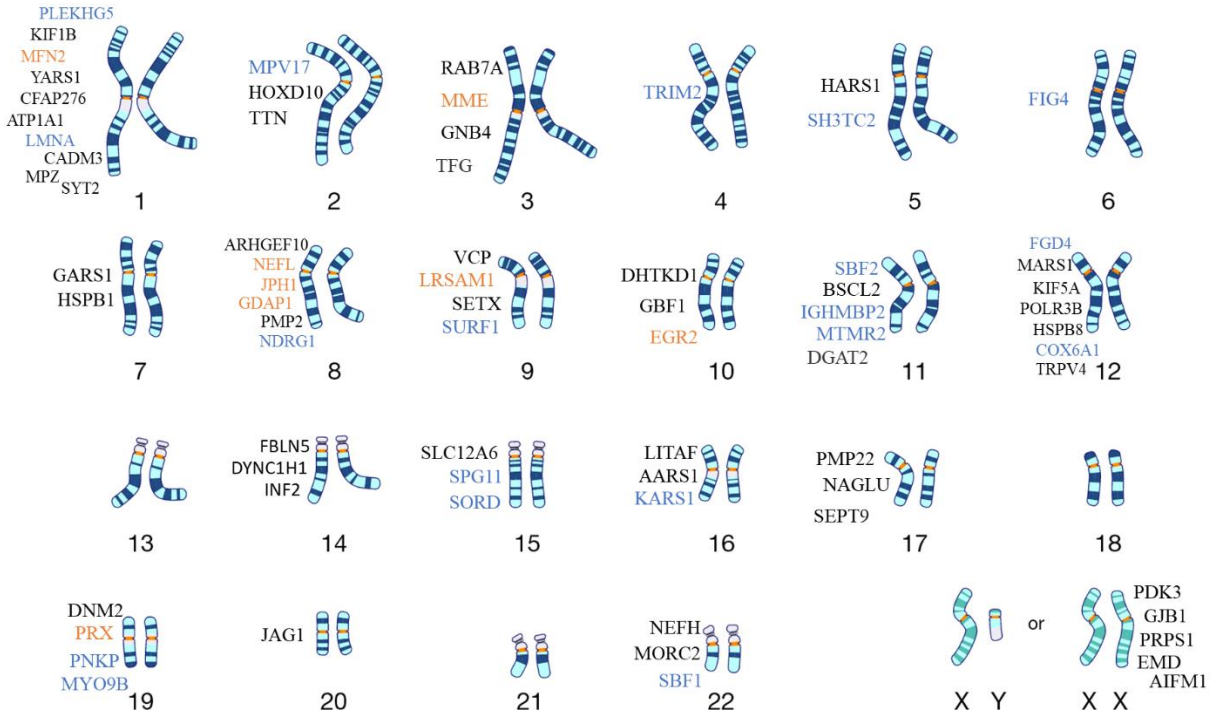


Figure 1.6 Complete set of human chromosomes with known CMT associated genes as listed in OMIM (https://omim.org/). The genes are listed at each specific chromosome they are located in, in black genes with autosomal dominant inheritance, in blue – autosomal recessive inheritance and in orange are genes that can have both recessive and dominant inheritance pattern. Karyotype schematic was adapted from National Human Genome Research Institute.

1.9.3 CMT and proposed pathomechanisms

1.9.3.1 Neuron

Neuron is a highly specialised cell that is electrically excitable and able to send electric signals in form of action potentials across itself and further through the neuronal network. This signal conduction is accelerated by the myelin sheath and the nodes of Ranvier that allow the signal to jump rather than travel in a straight line. Additionally, myelin serves as a protector to the axon – the longest part of the nerve cell. Other important parts of neuron anatomy are the cell body which extends to the dendrites. On the other side of the axon, lie the synapses, specialised connections that make neurotransmitter signalling possible. Neurons can be classified into sensory neurons which respond to sensory stimuli such as touch, sound and other, motor

neurons which control movement and interneurons which connect subsets of neurons in the same area.

1.9.3.2 Disease mechanisms

CMT can arise due to dysfunctions of various parts of the nerve and the pathomechanisms can involve the cell body (eg. *GDAP1*, *MFN2*), the axon (eg. *DCTN1*) or the myelin (*PMP22* and *GJB1*). Different faulty genes can cause damage through varied pathways such as axonal transport, myelination and signal transduction, mitochondrial stress, protein aggregation and numerous others (Estevez-Arias *et al*., 2022). Some genes and the pathogenic pathways they are implicated in to cause CMT are highlighted in fig.1.7.

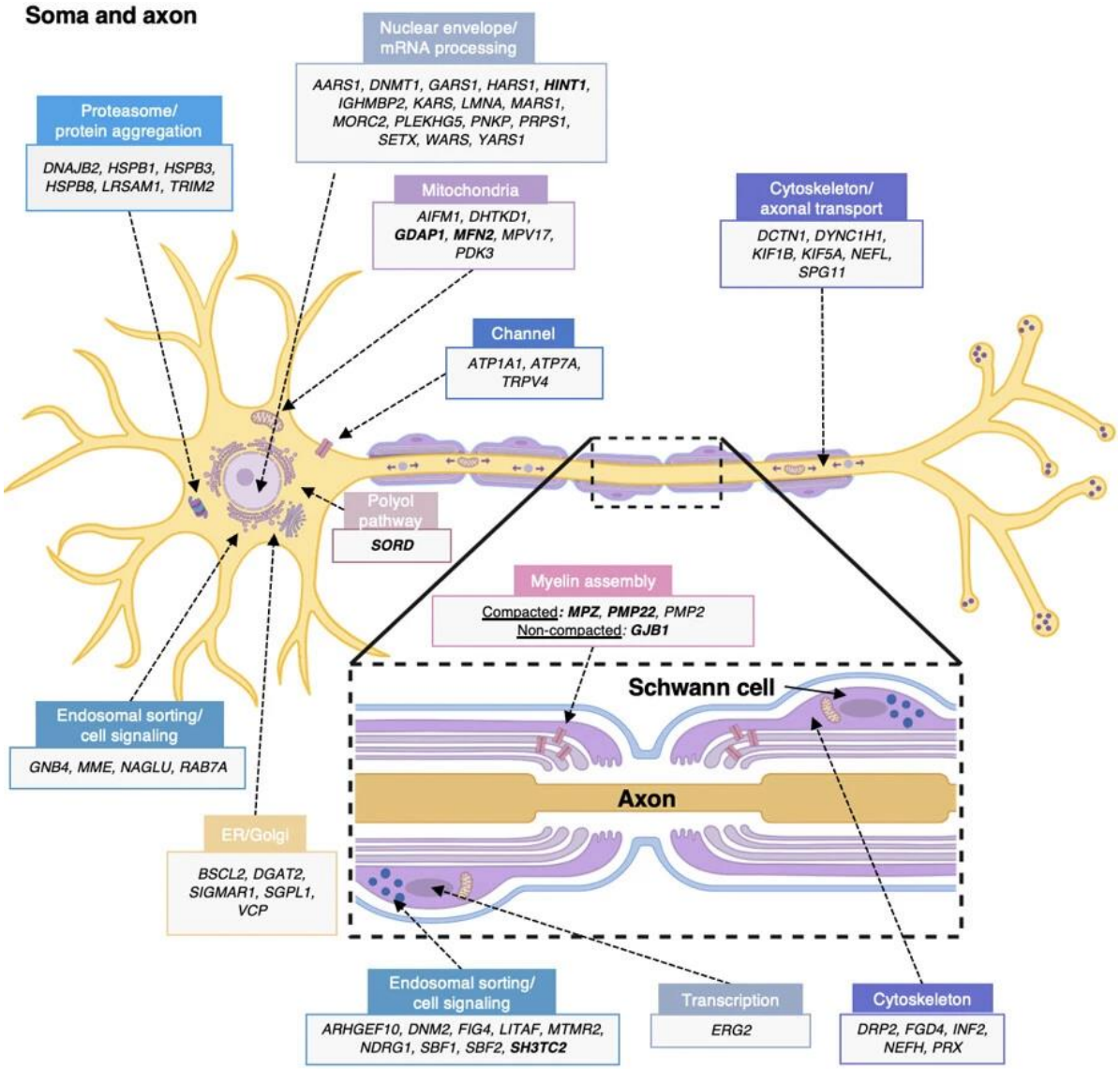


Figure 1.7 CMT genes and the nerve cell area proposed to act on. Most common genes are in bold. Figure from Esteves-Arias *et al*., 2022.

1.9.3.3 Therapies in CMT

CMTs are challenging to treat with no drug therapy available for most of the patients apart from physical therapy, rehabilitation and symptomatic treatment. Proposed treatments, some undergoing clinical trials, are based on gene dosage as in *PMP22*, regulation of myelin thickness in demyelinating neuropathies, or correction of lipid synthesis to name just a few (Pisciotta and Shy 2023). One of the most recently discovered CMTs, caused by mutations in *SORD*, is a potentially treatable neuropathy with defects in sorbitol pathway where sorbitol levels become elevated. It is also one of the most easily diagnosable CMTs, as sorbitol can act as a biomarker in the disease (Cortese *et al*., 2020). Clinical trials for SORD CMT are currently ongoing which focus on targeted inhibition of aldose reductase enzyme which converts glucose to sorbitol.

The candidate therapies for CMT and any other neurological disease are based on research into disease mechanisms, affected pathways and the consequence of their dysregulation. This is a very important point in the field of neurogenetics, as today's gene discoveries, may potentially lead to tomorrow's therapies.

1.9.4 Ataxias

Ataxias are a group of disorders that affect balance, movement, speech and vision. They can arise due to dysfunction of cerebellum – cerebellar ataxia; vestibular system – vestibular ataxia; and various parts of brain, spinal cord and sensory nerves can be implicated in sensory ataxia. Ataxias can be sporadic or hereditary and they show considerable genetic heterogeneity (Sun *et al*., 2019).

The advances in genomics field make possible discoveries in genetics of ataxias and to date up to 300 genes with proposed pathogenic variants have been described. However, there is still a considerable genetic diagnosis gap in hereditary ataxia as up to 75% patients lack genetic diagnosis (Chen *et al*., 2023).

1.9.4.1 Cerebellar ataxias

Cerebellar ataxias can be classified based on the mode of inheritance, and they include autosomal recessive cerebellar ataxias (ARCAs), autosomal dominant spinocerebellar ataxias (SCAs), episodic ataxias and X-linked ataxias (Manto *et al*., 2020).

Notably, there is an overlap of patient clinical presentations between different ataxia subtypes and often patient genetic material is tested concurrently for most common ataxia genes.

My interest lies in recessive inheritance, and I outline chosen ARCAS below and briefly mention SCAs.

1.9.4.2 Autosomal recessive cerebellar ataxias

Recessively inherited ataxias show a global prevalence of 3 in 100000 and they commonly present sporadically (Ruano *et al.*, 2014; Traschutz *et al.*, 2023).

1.9.4.2.1 Friedreich's ataxia

Friedreich's ataxia (FRDA) is the most common autosomal recessive hereditary ataxia with prevalence of about 1 in 30000 in central Europe (Vankan 2013) caused by biallelic repeat expansions of GAA trinucleotide in the first exon of frataxin encoding gene which leads to transcriptional deficiency of the gene. The normal expansion range for GAA FRDA is 14-34 triplet repeats and alleles with expansions larger than 90 repeats are considered pathogenic. Approximately 4% of patients with FRDA are compound heterozygous for expansion on one allele and a missense mutation on the other allele (Delatycki and Bidichandani 2019).

FRDA is characterised by an early onset progressive cerebellar ataxia, dysarthria, areflexia, loss of position sense and axonal neuropathy causing motor weakness. Late onset ataxia shows milder phenotype. Investigations for FRDA treatment included a number of trials to increase frataxin levels, therapies for gene and protein replacement, antioxidants and also as inflammation has been implicated in FRDA pathogenesis, modulation of inflammation has been trailed (Delatycki and Bidichandani 2019). Very recently, a new drug, Omaveloxolone, has been approved for the treatment of Friedreich's ataxia in adults in the USA and EU countries (https://www.ataxia.org.uk/omav-updates/).

1.9.4.2.2 Autosomal recessive spastic ataxia of Charlevoix-Saguenay

Autosomal recessive spastic ataxia of Charlevoix-Saguenay (ARSACS) is caused by biallelic mutations in SACS gene. The prevalence of ARSACS had been thought to be rare outside of Quebec Canada where it is estimated at 1 in 484 and is caused by founder effect due to French settlement, however, numerous studies had described patients outside of the region (Engert *et al.*, 2000).

This is a young onset disease and the first symptom at onset is unsteady walking gait which commonly begins when affected toddler are learning to walk. The patients have demyelinating neuropathy, progressive spasticity and cerebellar ataxia, and they will require walking aid or wheelchair assistance in adulthood.

There is currently no treatment for ARSACS, the pre-clinical studies to date have been focusing on enhancing mitochondrial transport as the mutated protein sacsin has downstream effect on disturbing mitochondrial fission.

1.9.4.2.3 Autosomal recessive spectrin repeat-containing nuclear envelope protein 1 ataxia

Autosomal recessive spectrin repeat-containing nuclear envelope protein 1 ataxia (*SYNE1*) can be caused by homozygous or compound heterozygous mutations in *SYNE1*. It presents with pure cerebellar ataxia in 20% of the patients while the majority of the patients have complex ataxia phenotypes with other neurologic and non-neurologic dysfunctions (Synofik and Nemeth 2018).

SYNE1 is one of the largest genes in the human genome and the protein in a structural protein implicated in formation of large, assembled complexes that are implicated in nuclear migration and anchoring to actin cytoskeleton. It is unknown whether this role of the protein contributes to the pathogenesis of ataxia and there are currently no treatments for SYNE1 ataxia.

1.9.4.2.4 Other recessive ataxias

A variety of genes and molecular pathways are found causative of cerebellar ataxias, and the list is likely to grow owing to discovery of repeat expansions and structural variations through long read sequencing and novel technologies.

Table 1.1 shows some of the most prevalent recessive ataxias with their corresponding known genes.

Disease	**Abbr.**	**Gene**	**Protein**
Friedreich Ataxia	FRDA	*FXN*	Frataxin
Ataxia telangiectasia	AT	*ATM*	Serine protein kinase
Ataxia with oculomotor apraxia type 1	AOA1	*APTX*	Aprataxin
Ataxia with oculomotor apraxia type 2	AOA2	*SETX*	Senataxin
autosomal recessive spastic cerebellar ataxia of Charlevoix-Saguenay	ARSACS	*SACS*	Sacsin
Sensory ataxic neuropathy, dysarthria, and	MIRAS/ SANDO	*POLG1*	DNA Polymerase subunit γ-1

ophthalmoparesis/mitochondrial recessive ataxia syndrome)			
Autosomal recessive cerebellar ataxia type 1	ARCA1	*SYNE1*	Nesprin-1
Spastic paraplegia type 7	HSP-SPG7	*SPG7*	Paraplegin
Autosomal recessive cerebellar ataxia type 2 with coenxyme Q10 deficiency	ARCA2	*CABC1/COQ8A*	Chaperone-activity of bc1 complex-like/Coenzyme Q8A
Autosomal recessive cerebellar ataxia type 3 caused by mutations in ANO10	ARCA3	*ANO10*	Anoctamin-10
Ataxia with vitamin E deficiency	AVED	*TPPA*	A-tocopherol transfer protein
Cerebrotendinous Xanthomatosis	CTX	CYP27A1	CYP27 Sterol 27-hydroxylase
Marinesco-Sjogren syndrome	MSS	*SIL1*	Nucleotide exchange factor SIL1
Infantile onset spinocerebellar ataxia (IOSCA)	IOSCA	*C10orf2*	Twinkle

Table 1.1 A list of the most common recessive ataxias with associated genes and protein products (Beaudin *et al*., 2019).

1.9.4.3 Spinocerebellar ataxias

The list of ataxias mentioned is by no means exhaustive as many other hereditary ataxias have been described. It is important to briefly mention spinocerebellar ataxia (SCA) which encompasses a large subset of ataxias. SCA is an autosomal dominant disease, often caused by repeat expansions and more than 40 genetic SCAs have been identified and they are classified according to the genetic loci with SCA1 first identified. The global prevalence of SCAs is 1-5 in 100000 individuals depending on geographical location (Moraes, *et al*., 2023). There is much research on SCAs, and they are commonly tested in diagnostic laboratories. Table 1.2 list currently known genes implicated in pathogenesis of specific SCAs.

The most recently discovered cause of Spinocerebellar ataxia 27B (SCA27B) is a triplet repeat expansion of GAA nucleotides in FGF14 encoding fibroblast growth factor 14 (Pellerin *et al*., 2023). This is a dominantly inherited disease causing late-onset cerebellar ataxia with pathogenic repeat size of at least 250 GAA repeats.

Disease Subtype	**Gene/Locus**	**Mutation type**
SCA1	*ATXN1*	*CAG repeat expansion*
SCA2	*ATXN2*	*CAG repeat expansion*
SCA3	*ATXN3*	*CAG repeat expansion*
SCA4	*ZFHX3*	*GGC repeat expansion*
SCA5	*SPTBN2*	*Point mutations*
SCA6	*CACNA1A*	*CAG repeat expansion*
SCA7	*ATXN7*	*CAG repeat expansion*
SCA8	*ATXN8*	*CAG repeat expansion*
SCA9	Not assigned	
SCA10	*ATXN10*	*ATTCT repeat expansions*
SCA11	*TTBK2*	*Point mutations*
SCA12	*PPP2R2B*	*CAG repeat expansion*
SCA13	*KCNC3*	*Point mutations*
SCA14	*PRKCG*	*Point mutations*
SCA15	*ITPR1*	*Point mutations*
SCA16	*ITPR1*	*Point mutations*

SCA17	*TBP*	*CAG/CAA repeat expansion*
SCA18	Not assigned	
SCA19	*KCND3*	*Point mutations*
SCA20	11q12	
SCA21	*TMEM240*	*Point mutations*
SCA22	*KCND3*	*Point mutations*
SCA23	*PDYN*	*Point mutations*
SCA25	*PNTP1*	*Point mutations*
SCA26	*EEF2*	*Point mutations*
SCA27a	*FGF14*	*GAA repeat expansion*
SCA28	*AFG3L2*	*Point mutations*
SCA29	*ITPR1*	*Point mutations*
SCA30	4q34.3-q35.1	
SCA31	*BEAN1*	*TGGAA repeat expansion*
SCA34	*ELOVL4*	*Point mutations*
SCA35	*TGM6*	*Point mutations*
SCA36	*NOP56*	GGCCTG repeat expansion
SCA37	*DAB1*	*TTTCA repeat expansion*
SCA38	*ELOVL5*	*Point mutations*
SCA40	*CCDC88C*	*Point mutations*
DRPLA	*ATN1*	*CAG repeat expansion*
SCA42	*CACNA1G*	*Point mutations*
ADCADN	*DNTM1*	*Point mutations*

Table 1.2. A list of Spinocerebellar Ataxias with the associated genes and mutation type (repeat expansions or point mutations). Adapted from OMIM (https://omim.org/).

1.9.5 Cerebellar ataxia and proposed pathomechanism

1.9.5.1 Cerebellum

Cerebellum is the largest part of the hindbrain and is located posterior to the brain. It has major roles in movement and balance control, allows for gait coordination and posture maintenance as well as voluntary muscle activity. Cerebellum has a very high neuronal content, with approximately 80% of all brain neurons located in the organ (Roostaei *et al*., 2014).

1.9.5.2 Disease mechanisms

A variety of pathomechanisms have been implicated in recessive ataxias and they may include defects in energy production, DNA repair or oxidative damage. Importantly, cerebellar ataxias cause dysfunction of cerebellum, and Purkinje cells, neurons specific to cerebellum, are one of the largest neurons in the nervous system. This means they have high energy requirement for proper functioning and may become vulnerable due to energy deficiency which can be potentially exacerbated if mitochondria are under oxidative stress. Further DNA may be damaged due to reactive oxygen species (ROS) produced in faulty mitochondria and additional defects in DNA repair may lead to more negative consequences (Beaudin *et al*., 2022).

It appears that cerebellum is particularly susceptible to these processes and some known cerebellar ataxia genes are highlighted in fig.1.8.

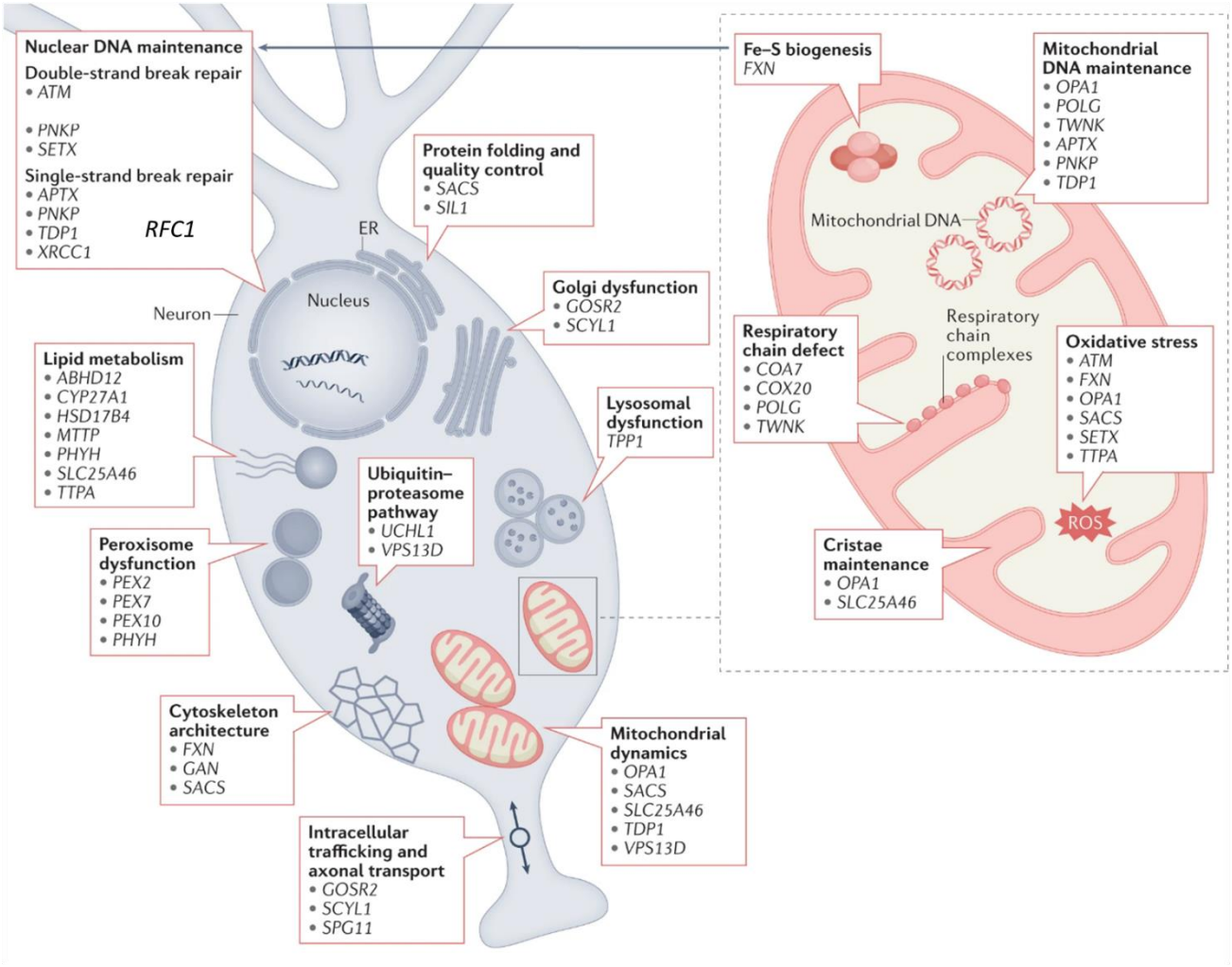


Figure 1.8 Ataxia associated genes with proposed pathways. Figure adapted from Beaudin *et al*., 2022.

1.9.5.3 Therapies

There are no cures for most of cerebellar ataxias and current therapeutic avenues for cerebellar ataxias are largely based on rehabilitation and vitamin supplementation as for example in ataxia with vitamin E deficiency.

However, importantly, a new and first treatment for FRDA has been approved. Omaveloxone, is a drug that activates erythroid 2-related factor 2 (Nrf2) which signalling is suppressed in FRDA. Treatment with the drug in the clinical trials significantly improved neurological function in the patients and patients saw improvement in specific clinical tasks. Notably, FRDA is a progressive disease thus improvement in the patient symptoms is very relevant (Kessler, Sharma and Lynch 2023).

1.10 Animal models in neuroscience research

One of the most important aspects in neurogenetics research is the characterisation of gene function within the human nervous system. Animal models of neurogenetic diseases are an important tool in enquiring about the function of a gene and they have enhanced our understanding of not only pathogenesis of many neurologic diseases but also normal human biology.

Different models can be used depending on variables such as phenotype, disease onset and duration, gene conservation between species and availability of a gene ortholog. However, other considerations must be taken into account, and they include ethical implications, correct facility and staff, investigator experience as well as affordability and time.

By far the most commonly used animal models in neuroscience include mice, rats, zebrafish, worm and fruit fly and they all share similar advantages:

- Controllability and standardisation: animal studies are conducted in pre-specified environmental conditions, following specific diets and controlled exposure to stimuli or drugs. This allows for investigating a chosen variable and its effect on biological processes of phenotypes of the models. Once described, the conditions can be used in replication experiments. This level of controllability and standardisation would not be possible in human studies and in many cases would be considered unethical.
- Reproducibility: due to use of standardised conditions, animal studies can be replicated in facilities all over the world which offers validation of experimental results as well as ability to collaborate.
- Manipulation and ethical considerations: some experimental manipulations and procedures can be conducted on animals that would be considered unethical on a human subject. Moreover, valuable preliminary data can be obtained by using animal models which can then justify further in-vivo research in humans for example by helping to first minimise risks.
- Translation to human health: preliminary and pre-clinical studies in animals can contribute to development of new treatments, adjusting dosage of medications or screening for toxicity of drugs and medical compounds.
- Longitudinal studies – animal models can be observed throughout their lifespan, and much can be learned from progression of their disease or response to treatment. Post-mortem tissue is also then available for further functional research.

Despite these advantages, all animal models have some limitations that may include

- Species differences: despite genetic and phenotypic similarities, there are differences between animals and humans. These can range from size – for example human is a large organism whose nerves are long and can exceed one meter in length (sciatic nerve), whereas mice are tens of times smaller – to different biological pathways
- Complexity: human genetics is complex and inherited disease can be multifactorial and animal models may not allow to fully replicate the genetic and phenotypic variation seen in humans
- Gene orthologs: not all genes expressed in humans can be found in a corresponding animal model
- Lifespan: whilst an animal can be observed throughout its life, it is an important consideration that the lifespan may not translate to human disease. Some late onset disease may not be modelled successfully in an animal due to its shorter lifespan and potentially not developing symptoms. In addition, if a disease develops over years in humans, it may not be accurately modelled in an animal which only lives for two years.
- Environmental factors: the environment where animal studies are conducted is highly controlled and may not translate to living conditions of a human subject
- Drugs response: there may be a different response to a drug in an animal model than in human due to differences in metabolism and other factors.
- Genetic manipulation: genetic manipulation such as gene knockdown can lead to unintended effects affecting the validity of a model

1.10.1 *Drosophila melanogaster* in studies of neurological disease

Drosophila melanogaster is an invaluable model in studying neurological disease. Due to its short life cycle of about 10 days from embryo to adult fly but also the availability of large number of animals that can be used in an experiment, as well as unparalleled genetic tractability, it allows for rapid phenotypic screening of various disease modelling mutant flies, including knock down or out flies or knock in flies.

Notable *Drosophila melanogaster* studies in recent years include elucidating gain of function of toxicity in *C9orf72* repeat expansion disorders (Sharpe *et al*., 2021), toxicity of CGG repeat expansions of *NOTCH2NLC* which included mitochondrial swelling in the fly model (Yu *et*

al., 2022) and elucidation of pathomechanisms in Frederich ataxia (Calap-Quintana 2018) to name just a few.

Notable *Drosophila melanogaster* models of CMT disease include *GDAP1* knock down and overexpression both of which caused mitochondrial dysfunction (Del-Amo *et al*., 2017), *KIF1A* knockout causing axonal transport disruption (Kern *et al*., 2013) and *SORD* orthologs neuronal specific knockdowns resulted in age-dependant locomotion deficits (Cortese *et al*., 2020).

However, and importantly, *Drosophila* axons do not have myelin sheath or Schwann cells therefore modelling demyelinating conditions is not possible with this model.

In my research, I use *Drosophila melanogaster* to study both *RFC1* and *ARHGAP19* knockdowns and I describe those models in chapter 5. In addition, *ARHGAP19* fly model is complemented by a collaboration with external laboratories who model the gene in *Danio rerio,* and this work is included in Chapter 4.

1.11 Thesis Aims

This thesis aims to use genetic and functional tools to gain increased understanding of ataxias and neuropathies focusing on Cerebellar Ataxia with Vestibular Areflexia syndrome (CANVAS) and a novel neuropathy gene, *ARHGAP19*. I will discuss researching correlation of repeat expansion size and disease onset and severity in RFC1 repeat expansion disorder and exploration of genetic heterogeneity in RFC1 disease. Furthermore, I will describe the discovery of rare biallelic variants in *ARHGAP19* that cause CMT neuropathy and further describe the functional studies undertaken to elucidate the role of *ARHGAP19*. Finally, I will talk about the use of *Drosophila melanogaster* as an animal model in neurogenetic disorders and how I explored loss-of-function mechanisms in both *RFC1* and *ARHGAP19* using this model.

1.11.1 Chapter 2 Aims

Primary aims of my work in this chapter were to contribute to a large multi-centre effort lead by Dr Andrea Cortese of screening and research diagnosing patients with RFC1 disease spectrum and subsequently to measure their allele sizes with Southern blotting. We extrapolated the data to explore the relationship of size of the AAGGG expansions in *RFC1* locus with the age of onset of the disease, the disease progression, and its severity; and by testing 27 families with multiple members, we explored anticipation in *RFC1* disease. A large part of this work involved optimizing and troubleshooting the Southern blotting protocol and implementing in the Institute as well as validating, a new technology for measuring repeat expansions, Optical Genome Mapping.

1.11.2 Chapter 3 Aims

Primary aims of my work in this chapter were to explore genetic heterogeneity in RFC1 disease spectrum. In this study we leveraged short read WGS from the Genomics England sequencing project to investigate the normal and pathologic variation of the *RFC1* repeat expansion and to identify additional pathogenic repeat configurations in *RFC1* causing CANVAS and disease spectrum. Further, I contributed to describing the novel pathogenic repeat expansions and the full sequencing of novel pathogenic repeats was further analysed by targeted long read whole genome sequencing and sizes of the expansions were measured by optical mapping and/or Southern blotting. Finally, I address the testing complexity in RFC1 disease and the importance of combining the testing approaches to gain most accurate results.

1.11.3 Chapter 4 Aims

Primary aims of my work in this chapter were to describe a novel CMT associated gene *ARHGAP19* discovered in Professor Henry Houlden's Laboratory. Thanks to international collaborations such as SYNAPS, ICGNMD and various gene depositories, we were able to collect a cohort of individuals with biallelic mutations in *ARHGAP19* and further provide genetic, clinical, and functional evidence for *ARHGAP19* to be added to the list of GTPase Activating Protein (GAP) genes associated with human neurological disease. Further, I contributed to establishing invaluable links with external collaborators for their expertise and help with discerning the mechanisms of *ARHGAP19* disease causing mutations. Finally, I learned cell culture assays, explored in silico prediction tools and used standard molecular biology approaches such as quantitative PCR (qPCR) and Western blotting for RNA and protein levels assessment.

1.11.4 Chapter 5 Aims

Primary aims of my work in this chapter were to establish *Drosophila melanogaster* models in collaboration with Professor James Jepson's laboratory for studying functional consequence of loss-of-function of both the genes described in this thesis – *RFC1* and *ARHGAP19*. I explored the use of RNA interference driven genetic knock downs, genetic knockdowns using ubiquitin degradation system and finally genetic knockouts with null alleles. I further validated the models and used the available tools for assessing the phenotypes of the flies. In addition, for *RFC1* model system, I stressed the model with a known DNA damaging agent – cisplatin – and researched the consequences of cisplatin treatment in the *RFC1* knockdown fly.

CHAPTER 2. AAGGG repeat expansions in *RFC1* spectrum disorder

2.1 Introduction

2.1.1 Short tandem repeats

Short tandem repeats (STR) are sequences of DNA of two or more base pairs (typically 2-6 nucleotides) that are repeated consecutively at a locus and constitute over 6.5% of the human genome (Chintalaphani *et al*., 2021). These repetitive fragments of DNA can be located either in coding or non-coding regions of human genome. STRs can become expanded or contracted during DNA replication due to various mechanisms such as DNA polymerase slippage, formation of secondary structures like hairpins or errors in DNA repair (Francastel and Magdinier, 2019). Therefore, STRs cause a variation in human populations. However, due to their relative instability, STRs may become abnormally expanded and become pathogenic and to date nearly 50 such conditions have been described with up to 40 exhibiting neurological disorders.

2.1.2 Repeat expansions in disease

Some of the most common neurological expansion disorders include autosomal dominant CAG expansion in *HTT* gene causing Huntington's disease (Huang *et al*., 2022), autosomal dominant CTG expansion in *DMPK* gene causing myotonic dystrophy type 1 (DM1) (Bird *et al*., 1993) or autosomal recessive GAA expansions in *FXN* gene causing Friedrich's ataxia (Lam *et al*., 2023). There are about 10 known disorders without primary neurological features (Chintalaphani *et al*., 2021), and they present with developmental abnormalities such as GCG expansions in *PHOX2b* gene in congenital central hypoventilation syndrome (Amiel *et al*., 2003) or sight loss such as in CTG expansion in *TCF4* gene in Fuchs endothelial corneal dystrophy 3 (Fautsch *et al*., 2021).

Due to the advancements in sequencing technologies, namely next generation sequencing and further long read sequencing and optical genome mapping, a list of repeat expansion disorders is growing rapidly. Recently discovered genes include *VWA1* (Pagnamenta *et al*., 2021), *NOTCH2NLC* (Ishiura *et al*., 2019), *FGF14* (Pellerin *et al*., 2023); and in 2019 *RFC1* which comprises my next two chapters (Chapter 2 and 3).

2.1.3 Proposed pathomechanisms of repeat expansion disorders

Tandem repeats can be located in coding and non-coding parts of the genome; therefore, the mechanism of pathogenicity will depend on the expanded STR locus and if located in the gene, on the function of the particular gene. However, broadly the mechanisms are categorized in two groups: loss-of-function (LOF) or toxic gain-of-function (GOF).

2.1.3.1 Loss of function mechanisms

Loss of function mechanisms include alteration to gene products that lead to inability of that product to function as the wild type gene product, this may occur at RNA or protein levels. Such alternations can include post-transcriptional modifications as methylation defects as for example in fragile X syndrome where CGG expansion mutations are associated with hypermethylation of promotor region of *FMR1* gene (Bassel and Warren, 2008); defective transcription of DNA to RNA may arise from secondary DNA structures formed in STRs that can lead to incomplete RNA product. Further, nonsense mediated decay may lead to elimination of mRNA products and this in turn may cause no protein product (Swinnen *et al*., 2020). Other examples of loss-of-function may include incorrect protein folding that can lead to its degradation or inability to function (Matsell *et al*., 2024).

Traditional molecular biology approaches to detecting loss of function can include quantitative PCR for comparison of RNA levels between an affected and unaffected individuals or Western blotting for detection of changes at protein levels.

2.1.3.2 Gain of function mechanisms

Toxic gain of function can arise from RNA toxicity where repeat expansions (RE) in either coding or non-coding regions may cause alternative splicing events and lead to formation of pathogenic RNA species. RNA can also form unusual secondary structures (Frasson *et al*., 2022). Insoluble RNA foci, which are aggregates of the toxic RNAs sequestering RNA binding proteins or essential cellular components, may also form and result in a dysfunction of the complex (Zhang and Ashizawa, 2017). For example, a hallmark in myotonic dystrophy pathology is formation of RNA foci in nucleus (Chapuis *et al*., 2022). Another mechanism may be repeat associated non-AUG (RAN) translation which occurs when mRNA is translated into a protein without containing a start codon but rather RNA is folded into secondary structures that promote translation. RAN-translation has been described in SCA8, C9orf72 diseases, Huntington's disease (Banez-Coronel *et al*., 2015; Rudich *et al*., 2020). Misfolding and proteinopathy are commonly associated with poly-Q disorders – expansions of exonic CAG trinucleotide coding for glutamine. Large CAG expansions of polyglutamine tract become

aggregated and are insoluble causing neuronal damage, often in spinocerebellar ataxias where poly-Q aggregates in cerebellum (Kratter and Finkbeiner, 2010).

2.1.4 Common concepts in repeat expansion disorders

Typically, most repeat expansions associated with disease must surpass a certain variable threshold to become pathogenic (Chintalaphani *et al*., 2021). The healthy range of STR numbers in each neurological disease associated gene varies, for example, in recessive FRDA the normal, non-pathogenic expansion size of GAA is in range of 8-30 repeats, intermediate premutation alleles are between 30 and 60 repeats and pathogenic sizes are over 60 repeats (Rodden *et al*., 2023). In addition, it is quite commonly found in RE disorders that large expansions cause earlier age of onset and more severe symptoms. For example, DM1 with 50-150 repeats of CAG has a late age of onset over 20 years old and milder phenotype than DM1 caused by 100-1000 repeats where the age of onset occurs in adolescence and presents with classical, severe clinical phenotype of weakness, myotonia blindness and heart problems (Peric *et al*., 2021). Smaller, STRs associated with milder or variable phenotype are termed premutation alleles.

Another important consideration in RE disorders, that directly affects counselling of the patients, is clinical anticipation. Due to the meiotic instability of RE they may expand when transmitted to offspring and cause more severe and earlier disease. This is the case in HD, where larger repeats are more unstable (Riddley *et al*., 1991). However, not all RE disorders show genetic anticipation like for example FRDA which has a recessive mode of inheritance.

On the other hand, somatic instability refers to the variability of repeat expansion size within different tissues of the same individual which often expand further with time. In HD, the huntingtin gene is ubiquitously expressed, however, medium spinal neurons and putamen which are affected in the disease, have been shown to be susceptible to CAG instability. (Kovalenko et al., 2012; Sabado et al., 2024).

Another striking example of this phenomenon is a recently described somatic instability of CTG repeats in TCF4 gene causing Fuchs endothelial corneal dystrophy. There, CTG expansions can be detected in patient blood samples and measure between 54-160 triplet repeats, however, the repeat sizes in the affected corneal endothelial cell -derived samples of these patients ranged between 1800-11,900 repeats highlighting the importance of somatic instability in repeat expansion disorders and the relevance to the pathomechanisms of these diseases (Zarouchlioti et al., 2024).

2.1.5 Pathogenic cut-off and large expansions

Lower pathogenic threshold of RE disorders may be uncertain and debatable. It is not uncommon for an expanded allele to present in a healthy population like in a case of some SCAs (Sulek *et al*., 2004) or C9orf72 disease (Ishiura *et al*., 2019). Additionally, sizing of those repeat expansions may be inaccurate, especially in cases with very large expansions which may hinder investigations on correlation of size with severity and age of onset or anticipation.

2.1.6. Repeat expansions in *RFC1* cause Cerebellar Ataxia with neuropathy and vestibular areflexia syndrome

Cerebellar Ataxia with neuropathy and vestibular areflexia syndrome (CANVAS), prior to discovery of its genetic cause, had been described as a clinical observation with first description in 1991 (Bronstein *et al*., 1991) where clinical symptoms combined vestibular, neuropathic and cerebellar dysfunctions with broken pursuit of eye movements. In 2004, 4 patients with cerebellar ataxia and bilateral vestibulopathy were reported (Migliaccio *et al*., 2004); and finally, a carful phenotyping of 23 patients identified a typical triad of symptoms affecting cerebellum, sensory neuron, and vestibular system and thus the name CANVAS had arisen (Szmulewicz *et al*, 2011).

2.1.6.1 RFC1 gene discovery

In 2019, Cortese *et al*., using non-paramertic linkage analysis and WGS in 11 CANVAS families with 29 individuals of whom 23 were affected, identified tandem repeats in the second intron of replication factor C subunit 1 (*RFC1*) gene (hg19 chr4:39350045–39350103). The group of Rafehi *et al*., also identified non-reference tandem repeats in *RFC1* locus by using a bioinformatic based approach.

A normal allele consists of AAAAG pentanucleotide sequence that is repeated 11 times. However, the sequence can change to AAAGG or AAGGG and the pentanucleotides can further become expanded (fig.2.1). At discovery of the genetic cause CANVAS was described to be caused by biallelic AAGGG expansions with the number of repeats ranging from several hundred to more than 2000 repeats. Majority of the patients carried expansions of about 1000 pentanucleotide repeats.

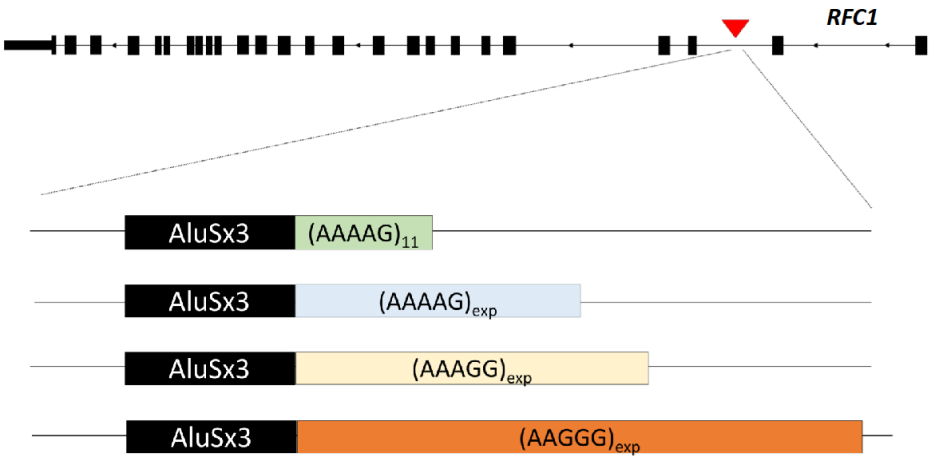


Figure 2.1 Schematic of *RFC1* gene with the expansion locus and the genetic heterogeneity seen at discovery of the repeat expansion in *RFC1*. A normal reference *RFC1* repeat locus consist of pentanucleotide AAAAG repeated 11 times. This pentanucleotide can become expanded. The sequence can also change to AAAGG or AAGGG which pentanucleotides can also become expanded. Figure adapted from Cortese *et al*., 2019.

The CANVAS patients shared a common haplotype which was estimated to have arisen about 25000 years ago, likely in Europe (Rafehi *et al*., 2019). Cortese *et al*., speculated that the change of the sequence from AAAAG to AAAGG or AAGGG could have arisen from an ancestral founder effect and the pathological expansions followed, possibly as a result of high Guanine-Cytosine content in the repeat.

CANVAS is a recessive disease, and it can be either sporadic or occur in siblings. Notably, few families with cousins affected, suggesting a pseudo-dominant inheritance, were also reported; those individuals were biallelic for expanded AAGGG with one of the alleles coming from another branch of the family. The pathogenic AAGGG is fully penetrant where all individuals with biallelic AAGGG RE develop the disease.

2.1.6.2 Patient phenotype

Late-onset ataxia is a common neurological condition where failure of systems controlling motor coordination occurs. This can lead to falls because of gait and stance ataxia and severe limitations in daily life. The disorder can be acquired, hereditary or non-hereditary; and up to 60% of familial and 19% of sporadic cases could have a genetic basis (Muzaimi *et al*., 2004, Gebus *et al*., 2017, Lieto *et al*., 2019) and in most patients, it can present without an obvious familial background (Klockgether 2010). CANVAS is a common cause of late-onset progressive ataxia, and the CANVAS patients suffer from ataxia, sensory neuronopathy or neuropathy as well as vestibular dysfunction (Szmulewicz *et al*, 2011).

Efforts have been made to piece together the syndromic clinical features of CANVAS with the genetic information to allow for more accurate clinical diagnosis. In 2020, Cortese *et al*.

reported the clinical features in the first 100 genetically confirmed *RFC1* CANVAS cases. The mean age of onset appeared to be just over 50 years old. Progressive unsteadiness was the most common complaint at disease onset and universally present during disease progression. A sensory neuropathy was identified as a common feature in all cases carrying biallelic AAGGG *RFC1* expansions. Patients often reported symptoms including loss of feeling, neuropathic pain, 'pins and needles' (paraesthesia) and unpleasant sensation in response to touch (dysesthesia), pointing to a damage to peripheral nerves. Notably, in some patients the disease manifested as isolated sensory neuropathy. Cerebellar involvement was observed in two thirds of patients, showing nystagmus, dysmetric saccades and broken pursuits and leading, as the disease progresses, to dysarthria and dysphagia. A characteristic radiological pattern of cerebellar atrophy affecting the vermis and hemispheric crus I was identified and further confirmed on post-mortem brains (Szmulewicz *et al*., 2011). Vestibular areflexia is also often present and probably its frequency is still underestimated. Patients may complain of oscillopsia and, when clinically tested, vestibulo-ocular reflex is often bilaterally impaired. Interestingly, over 60% of CANVAS patients experience dry cough (fig. 2.2) whose cause remains unexplained. The cough is reported up to thirty years before neurological onset (Cortese el al., 2020, Infante *et al*., 2018), and it is hypothesised to be arising either as hypersensitivity syndrome due to a peripheral mechanism where dysfunction of C fibres at level of upper way or oesophagus occurs; or due to cerebellar circuitry impairment (Infante *et al*., 2018). Nerve conduction studies showed non-length dependent sensory neuropathy in all the tested patients. Motor nerve conduction is preserved (Cortese *et al*., 2019). Visualisation of symptoms of 100 genetically confirmed CANVAS cases can be seen in fig.2.2 (Cortese *et al*., 2020).

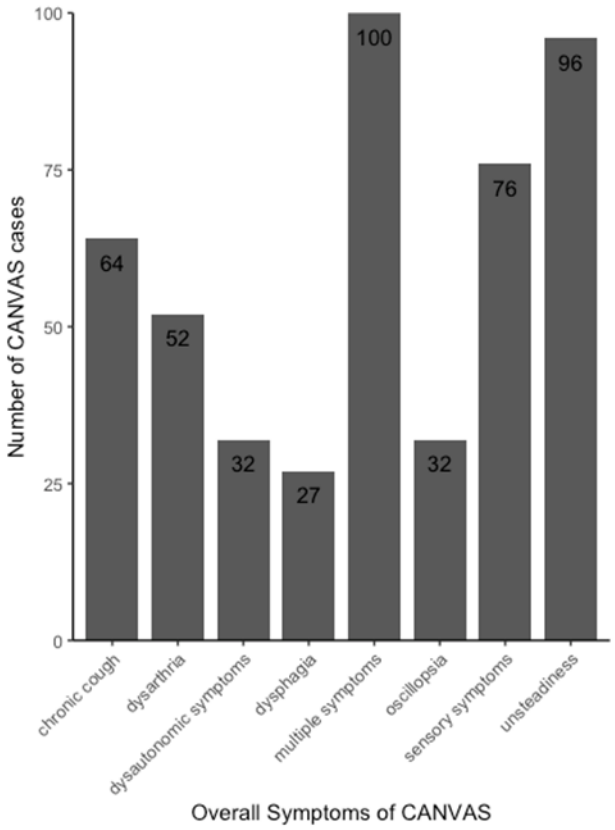


Figure 2.2 Overall symptoms of CANVAS. Symptoms during the manifestation of the disease in 100 biallelic RFC1 expansion cases. Listed are the number of patients reporting specific symptoms and a combination of 2 or more symptoms (multiple symptoms). Adapted from Cortese *et al*., 2020.

2.1.6.3 RFC1 expansions around the globe

The first cohort originating from 11 families with a CANVAS diagnosis studied by Cortese *et al*. consisted of 29 individuals of whom 23 were affected and six unaffected (Cortese *et al*., 2019). Additional cohort of 150 sporadic cases with late onset ataxia were screened and 22% of them were found to have the expanded AAGGG present and if only the individuals with sensory neuronopathy and/or bilateral vestibular areflexia were considered, the percentage would have been higher. The patients were of European ancestry (Cortese *et al*, 2019). Another cohort studied by Cortese included 363 Caucasian individuals with late-onset ataxia of whom 105 patients were identified to carry the biallelic $(AAGGG)_{exp}$ (Cortese *et al*., 2020). In a bioinformatics-based approach to screening the repeat expansions, a cohort of 35 individuals with clinically diagnosed CANVAS was recruited. Of those, 30 were found to carry the mutant biallelic repeat expansion and most of the individuals were of European ancestry and a few were of different ethnic backgrounds (Rafehi *et al*., 2019). Further, since the discovery of genetic cause of CANVAS and the first publications noted above, it has emerged that *RFC1* repeat expansions are common cause of cerebellar ataxia with neuropathy and it is underdiagnosed or misdiagnosed due to a range and variety of symptoms and relatively recent

characterisation of the disorder. The frequency of the $AAGGG_{exp}$ allele differs between populations worldwide. In a cohort of European descent, the allelic distribution for $AAGGG_{exp}$ was concluded to be 0.7% and conversely, the wild type, non-expanded AAAAG allele frequency equals 75.5% (Cortese *et al*., 2020). In a Canadian cohort of 163 control individuals, the frequency of expanded AAGGG was 4% and non-expanded AAAAG, 84.6% (Akcimen *et al*., 2019). In 490 healthy Chinese Han individuals, the frequency of $AAGGG_{exp}$ was found to be 2.24% and for $AAAAG_{11}$, 70.82% (Fan *et al*., 2020). Based on allele frequency the estimated disease prevalence at birth ranges from 1:10,000 to 1:650 individuals (Cortese *et al*., 2019). Wu *et al*. estimated the disease prevalence in Auckland, New Zealand to be nearly 1:100,000 (Wu *et al*., 2014), suggesting that disease is either under diagnosed or has reduced penetrance in the population.

As a result of the discovery there has been a high demand for *RFC1* screening in various populations across the globe and it is transpiring that the frequency of expanded AAGGG allele is as high as 7% (ranging from 0.7% to 6.5% in indifferent control populations (Davies *et al*., 2022), table 2.1). Furthermore, much can be learned from diverse populations, and I further talk about *RFC1* genetic heterogeneity in my next chapter.

Studies	Ethnicity	Cohort	$AAGGG_{(n)}$ frequency	Method of identification
Cortese et al., 2019[2]	European	304 controls	0.7%	PCR
Rafehi et al., 2019[3]	European	31 controls	6.5%	Bioinformatics
		69 Coriell WGS	4.3%	
		133 GTEx WGS	4.1%	
Akcimen et al., 2019[26]	Canadian	163 controls	4.0%	PCR
Fan et al., 2020[68]	Chinese Han	245 controls	2.2%	PCR
Wan et al., 2020[19]	Chinese	203 controls	1.0%	PCR

Table 2.1 Carrier frequency of Pathogenic AAGGG alle in healthy populations studied up to 2020. The AAGGG carrier frequency can range between 0.7% to 6.5% in healthy populations depending on control population studied. Adapted from Davies *et al*., 2022.

2.1.6.4 CANVAS and ataxias with similar features

Clinical diagnosis of CANVAS may be difficult due to the symptom overlap with several ataxic disorders which include FRDA, SCAs and Multiple System Atrophy (MSA) as shown in the table 2.2. It is therefore advised that only acquired causes of ataxia and neuropathy in patients are excluded, but also the patients are screened for FRDA and SCAs expansions (Dominik *et al*., 2020).

Disease	CANVAS	Friedreich ataxia	Spino-cerebellar ataxia	Multi system atrophy
Gene	RFC1	FXN	ATXN1-2-3, CACNA1, several others	No definite gene identified
Cerebellar ataxia	Frequent	Yes	Yes	Yes (MSA-C)
Neuropathy	Sensory neuropathy always present	Frequent sensory or sensory-motor neuropathy	Possible sensory or sensory-motor neuropathy depending on subtype	Usually absent
Vestibular areflexia	Frequent	Possible	Possible (SCA2)	Usually absent
Dysautonomia	Mild	Usually absent	Usually absent	Severe
Onset	Usually late onset	Usually early onset, but late onset possible	Usually early onset, but late onset possible	Usually late onset
Additional neurological features	Cough	Optic atrophy, hearing loss, pyramidal tracts involvement	Pyramidal tracts involvement, parkinsonism, cognitive impairment, visual impairment, variably associated depending on subtype	Parkinsonism, rapid progression, REM behaviour disorder
Extra neurological involvement	No	Cardiomyopathy, diabetes, scoliosis	No	No

Table 2.2. CANVAS, FA, SCA and MSA patients share a number of complaints which may include ataxia, sensory neuropathy, dysarthria and dysphagia (Dominik *et al*., 2020, Szmulewicz *et al*., 2011, Cortese *et al*., 2020., Delatycki *et al*., 2012, Palma *et al*., 2018)

2.1.7 Gold standard molecular techniques for RE disorder diagnosis

The gold standard techniques for detecting repeat expansion disorders have been repeat primed PCR (RP-PCR) for the detection of specific expansion at a known gene locus and Southern blotting for confirming the results of PCR and measuring the size of the expansions (fig.2.3).

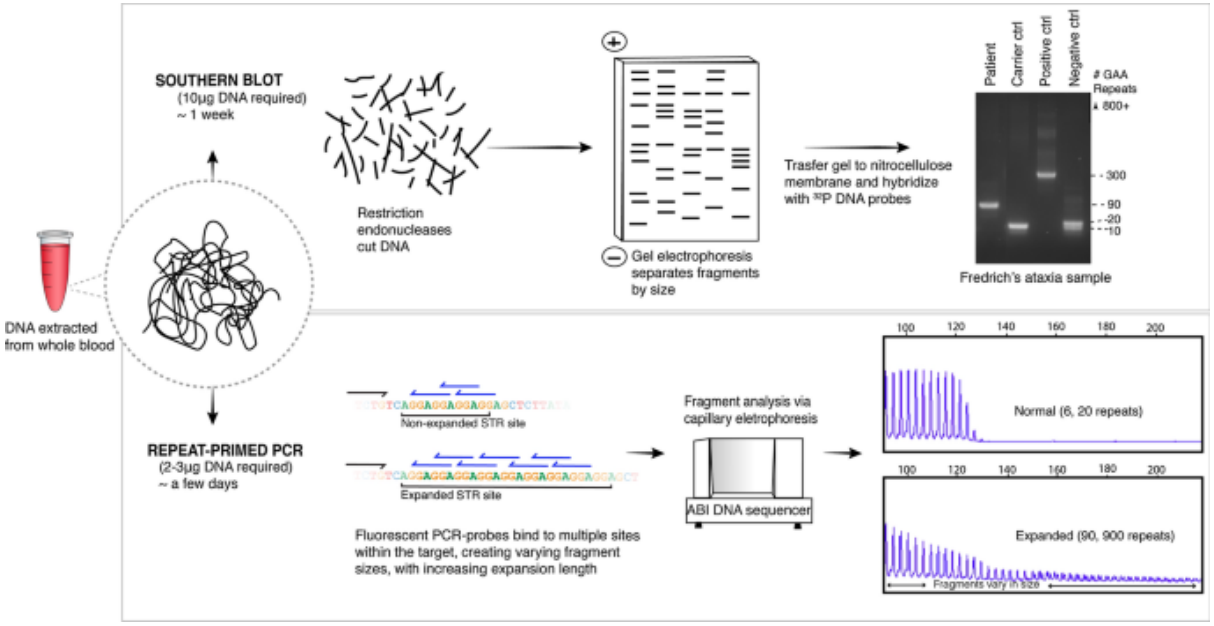


Figure 2.3 Example of RE testing flowchart. Genomic DNA is used for testing of presence of a specific expansion motif at a given locus and Southern blotting is performed for confirmation of PCR results and sizing the expansion. Adapted from Chintalaphani *et al.*, 2021

And indeed, current diagnostic strategy for *RFC1* testing relies on polymerase chain reaction, including flanking PCR and RP-PCR (Cortese *et al*., 2019; Dominik *et al*., 2020; Chintalaphani *et al*., 2021; Facchini*, Dominik* *et al*., 2023). However, given the large size and high Guanine – Cytosine (GC) content of the pathogenic AAGGG motif, PCR-based techniques fail to amplify the full expanded repeat. Therefore, demonstration of the presence of two expanded alleles and measurement of their size was only possible with traditional Southern Blotting (SB). SB utilises a pre-designed probe that only binds to a specific locus flanking the *RFC1* repeat, and the expansion sizing is based on the visual comparison between the sample track and a reference ladder track (Cortese *et al*., 2019). Despite being clinically very useful, SB is a time-consuming technique which requires considerable amount of work and a dedicated laboratory setup.

Southern Blotting has been a gold standard technique for measuring allele sizes in various conditions such as C9orf72 repeat expansion disorders (DeJesus-Hernandez *et al*., 2011) myotonic dystrophy type 1 (DM1) (Joosten *et al*., 2020), and fragile X syndrome (Curtis-Cioffi *et al*., 2012). SB is a cumbersome method and studies have been carried out whether more

convenient and high-throughput methods, so far mainly limited to PCR, can replace or minimise the need for SB (Grasso *et al*., 2014; Chen *et al*., 2010). However, PCR cannot amplify large repetitive sequences; therefore, it is not possible to use it for sizing of repeat expansions.

Optical Genome Mapping (OGM) is a new technology which enables accurate detection of large (>500 nucleotides) Structural Variants, based on the measurement of the distance between fluorophore-labelled probes which tag ultra-high molecular weight DNA molecules. The advantages of this technique include the following: (1) a more streamlined laboratory protocol; (2) the possibility of mapping the entire genome for each sample, instead of a single locus; (3) the possibility of automatizing the data analysis. The main commercial implementation of OGM is currently provided by Bionano Genomics. Until implementation of OGM at UCL and before research included in this thesis, Bionano OGM was able to reliably detect the presence of repeat expansion in DM1 and SCA10 (Otero *et al*., 2021; Torres *et al*., 2022). In addition, OGM was successfully used to confirm the presence of biallelic *RFC1* expansions in seven Dutch patients carrying *RFC1* expansions (Ghorbani *et al*., 2022). However, a systematic comparison between OGM and SB was never performed. Therefore, in addition to using Southern blotting, I explored OGM for measurement of *RFC1* repeat expansions.

2.1.8 Pathomechanisms

The underlying pathomechanisms of CANVAS is still unknown, however, it is suspected to be a loss-of-function mechanisms due to the recessive mode of inheritance and lately the discovery of truncating variants in trans with the pathogenic AAGGG motif in 2022 (Ronco *et al*., 2022) that my work contributed to (further described in chapter 3). Furthermore, so far there has not been evidence for common RE gain-of-function mechanisms in RFC1 disorder such as no evidence for RAN translation or intron retention and formation of RNA foci. However, investigations of bulk tissues (peripheral tissues and post-mortem brain samples) from CANVAS patients, there appears to be no evidence for reduction of canonical *RFC1* transcript at mRNA level or no loss of protein product highlighting that the loss-of-function hypothesis may be limited to a subset of cells or masked by bulk tissues.

Investigations on postmortem brains of CANVAS patients revealed substantial loss of Purkinje fibres, most severe in vermis. This is in line with established knowledge that ataxia is marked with neuronal loss in cerebellum. In addition, available nerve biopsies of CANVAS patients revealed loss of small and large myelinated fibres (Cortese *et al*., 2019).

2.1.9 Replication Factor C complex

The replication factor C complex (RFC) is a protein complex composed of RFC1-5 and is essential for DNA replication and DNA damage response. RFC1 is the biggest protein of the complex at 140kDa (others being at 36-40kDa in size) and it contains the main DNA-binding region and interacts with proliferating cell nuclear antigen (PCNA) (Majka and Burgers, 2004).

The complex loads PCNA and DNA polymerase in presence of adenosine triphosphate (ATP) onto DNA and allows for its elongation in presence of dNTPs and other involved proteins (fig.2.4 A). Moreover, the complex can act in DNA damage response, specifically in mismatch repair and excision repair. Importantly, it has been shown that RFC1 consists of three domains that provide a binding site for DNA, and it is used for binding of gapped or nicked DNA (fig.2.4 B) (Li *et al.*, 2018; Liu *et al.*, 2022) .

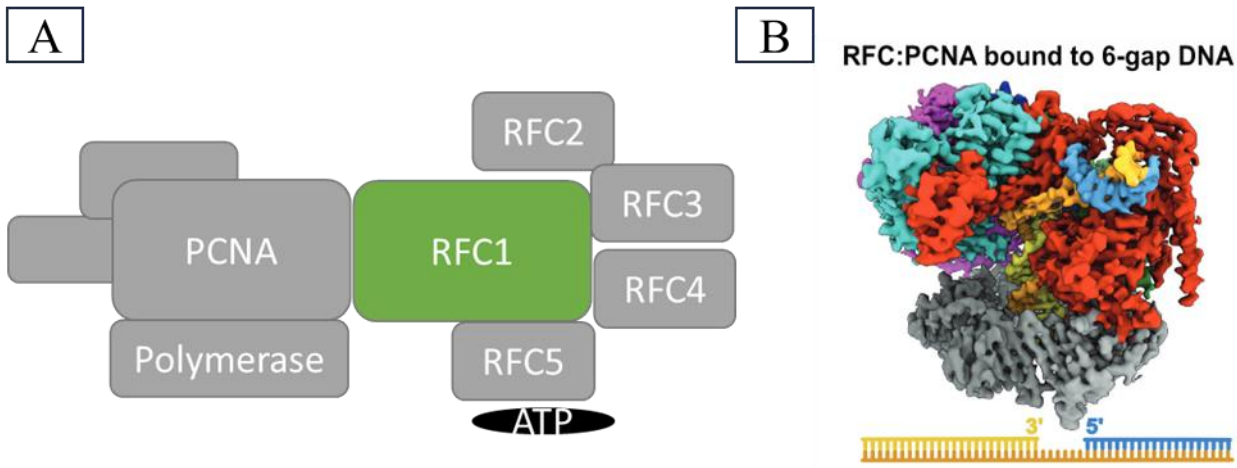


Figure 2.4 A schematic of RFC complex with RFC1 binding PCNA in presence of ATP (A) and an example of RFC:PCNA complex binding to a gapped DNA of 6 nucleotides (B). Adapted from Liu *et al.*, 2022.

To date, as described above, there has been no evidence of decreased RNA or protein levels of *RFC1* in CANVAS patients with biallelic AAGGG repeat expansions, and this is also the case for other RFC1 proteins forming the complex. Interestingly, dysfunction of the complex or its members has been implicated in numerous cancers such as breast cancer, acute myeloid leukaemia, ovarian and many others (Li *et al.*, 2018).

2.2 Materials and methods

Main contributors to the methods (and results) in this chapter are listed in table 2.3

Contribution table	
Methods used	**Contributors**
CANVAS Screening (PCR & Southern blotting)	**Natalia Dominik**, Riccardo Curro, Roisin Sullivan, Valentina Galassi Defoire
Southern blotting optimisation and troubleshooting	**Natalia Dominik**
Bionano Optical Genome Mapping	**Natalia Dominik**, Stephanie Efthymiou
Clinical Examinations	Andrea Cortese, Henry Houlden, Riccardo Curro, RFC1 repeat expansion study group
Bioinformatic analyses	Stefano Facchini

Table 2.3. Main contributors to the methods (and results) in this chapter

2.2.1 Patients and rationale

Multi centre cohort of 2334 patients with clinical diagnosis of idiopathic sensory neuropathy, late-onset (> 25 years old) cerebellar ataxia, complex neuropathy or CANVAS were collected. Where testing was available, other causes of spinocerebellar degeneration were excluded such as acquired causes, SCAs, FRDA. Availability of WGS was not an inclusion criterion in this study.

Using standard screening methods for *RFC1* expansions which included flanking PCR concurrently with RP-PCR for canonical AAGGG, AAAGG and AAAAG, I screened 1531 patient samples. Samples with no band on flanking PCR, positive sawtooth pattern on AAGGG RP-PCR but negative on the non-pathogenic AAAAG and AAAGG configurations were subjected to Southern blotting if sufficient DNA was available (fig.2.5), and I performed Southern blotting for 395 patients. In addition, to compare OGM and Southern blotting techniques, 17 CANVAS patients' blood samples with biallelic pathogenic repeat expansions were subjected to Bionano Optical Genome Mapping. I performed OGM on 10 patient samples.

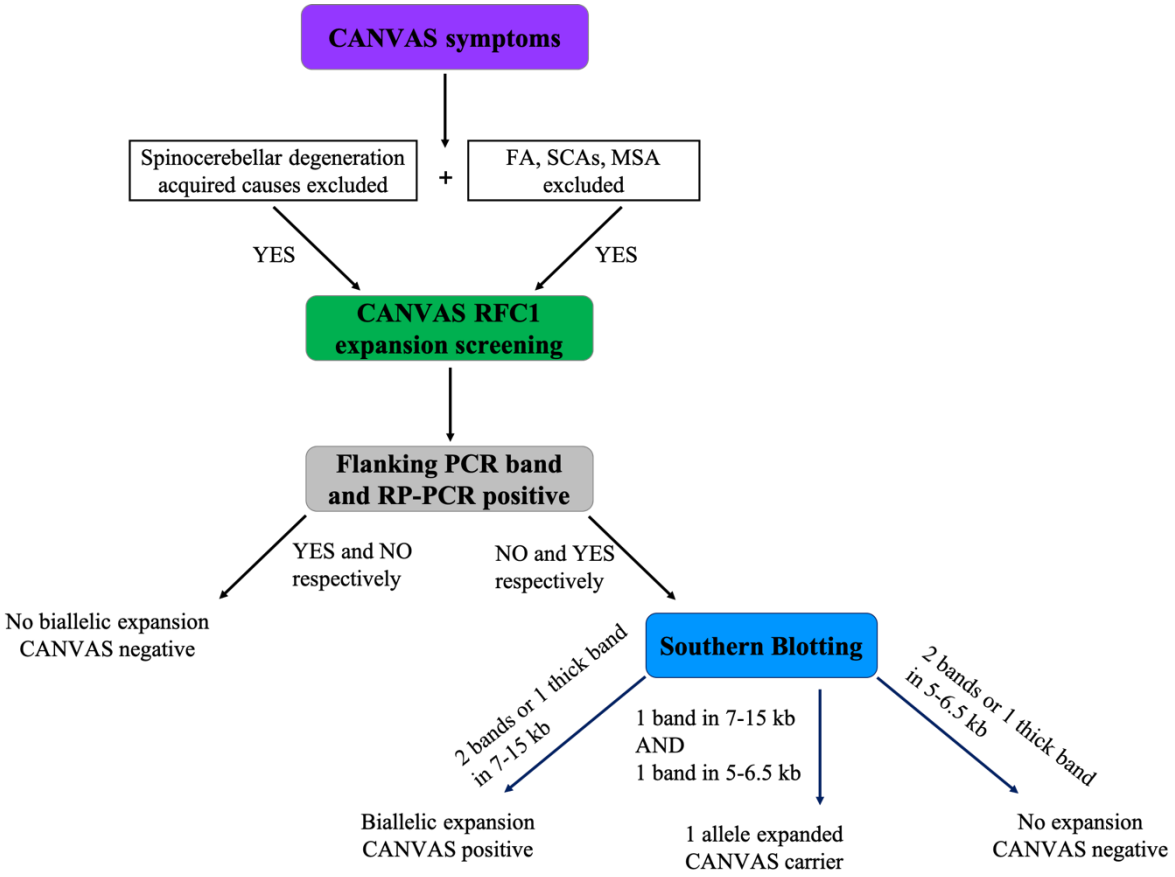


Figure 2.5 Workflow diagram representing repeat expansion screening methodology. Flanking PCR and RP-PCR are used simultaneously on patient DNA to identify which are more likely to have two expanded alleles. If flanking PCR shows no amplifiable product and RP-PCR shows typical saw-tooth pattern, Southern blotting is carried out on additional patient DNA if available (Dominik *et al*., 2020)

2.2.2 Clinical details

Clinical and demographic data for patients testing positive for *RFC1* expansions was collected using a standardised table completed by all referring clinicians. The data included family history, age at neurological onset, age at onset of sensory, cerebellar or vestibular symptoms, use of walking aids, and detailed first and last available neurological examinations. Based on the information, patients were divided into three phenotype categories: sensory neuropathy, complex ataxia/neuropathy or CANVAS. Disease severity and progression were indicated by difficulty in walking resulting in a need for walking aid and presence of dysarthria and dysphagia. Only patients of Caucasian ancestry were considered in the analysis.

Considering that as many as 70% of CANVAS patients may present with chronic cough, this data was also recorded, but not considered as first neurological symptom for the age of onset of the condition.

2.2.3 Flanking PCR

1µl of DNA of concentration of 25-150ng/µl was added to 7.5µl FastStart mastermix (Roche), 5µl of PCR grade water and 1µl of 10uM Forward and Reverse primers and ran in PCR conditions as in table 2.4. 7µl of PCR product was subjected to gel electrophoresis on 2% agarose gel against a 100bp DNA ladder (Gel Pilot) at 90V for 30min. The results were recorded as no amplifiable product where no band was seen, reference where the band was of wild-type (WT) size of 350bp and intermediate as a band seen at higher size than WT.

2.2.4 RP-PCR

1µl of DNA of concentration of 25-150ng/µl was added to 7.5µl of Phusion Mastermix (Biolabs) with 5µl of PCR water and 1µl each of 10uM FAM labelled forward primer, anchor primer and reverse primer (table 2.4). Two reverse primers are used in repeat-primed PCR: an anchor and a reverse primer (fig.2.6A). The reverse primer binds to a specific repeat motif that is being tested and may bind to the repeat in any place of the repeat tract which results in the positive 'sawtooth' pattern (fig.2.6B). The reverse primer contains a stretch of DNA sequence termed 'a clamp' which is complementary to the DNA sequence immediately after the repeat expansion sequence. This allows for amplification of the entire repeat expansion tract in the alleles with a lesser number of repeat expansions (<~60). The anchor primer that binds to the tail of the reverse primer for an amplification in further cycles of PCR when the reverse primer becomes depleted. The tail of the reverse primer and the anchor primer are designed not to be complementary to any DNA sequence in human genome to avoid unspecific amplification.

Reverse primer and cycler conditions were changed according to repeat expansion tested (table 2.4).

	Sequence	Reagents	PCR conditions
Short range flanking PCR	Fw: TCAAGTGATACTCCAGCTACACCGTTGC Rv: GTGGGAGACAGGCCAATCACTTCAG	7.5µl Fast-start Master Mix 5ul H2O 1ul forward primer 10 µm 1ul reverse primer 10 µm 1ul DNA (around 50ng/µl)	95°C 4 mins; [95°C 30 secs 63°C 50 secs 72°C 60 secs] x35; 72°C 5 mins
RP-PCR	Fw FAM: TCAAGTGATACTCCAGCTACACCGT Anchor: CAGGAAACAGCTATGACC RFC1-WT CAGGAAACAGCTATGACCAACAGAGCAAGACTCTG TTTCAAAAAAGAAAAGAAAAGAAAAGAAAA RFC1-V CAGGAAACAGCTATGACCAACAGAGCAAGACTCTG TTTCAAAAAAGGAAAGGAAAGGAAAGGAAA RFC1-MUT CAGGAAACAGCTATGACCAACAGAGCAAGACTCTG TTTCAAAAAAGGGAAGGGAAGGGAAGGGAA	Phusion Flash High-Fidelity PCR Master Mix 2X (Thermo-Fisher) Primers 0.5µM gDNA 50ng	98°C 3 mins; [98°C 10 secs 70°C (M)/ 65°C (V)/ 55°C (W) 50 secs 72°C 2 mins] x35; 72°C 1 min
DIG probe generation	Fw: ATTAGGTGTCTGGTGAGGGC Rv: GAAGAATGGCCCCAAAAGCA	Faststart Master Mix 2X (Roche) Primers 0.5µM Plasmid 50mg Roche PCR DIG probe synthesis mix vial no.2	95°C 4 mins; [95°C 30 secs 63°C 30 secs 72°C 1 min] x35;

Table 2.4. Thermocycler conditions for Flanking PCR and RP-PCR

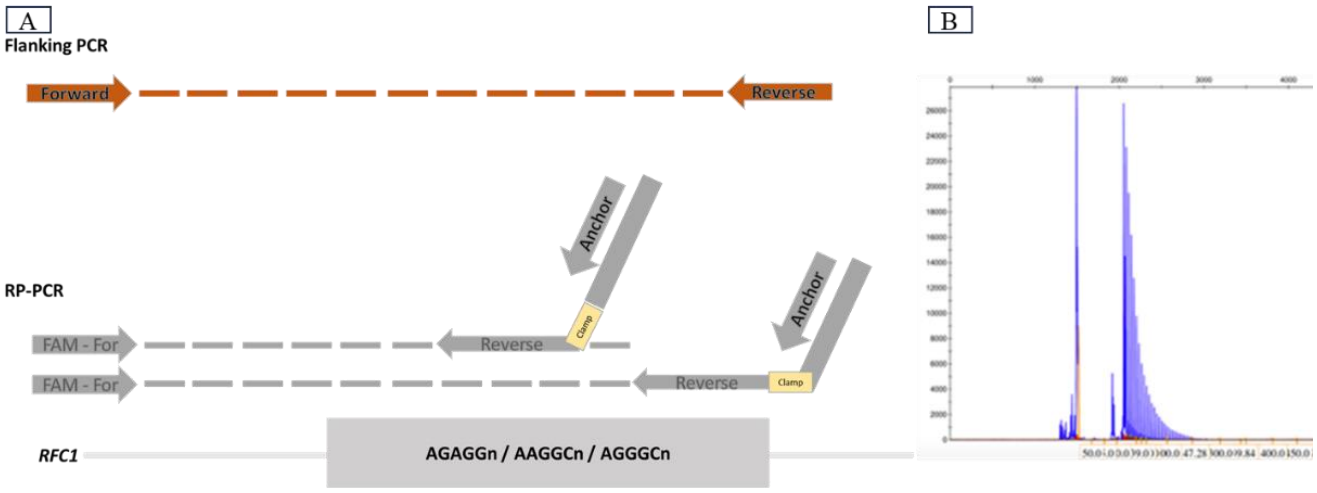


Figure 2.6 A schematic of binding of *RFC1* primers used and a positive RP-PCR pattern example. Flanking PCR (top) uses a forward and reverse primer to amplify across the second intronic region of *RFC1* where expansions may be found. RP-PCR (bottom) uses a fluorescently labelled forward primer and two reverse primers – a reverse primer that binds to the specific repeat motif tested and an anchor primer that binds to the tail of the reverse primer for an amplification in further cycles of PCR when the reverse primer becomes depleted. The reverse primer contains a stretch of DNA sequence termed 'a clamp' which is complementary to the DNA sequence immediately after the repeat expansion sequence. The tail of the reverse primer and the anchor primer are designed not to be complementary to any DNA sequence in human genome B) RP-PCR with primers targeting the AAGGG pentanucleotide repeated unit. An ABI 3730 DNA Analyser was used to separate the products, and these were visualised using GeneMapper. The presence of a 'sawtooth' pattern is characteristic of a possible affected individual

Fragment analysis was performed on ABI 3730xl Genetic Analyser (Applied Biosystems) and analysed using Geneious Prime or GeneMapper software. Expansions are visualized as a decremental 'sawtooth' pattern (fig 2.6B).

2.2.5 Sanger sequencing

Sanger sequencing was performed when additional confirmation of repeat expansion was warranted due to technical difficulties or inconclusive results.

1µl of gDNA of concentration 25-150ng/µl was added to 35µl of Phusion Mastermix (Biolabs) with 25µl of PCR water, 3% DMSO and 1µl of forward and reverse flanking primers as in table 2.4. PCR cycling conditions are available in the table. The PCR products were enzymatically purified with FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific) prior to sequencing at Source Bioscience. The electropherograms were visualized with Geneious Prime.

2.2.6 Digoxigenin labelled probe synthesis

Digoxigenin (DIG)-labelled DNA probe for Southern blotting was prepared using PCR DIG Probe Synthesis Kit (Roche). One microliter of 50ng/µl of plasmid containing 1kB DNA sequence flanking the *RFC1* repeat locus was used to mix with 20µl Fast Start Master Mix,

17µl of PCR grade water, 1µl of 10pMol forward and reverse primers (table 2.4) and 3µl DIG Probe synthesis mix. PCR conditions are shown in table 2.4. DIG labelled PCR products are the probes used for detection of repeat expansions in *RFC1* on Southern blotting membrane and they further bind anti-DIG antibody for signal amplification for detection of the alleles on X-ray film.

2.2.7 Southern blotting

2.2.7.1 DNA preparation and gel electrophoresis

Up to 18 samples can be processed per blotting (18 samples and 2 ladders). Five micrograms of genomic DNA at concentration of 150ng/µl was diluted with water up to 33µl total volume on a 96-well plate. The workflow of southern blotting is shown in figure 2.7.

A master mix of 4µl 10X CutSmart buffer (NewEngland Biolabs), 2µl Spermidine (Sigma) and 1µl of Eco RI 100,000 U/ml (NewEngland Biolabs) was prepared per sample. Subsequently, 7µl of the mix was added to the samples diluted on 96-well plate and mixed well by pipetting up and down. The plate was put in a PCR cycler and ran at 37ºC for 1 hour after which an additional 1µl of Eco RI 100,000 U/ml added per sample and the plate ran at 37ºC for additional 2 hours.

A 1.5% agarose gel was prepared in a large (25cm) casting tray. 8.6µl of Blue/Orange 6X loading dye (Promega) was added to the enzymatically cut samples and pipetted thoroughly. Molecular Weight Markers II and III DIG-labelled (Roche) were prepared for 2 wells by mixing 4µl of water, 1µl of the marker and 1µl of the blue-orange dye, per well.

The samples and the markers were loaded on the agarose gel and ran overnight at 40V for around 15 hours.

2.2.7.2 Gel treatment and blotting

After the sample front reached the bottom of the gel, the gel was washed in distilled water for 5 minutes; and in depurination solution (475 ml water + 25 ml concentrated HCl) for 45 minutes. Denaturing solution was prepared by mixing 10g NaOh + 29.2g NaCl in 500ml of distilled water and after another 5 minutes wash in distilled water, the gel was washed in denaturing solution for 45 minutes. The gel was then again washed in distilled water for 5 minutes. A 45-minute wash with pre-made neutralising solution followed (Thermo Scientific).

Following the washes, the blotting was assembled by creating a sandwich of: 4 long (~50cm) 3MM Whatmann paper pieces in a non-spill tray, soaked in 1 litre of 10XSSC solution (Lonza).

The gel was put on the soaked paper and a 20x20 cm positively charged nylon membrane (Roche) was carefully positioned on the gel. The edges of the sandwich were masked with cling film and a 20x20 square of 1 CHR paper soaked in 2XSSC was put on the membrane. Then 4 pieces of dry 3MM Whatmann paper and a stack of about 20cm paper towels were positioned on top. A glass piece was placed on the sandwich and a bottle weight on top. The DNA was allowed to transfer overnight for about 15 hours.

2.2.7.3 Membrane pre – and hybridisation

After the blotting, the sandwich was disassembled and the membrane inspected, placed on 20x20 square of 1 CHR paper soaked in 2XSSC (Lonza) and placed in UV transilluminator for 3 minutes. Subsequently, the membrane was placed in a plastic container and pre-hybridisation with 40ml DIG Easy-Hyb (Roche) solution followed for 4-5 hours in shaking incubator at 49ºC.

A probe mixture of 60µl 1kb *RFC1* probe and 150µl salmon sperm (Agilent) was boiled for 5 minutes. It was then immediately snap-cooled on ice and added to 30ml Roche DIG Easy-Hyb solution to make hybridisation mixture. The pre-hybridisation solution was poured off the membrane and replaced with the hybridisation mixture. The box was placed in shaking incubator at 49ºC overnight for about 15 hours.

A 2 litre 0.1X SSC / 0.1% SDS solution was prepared by mixing 1970 ml distilled water, 10ml 20X SSC and 20ml of 10% SDS. The solution was placed in 65ºC oven overnight for use the next day.

2.2.7.4 Membrane washing and detection

The hybridised membrane was washed twice for 15 minutes with 500ml of 2XSSC/0.1%SDS solution prepared by mixing 890 ml of distilled water with 100ml 20XSSC and 10ml 10% SDS.

Subsequently, it was washed 4 times for 15 minutes with the 65ºC 0.1X SSC / 0.1% SDS, in a shaking incubator at 65ºC. A 2-minute wash at room temperature with 200ml washing buffer (Roche) followed. The membrane was then incubated with 200ml block solution made by mixing 194.4ml distilled water with 21.6ml 10X Maleic acid buffer (Roche) and 24ml 10X block solution (Roche) at room temperature, shaking.

The antibody solution was prepared by centrifuging Anti-DIG AP antibody (Roche) for 10 minutes at 10000rpm and 4µl were immediately aliquoted from the top of the solution into 40ml Block solution. The membrane was drained from the blocking solution and incubated in the antibody solution for 30 minutes at room temperature.

The membrane was washed twice for 15 minutes with 200ml washing buffer at room temperature, then for 5 minutes with detection buffer and the membrane was then drained.

To visualize the DNA bands, a chemiluminescent CDP-STAR substrate (Roche) was put on the membrane and the membrane was wrapped with cling film and transferred to X-ray film. X-ray developer was used to visualize the film and the bands were measured.

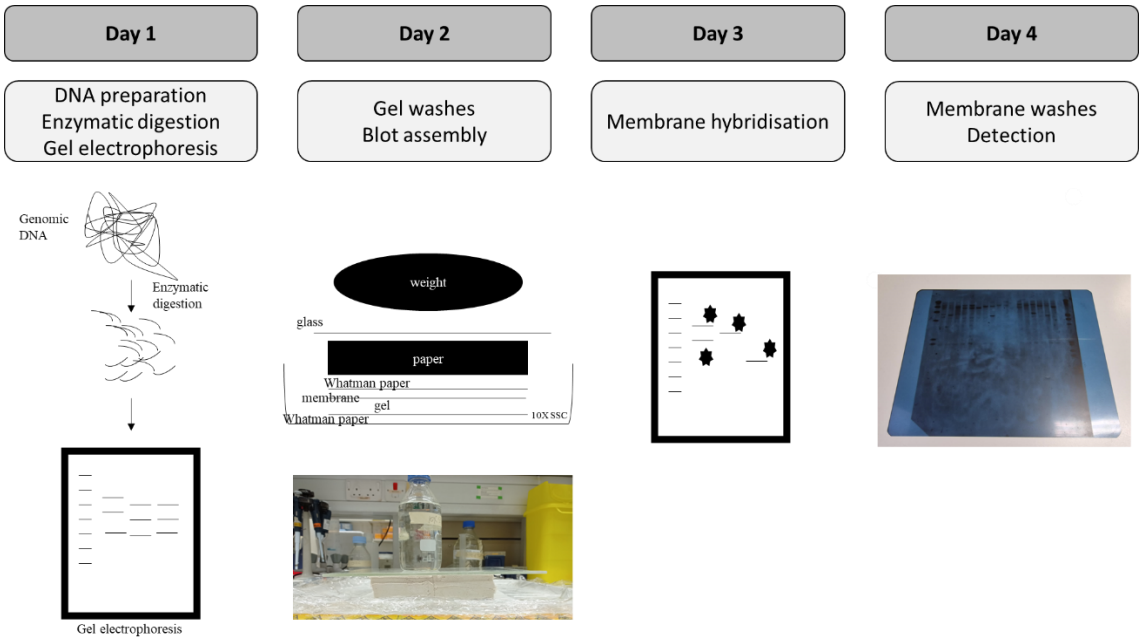


Figure 2.7 A schematic of Southern blotting workflow. Genomic DNA is enzymatically digested and separated on agarose gel. The DNA is then transferred onto positively charged membrane and the membrane hybridised with *RFC1* specific probe. The bands are visualised with chemiluminescent substrate.

2.2.7.5 SB: Repeat expansion size measurements

Repeat expansion size were measured using a standardised algorithm. Distances between the ladder track were measured with a ruler and the visible patient bands were measured according to a line drawn between 6.5kb ladder markers on two points of the blotting. The sizes were calculated by plotting the ladder points on linear regression graph and using the corresponding equation and subtracting the size of the probe binding to normal allele (5000bp). Sizes are presented in repeat numbers rather than base pair size.

2.2.8 Southern blotting optimizing

One of the first objectives of my PhD was optimising the Southern blotting technique described above. The original protocol produced good images; however, we observed incomplete transfer of DNA to the positively charged nylon membrane in several cases. Therefore, the transfer time was extended from approximately 4 hours to 15 hours (overnight).

After transfer, pre-hybridisation was extended to 5 hours from the original 3 hours and the temperature was increased to 49°C from 46°C. The increase of temperature was also implemented for the overnight hybridisation step.

Biallelic expansions in affected individuals are seen as two bands between 6.5 to 15kb, or one thicker band if the expansions on both alleles are the same or similar size. Unaffected individuals who carry the mutation can either have one band in normal, wild type, range of 5kb and one band in expanded range of 6.5 to 15kb, or two expanded alleles - one in the non-pathogenic range of up to around 6.5kb and one in the pathogenic range (fig.2.8). Although Southern blotting remains the gold standard technique to confirm the presence and establish the size of biallelic *RFC1* expansions, it has several limitations. Firstly, it requires a relatively large quantity (5µg or more) of good quality (260/280 ratio of 1.8-2 and 260/230 ratio of 2-2.2) DNA. Secondly, it is a time-consuming and labour-intensive technique that requires a specific laboratory set-up. Third it is rather cumbersome; a skilled and trained operator is needed to ensure final readout is trustworthy.

A new technology, Bionano Optical Genome Mapping, has been developed to detect structural variants in DNA and I have used Southern Blotting and Bionano for *RFC1* expansion sizing of the same individuals to compare the methods.

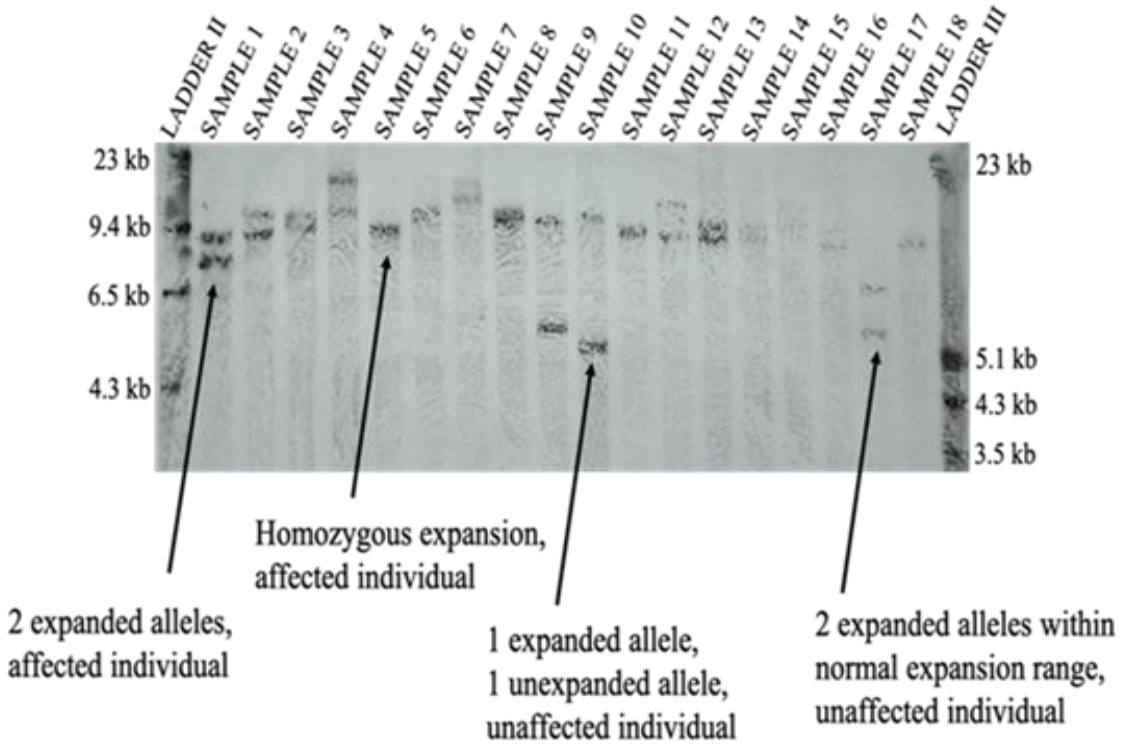


Figure 2.8 Patients are characterised by either one overlapping band or two bands within the region 7 to 15kb. Carriers are identified with one band residing between 7 and 15kb and the other at 5kb-6.5kb, equivalent to the non-expanded AAGGG sequence or a small AAGGG expansion. Non-affected individuals exhibit 2 bands in regions between 5-6.5kb. Two ladders are needed for accurate measurements: DIG-labelled DNA Molecular Weight Marker II (Roche) (labelled as LADDER II) and DIG-labelled DNA Molecular Weight Marker III (Roche) (labelled as LADDER III). The left- and right-hand side of each panel documents the molecular weights represented by LADDER II and LADDER III respectively. Figure from Dominik *et al*., 2020

2.2.9 Bionano optical genome mapping

2.2.9.1 DNA extraction

Ultra-high molecular weight DNA was extracted from venous blood using Bionano Genomics provided kits as described in Bionano Prep SP Frozen Human Blood DNA Isolation Protocol v2. Briefly, frozen blood was thawed at 37°C and white blood cells (WBC) were counted using Heamocytometer. And an appropriate volume of blood was taken to transfer 1.5 million WBC which were pelleted for 5 min at 4000g. DNA stabilising buffer and proteinase K were added to the pellet and resuspended. Then samples were rotated on hoolamixer with Lysis and Binding Buffer (LBB) for 15 min. Following that PMSF was added and after a 10-minute incubation, a nanodisk was dropped into the tubes as well as isopropanol. Samples were incubated for 15 minutes in hoolamixer and the DNA precipitated and attached to the magnetic nanodisks. The disks were subjected to washes with the wash buffers included in the kits and later the DNA was eluted from the nanodisk using Elution Buffer provided. DNA was carefully sheared by pipetting up and down 5 times at 60 seconds per aspiration and release.

2.2.9.2 DNA labelling

The concentration and homogeneity of the DNA samples was estimated using Qubit and the homogeneous high molecular weight DNA was labelled using Bionano Prep Direct Label and Stain (DLS) Protocol with kit provided. Briefly, 21µl of DNA was incubated at 37°C with the label mixture provided and for further 30 min with proteinase K. The labelled DNA was cleaned for 60min using membranes provided and the homogeneous labelled DNA was loaded onto a Saphyr chip.

2.2.9.3 Chip loading

8.5µl of labelled DNA was loaded into the inlet of the flowcell of the chip and allowed to migrate for 2 minutes. Subsequently, 11µl was of DNA was loaded into the outlet of the chip. A few drops of PCR grade water were added on top of the DNA in the inlet and outlet to form a convex. The flowcell was then closed with custom tops and the remaining 2 flowcells were loaded with another patient DNA following the same procedure. The chip was then closed with a clip and positioned in the Saphyre machine. The workflow for OGM is shown in figure 2.9.

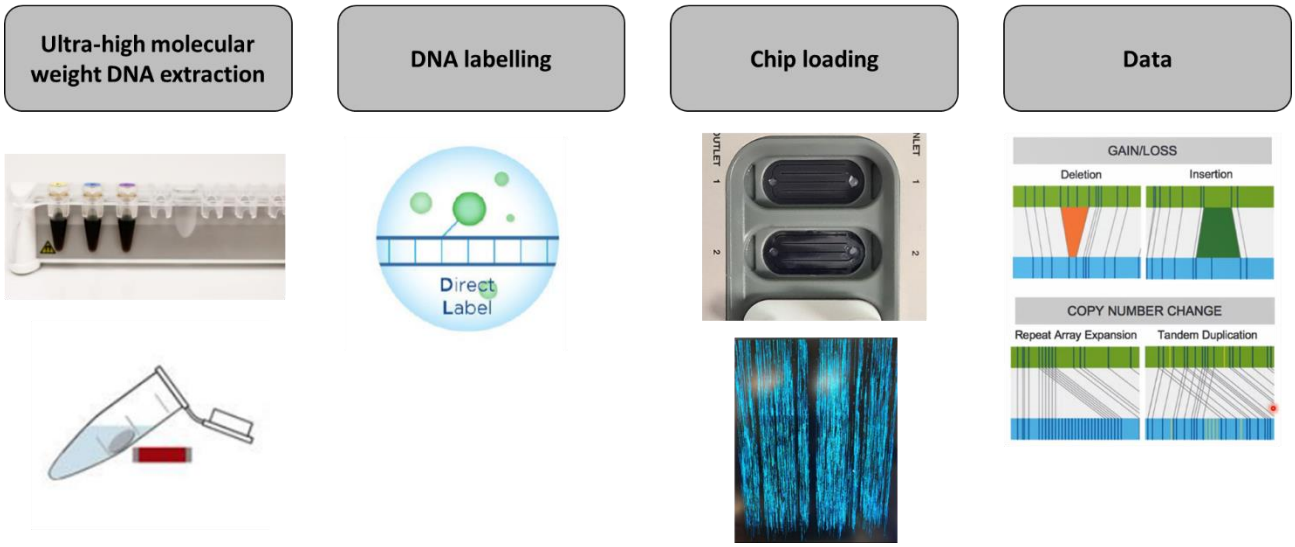


Figure 2.9. A schematic of Bionano optical genome mapping workflow. Bionano kits are used for extraction of ultra-high molecular weight DNA and for subsequent labelling of the DNA. The labelled DNA is inserted onto a chip which allows the molecules to be linearised and labels visualised on DNA backbone. The data can be seen on Bionano Access interface and the differences between the labels can infer presence of large structural variations.

2.2.9.4 OGM: Repeat expansion size measurements

Size measurements in OGM rely on fluorescent labels that bind to DNA motifs of CTTAAG that are present in the genome around 15-20 times per 100 kbps. Fig.2.10 A shows an example of a normal allele aligning to the reference genome, further presenting that the intervals between the labels vary between different loci therefore enabling correct mapping of the labelled DNA molecules. Multiple DNA molecules are mapped to the reference and the distance between markers 7723 and 7724 which contain *RFC1* repeat locus (fig.2.10 B), can be accessed for all the molecules. The data is then assessed as Gaussian distribution for repeat expansion size visualization.

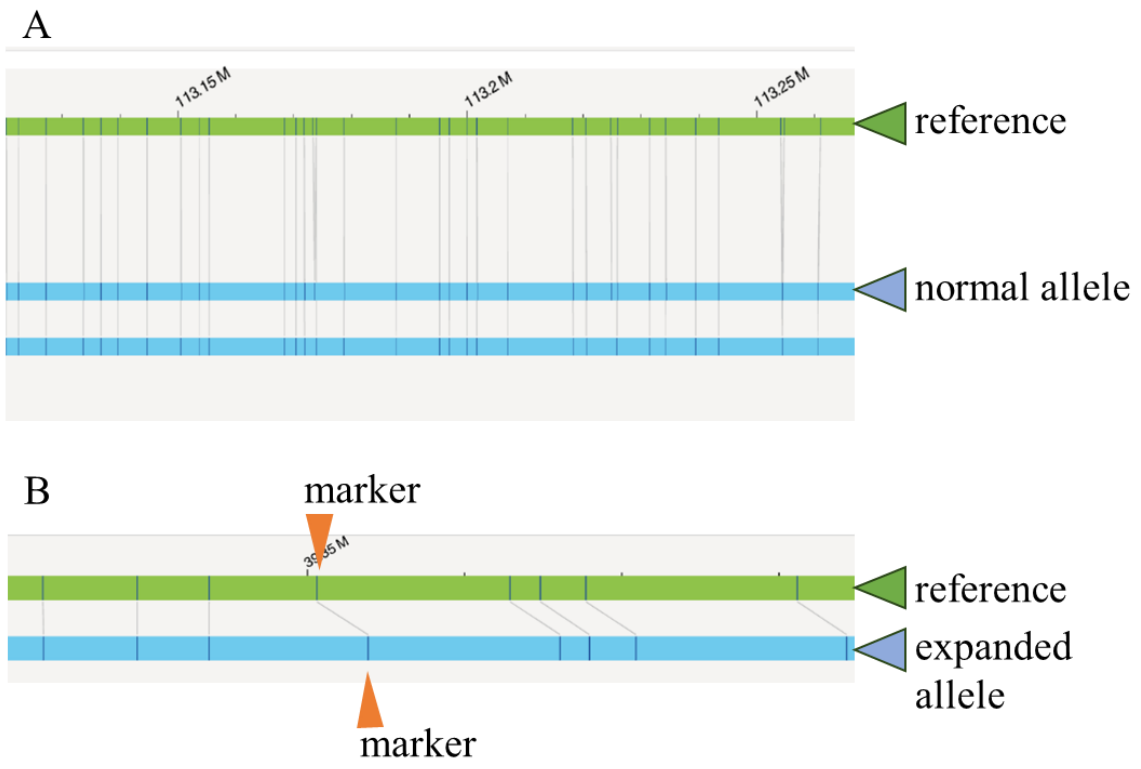


Figure 2.10. example of Bionano optical genome output. A) a section of DNA with all labels correctly aligning to the reference – no SV present. B) a section of DNA with a label at a distance from the reference label (shown in orange) – an expansion present.

2.2.10 Meiotic and somatic instability

Where consented, affected and unaffected first-degree relatives of probands were tested by Southern blotting totalling 27 families. Expansion size in *RFC1* locus was compared within the families to assess the stability of the AAGGG expansion during the transmission between the generations. Optical genome mapping was performed to assess the presence of somatic instability in affected (vermis, cerebellar hemispheres) versus unaffected tissues (frontal cortex, muscle, fibroblasts) of 4 patients carrying biallelic *RFC1* expansions. Blood-derived DNA

from a patient with C9orf72 GGGGCC expansion was also included as positive control for repeat instability.

2.2.11 Statistical analysis

Student's t-test or Mann-Whitney tests were performed for pair-wise comparisons of parametric or non-parametric variables, respectively. Linear correlation was measured with Paerson's coefficient. We analysed the correlation of the size of major (or larger), minor (or smaller) and both alleles with the age of neurological onset (cough excluded). Grade of correlation was defined as follows: $r < \pm 0.30$ = mild correlation; r from ± 0.30 to ± 0.70 = moderate correlation; $r > \pm 0.70$ = strong correlation. Plots and graphs were created with GraphPad Prism version 9.4.0 for Windows, GraphPad Software, San Diego, California USA, www.graphpad.com.

2.3 Results

2.3.1 Southern blotting optimisation

Successful optimisation of Southern blotting was implemented, and the improved protocol has been used in this thesis and contributed to numerous manuscripts and expansion discoveries (Scriba *et al*., 2020; Curro *et al*., 2021, Ronco *et al*., 2022, Dominik *et al*., 2023, Curro *et al*., 2024)

While five micrograms of good quality genomic DNA are still needed, bands are generally better visible, and more blotting are successful. Moreover, shorter exposure time for visualising the bands on fluorescent detection film can be used which leads to less background and higher quality image overall, we find that 10 minutes exposure produces clearest image (fig.2.11) (Dominik *et al*., 2020).

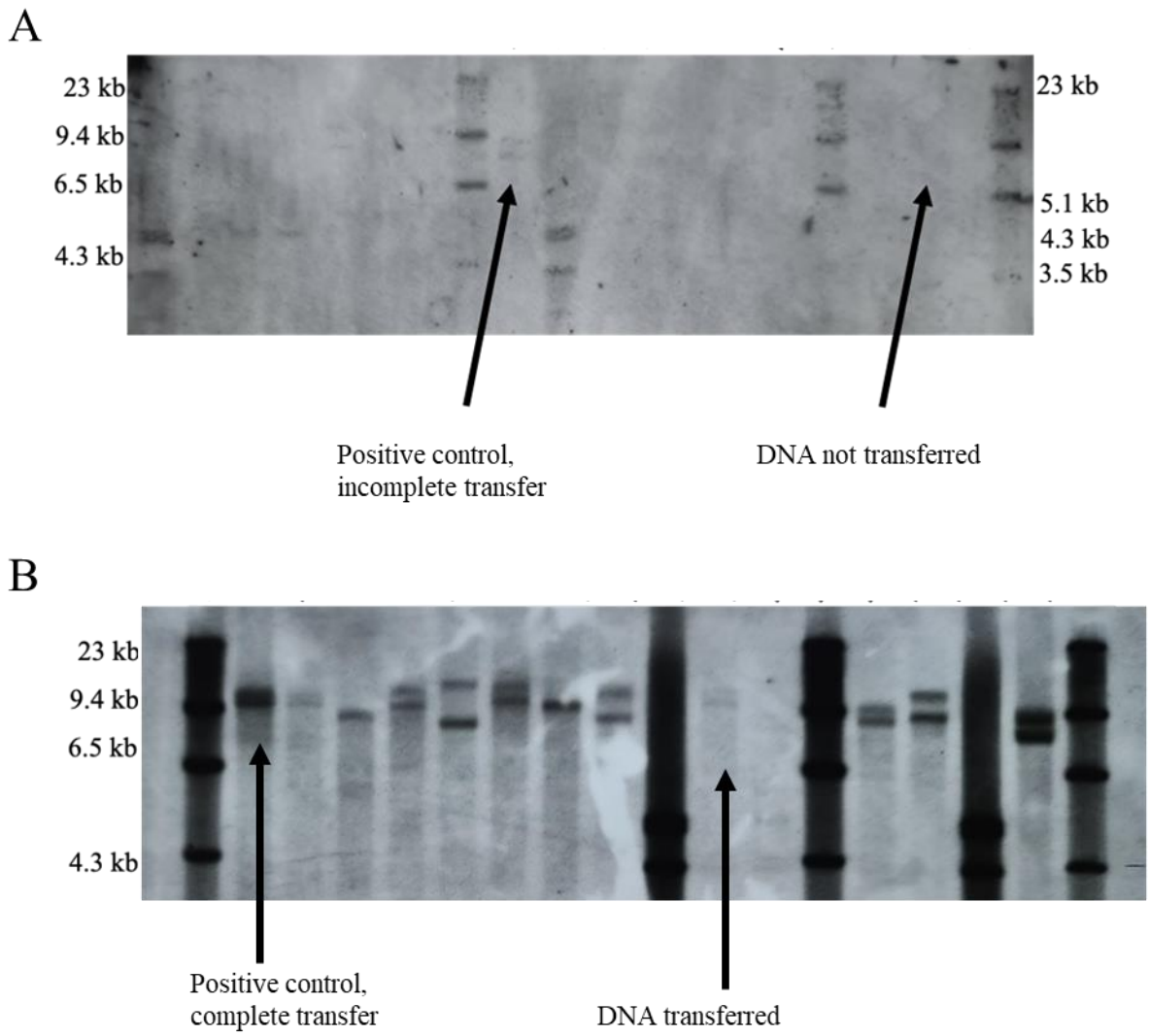


Figure 2.11 Southern blotting optimising. A) An example of Southern blotting with incomplete transfer where the positive control is poorly visible, and many DNA samples did not transfer onto the membrane. B) an example of Southern blotting after protocol optimisation where the positive control is fully transferred as seen by a strong band.

2.3.2 Southern blotting technical considerations

During the course of research contributing to my thesis, several technical issues resulted in failure of Southern blotting and the need for repetitive troubleshooting. An example of the technical hurdle I experienced was blank developed membranes with no trace of either sample bands or ladders. In order to troubleshoot the experiment, a thorough and well thought through strategy was needed due to the complexity and duration of the procedure (fig.2.12). Blank membrane with no ladder present could suggest problems at DNA transfer step from gel to membrane and this was assessed by visual inspection of the membrane post-transfer and presence of a blue colouration suggestive of transfer of the gel front therefore this issue was excluded. Another problem resulting in blank membrane could have been too stringent washes

and variable SSC buffers were assessed with different temperature combinations, however, this yielded no improvement. Finally, membranes were subjected to different saline-sodium citrate (SSC) concentrations (lanes 1, 3 and 5 with standard 2XSSC; lanes 2, 4 and 6 in 5XSSC) concurrently with UV light in different transilluminator machines (lanes 1 and 2 in the standard UV machine, lanes 3 and 4 in benchtop UV transilluminator and lanes 5 and 6 in (BioRad Gel Doc XR+) for purpose of DNA crosslinking on the membrane. It transpired that the UV machine that had been used previously became faulty and no UV light was produced resulting in no DNA crosslinking on the membrane and therefore subsequent washing off of the ladder and samples of the membrane. Lane 5 produced the strongest band therefore these conditions were taken forward into the next experiments.

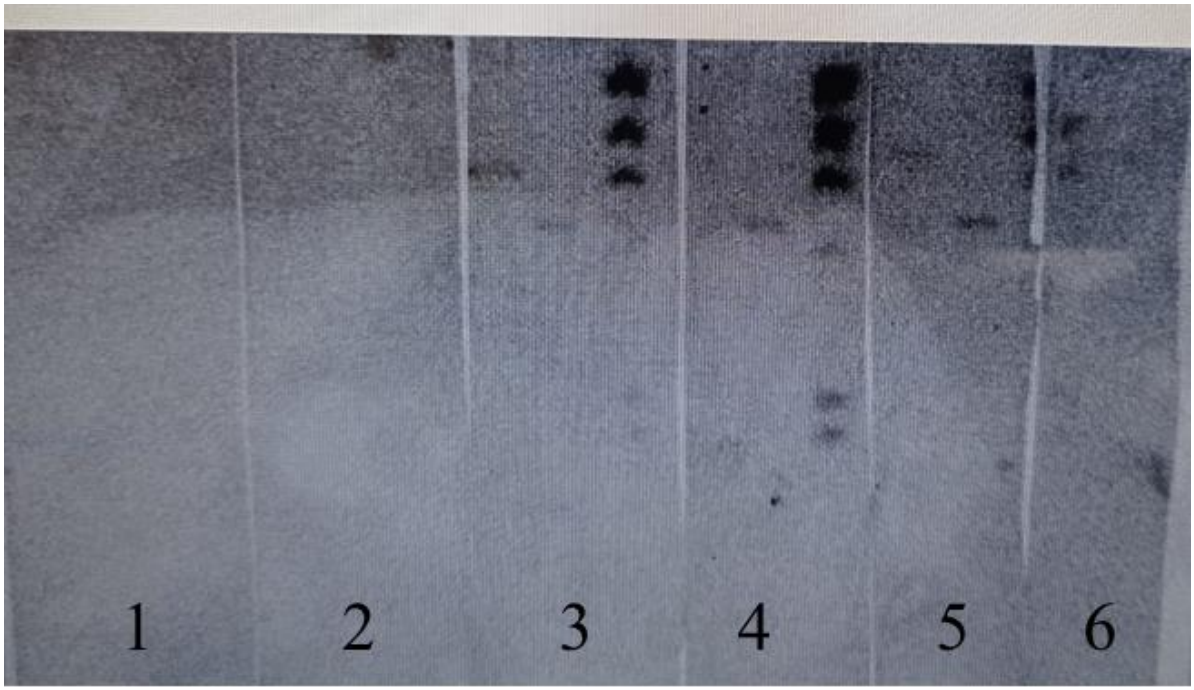


Figure 2.12. An example of Southern blotting troubleshooting. 6 separate strips of positively charged membrane were used for transfer of a control DNA and a marker. Lanes 1 and 2 were UV crosslinked in the standard UV machine and washed at different concentrations of SSC buffer (lane 1 standard 2XSSC and lane 2 5XSSC), there were no DNA nor ladders at detection. Lanes 3-6 were UV crosslinked in 2 different UV machines (lane 3 and 4 in benchtop UV transilluminator and lane 5 and 6 in (BioRad Gel Doc XR+) and washed with different concentrations of SSC buffer (lanes 3 and 5 with 2X SSC and lanes 4 and 6 with 5X SSC). DNA and ladders are visible in lanes 3-6 pointing to a malfunction of the UV machine used.

The above further highlights the complexity of Southern blotting as a technique and the need for specialised equipment and trained staff. Despite the technical hurdles, Southern blotting provides an invaluable information of sizes of patient alleles that can further be leveraged for discerning the role of repeat expansion size on age of onset of disease, disease progression and clinical characteristics as described in following sections of this chapter.

2.3.3 Genetic testing

I screened 1531 patients using flanking PCR and RP-PCR methods, and the patients with no amplifiable PCR product at flanking PCR, positive saw-tooth pattern for AAGGG RP-PCR and negative for AAAGG and AAAAG motifs were considered likely positive for biallelic AAGGG repeat. Furthermore, I contributed to standardized data collection and analysis of the entire study cohort of 2334 patients.

In total, we identified 556 patients likely positive for biallelic AAGGG repeat expansions and performed Southern blotting for 395 patient samples with sufficient DNA and with clinical information available. I performed Southern blotting for 315 patients. The data included in my thesis is based on the entire cohort of 2334 patients and published in Brain and fully described (Curro *et al*., 2024).

The presence of biallelic expansions was confirmed in 392 cases (99.3%). For the 3 patients who were not confirmed as carrying biallelic AAGGG repeat expansions, I performed Sanger sequencing and intermediate expansions of non-pathogenic repeat motifs of less than 100 repeats were observed and included $AAAAG_{exp}$, $AAAGG_{exp}$ and $AAAGGG_{exp}$ motifs.

2.3.4 Clinical details

Clinical data for 392 patients confirmed to carry biallelic expansions in *RFC1* by Southern blotting is available in table 2.5

Demographic		
N. of males (%), females (%)	195 (50%), 197 (50%)	
Positive family history	45 (11%)	
Current age (min-max)	70 years (42-90)	
Age at neurological onset (min-max)	54 years (25-80)	
Deceased	32 (8%)	
Symptoms	Last examination (N/total)	Age at onset
Unsteadiness	366/388 (94%)	56 years (30-80)
sensory symptoms	276/383 (72%)	55 years (25-75)
Dysarthria/Dysphagia	196/381 (51%)	64 years (30-85)
Oscillopsia	94/352 (27%)	62 years (36-81)
Chronic cough	267/358 (75%)	40 years (15-83)
Use of walking aid	203/379 (54%)	67 years (37-88)
Disease group	Patient number	
Isolated neuropathy	54 (14%)	
Complex neuropathy	131 (33%)	
CANVAS	195 (50%)	
Not assigned (incomplete clinical data)	12 (3%)	

Table 2.5 Demographic and clinical data of biallelic AAGGG patients confirmed by Southern blotting in our cohort

392 patients were confirmed to carry biallelic expansions with Southern blotting. Both genders were represented equally in the cohort. A positive family history for CANVAS-like symptoms was reported in 45 patients and the other 347 cases were sporadic. Age of onset of neurological symptoms (excluding cough) was 55 years for sensory symptoms, with unsteadiness as most common complaint between the patients (94%). Fig.2.13 shows the most common complaints and use of any walking aid at the last neurological follow up. Majority of the patients suffer from unsteadiness and sensory symptoms. Oscillopsia is the least common complaint but still present in one fourth of the patients. The cough that was present in 75% of the patients was the presenting symptom in half of the cases.

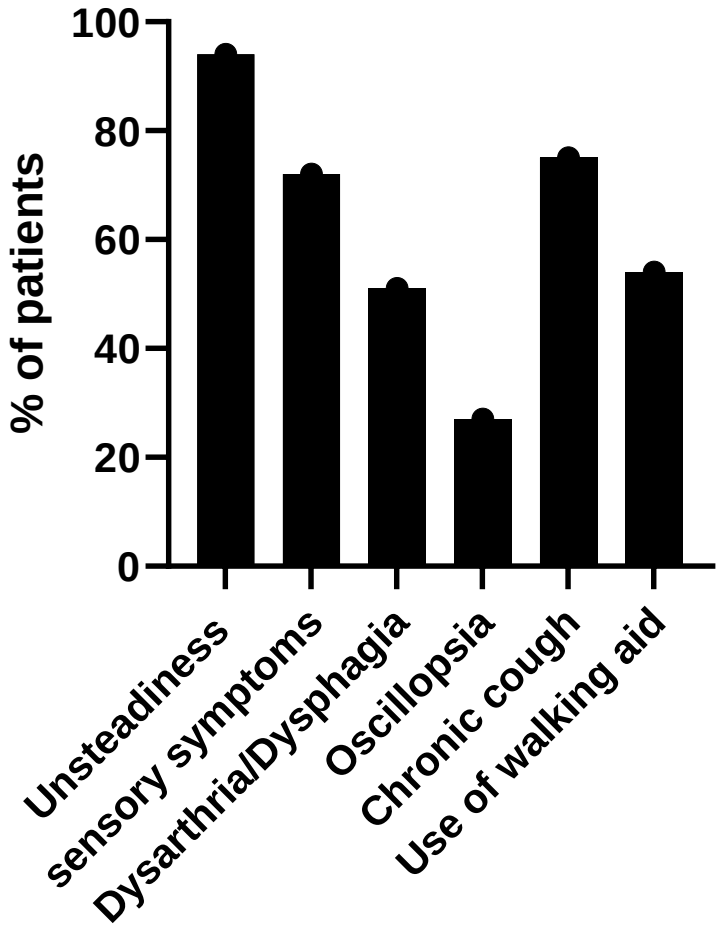


Figure 2.13 Patient symptoms at last examination. Most commonly reported symptom was unsteadiness (94%) followed by chronic cough (75%) and sensory symptoms (72%)

Patients could be divided into disease subgroups depending on their symptoms (fig.2.14), but full-blown CANVAS was still the predominant phenotype (50%). Importantly, all cases had sensory neuropathy.

Our cohort is of Caucasian ancestry, mostly of European descent, however, 18 patients originated from Turkey, and one patient each from Brazil, Iran, Iraq, Algeria and Lebanon.

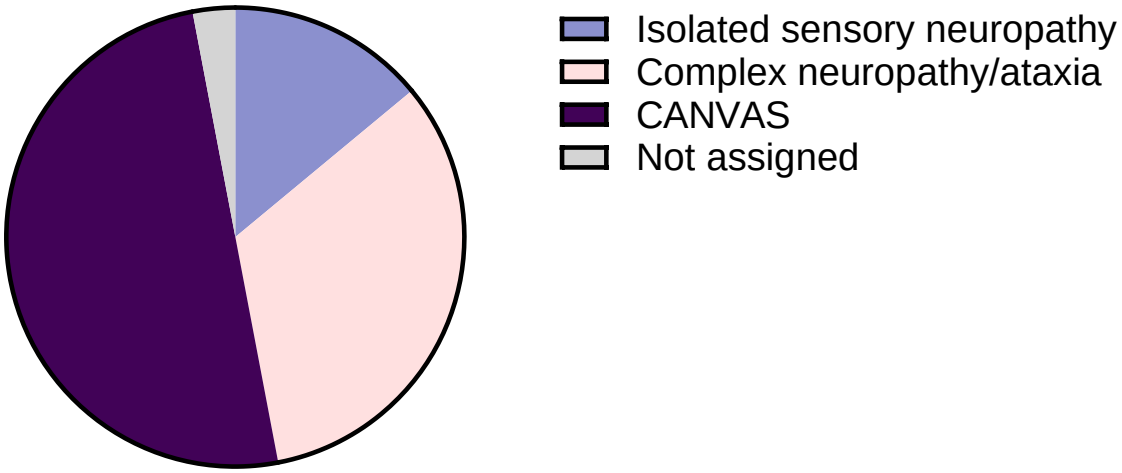


Figure 2.14 Patients could be divided into three subgroups depending on the symptoms. 50% of patients suffered from full blown CANVAS, followed by complex neuropathy/ataxia at 33% and isolated sensory neuropathy at 14% and 3% of the cases were not assigned a diagnostic category

2.3.5 Repeat expansions size and age of onset and disease phenotype

In total, 392 patients with biallelic AAGGG expansions in *RFC1* had a successful Southern blotting performed to total 784 alleles available for sizing. 36% of the patients (143 patients) had showed one band on Southern blotting suggesting alleles of the same or similar size, within the limits of detection of the technique.

We used the sizes of minor and major alleles as well as both in combination to investigate any possible correlations between age of onset of neurological symptoms and the expansion size. We observed that patients with larger expansions tended to have younger age of onset which was stronger for the minor allele (fig.2.15). Also, we investigated the size of alleles in disease subgroups isolated neuropathy, complex neuropathy/ataxia and full CANVAS.

Patients with smaller repeat expansions tended to have isolated neuropathy (minor allele=smaller allele=770 ± 260 repeat units; larger allele=1062 ± 364 repeat units) while people with complex neuropathy/ataxia or CANVAS had significantly larger alleles (complex ataxia: minor allele= smaller allele=1006 ± 324 repeat units, $p<0.001$; larger allele=1305 ± 515 repeat units, $p<0.001$ CANVAS: (smaller allele=1018 ± 329 repeat units, $p<0.001$; larger allele= 1294 ± 497 repeat units, $p<0{\cdot}001$). There was no significant difference between CANVAS and complex neuropathy/ataxia.

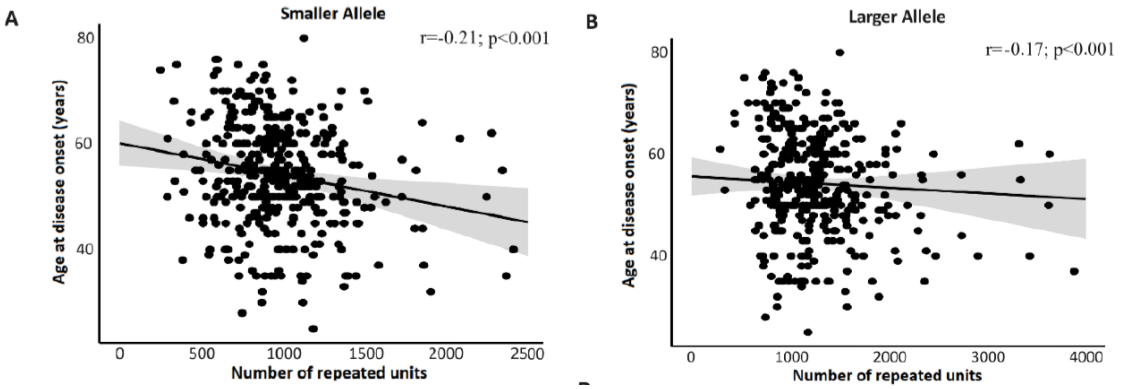


Figure 2.15 Age of onset of neurological symptoms and the expansion size for minor and larger allele. The scatter plots illustrate the strength and the direction of the correlation between the age at neurological onset of the disease (y-axis) and the repeat size of the smaller or the larger allele (x-axis). Pearson's correlation. Adapted from Curro *et al.*, 2023.

2.3.6 Meiotic and somatic instability of AAGGG repeat expansions in *RFC1*

27 families were available for assessment of stability of AAGGG expansion in transmitted allele. In total, 69 affected and unaffected individuals were explored and included 27 probands, 22 siblings, 18 offsprings and 2 parents. An example of Southern blotting within families is available in fig.2.16 and an intrafamilial repeat expansion number between the proband and family member in figure B. AAGGG appears stable between siblings and generations (r^2=0.95), with a median intra-familial variation of 25 repeats (*min max*=-250/+510). Expansions and contractions of expanded alleles occurred with the same frequency and there was no evidence of larger expanded alleles in offsprings.

We compared the repeat size in RFC1 locus in vermis – the most affected tissue in RFC1 disease, cerebellar hemispheres, frontal cortex and peripheral tissues (fig. 2.16 C&D) and in C9orf72 patient blood sample as a positive control for somatic instability. We saw a variation in size of the repeats between −97 and +190 repeats (−5%/+7%) compared to mean size. Furthermore, mean dispersion of the repeat length was ±1.7% for vermis, ±2% for cerebellar hemispheres and ±2.7% for frontal cortex, as opposed to a dispersion of ±36% in an individual carrying C9orf72 expansion. This suggests limited somatic instability in the affected and unaffected bulk tissues.

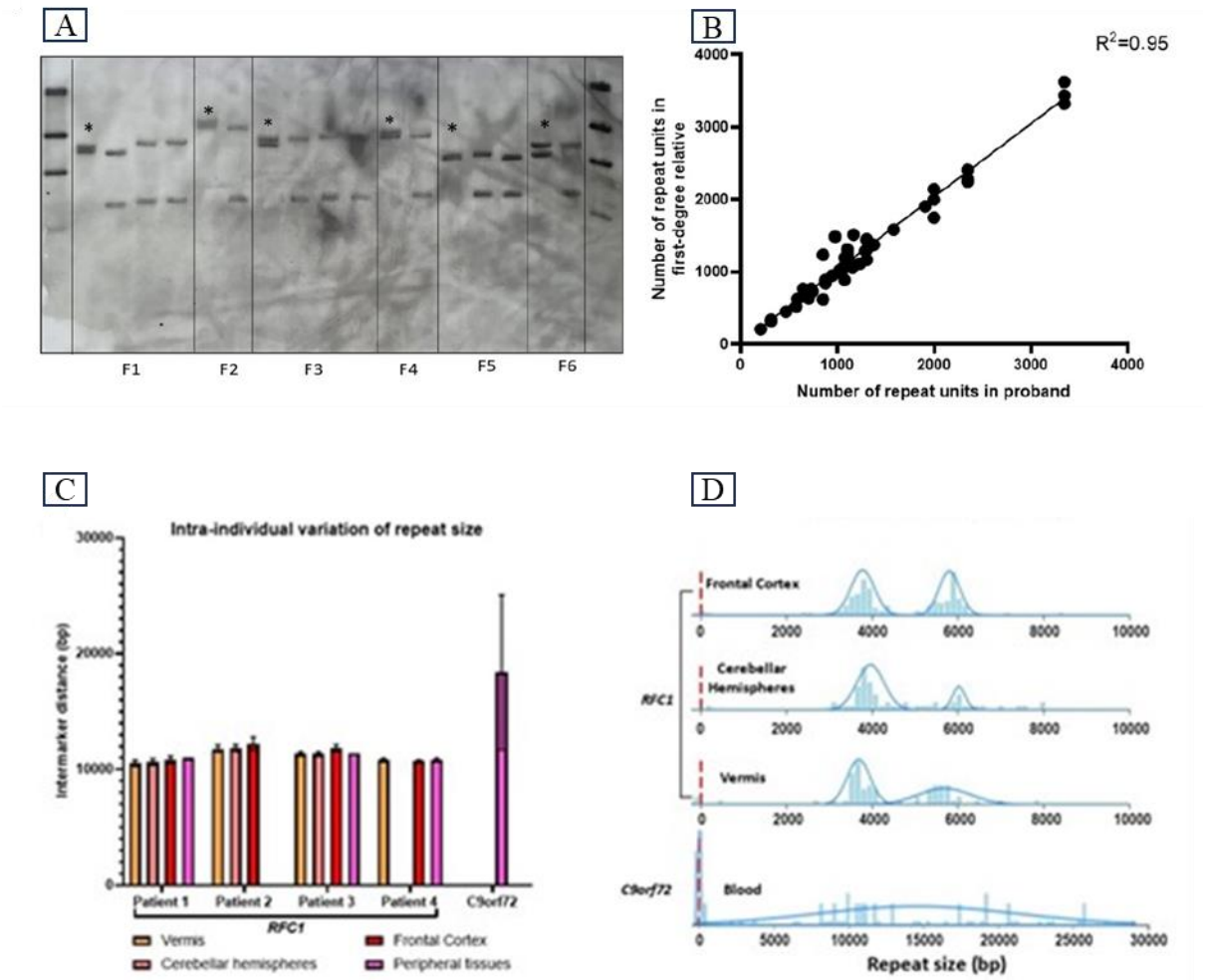


Figure 2.16 limited meiotic and somatic instability of the AAGGG repeat expansions. A) Representative Southern blotting showing 6 probands (with asterisks) and their unaffected family members; and B) a correlation plot of the repeat expansion sizes in the 27 families where each dot represents a meiotic event. C) The repeat size among different brain areas and peripheral tissues of four patients with *RFC1* biallelic expansions and in one patient with *C9orf72* expansion and an example of patient RFC1 repeat expansion size measured in different tissues by OGM Adapted from Curro *et al*., 2023

2.3.7 Validation of a new technology for repeat expansion testing in *RFC1* disease

We measured the size of expanded alleles using both optical genome mapping and Southern blotting in 17 cases (table 2.6 and fig.2.17) to compare the resulting sizes between the techniques. This work has been published in Biomolecules, an MDPI Journal (Facchini*, Dominik* *et al*., 2023)

Optical genome mapping (OGM) is a new technology that allows for visualizing fluorescently labelled DNA with a camera system (Saphyr) provided by Bionano and subsequent detection of structural variants larger than 500 base pairs.

All CANVAS samples were confirmed to carry biallelic *RFC1* repeat expansions with both methods (Table 2.6). Patients 6, 9 and 17 showed a presence of only one thick band on the corresponding Southern blotting suggesting alleles of the same or similar sizes, however, Bionano OGM detected two distinct alleles in those cases. An example of OGM better resolving 2 alleles of similar size is presented in fig.2.18. In only one case, OGM showed a presence of homozygous alleles while SB showed 2 distinct expansion sizes. 5 cases showed homozygous alleles using both the techniques. In addition, 2 control samples were subjected to OGM analysis, confirming the absence of biallelic expansions (Control 1 has one expanded allele; Control 2 has two unexpanded alleles). An excellent linear correlation was observed between the two methods (fig.2.17), with r2 = 0.97. However, the linear coefficient is 0.62 [0.58–0.66] at 95% confidence interval (C.I.), and the intercept is 232 [181–226] at 95% C.I. suggesting under or over- estimation of allele sizes by one of the methods.

Patient	SB Allele 1	SB Allele 2	OGM Allele 1	OGM Allele 2
Pt 1	765	1242	677 ± 41	955 ± 45
Pt 2	598	1035	622 ± 34	841 ± 36
Pt 3	989 (Homozygous)		894 ± 29 (Homozygous)	
Pt 4	1127	1593	866 ± 48	1182 ± 70
Pt 5	1447	1838	1017 ± 57	1180 ±40
Pt 6	917 (Homozygous)		664 ± 24	730 ± 22
Pt 7	1400 (Homozygous)		1223 ± 36 (Homozygous)	
Pt 8	991 (Homozygous)		829 ± 53 (Homozygous)	
Pt 9	1185 (Homozygous)		880 ± 46	943 ± 29
Pt 10	1256	4746	1055 ± 79	3226 ± 163
Pt 11	249	810	333 ± 20	831 ± 35
Pt 12	724 (Homozygous)		792 ± 63 (Homozygous)	
Pt 13	294 (Homozygous)		406 ± 32 (Homozygous)	
Pt 14	640	794	652 ± 40	759 ± 24
Pt 15	605	714	640 ± 51 (Homozygous)	
Pt 16	794	2386	745 ± 51	1646 ± 97
Pt 17	810 (Homozygous)		582 ± 35	654 ± 24
Control 1	/		−4 ± 26	450 ± 22
Control 2	/		−6 ± 30 (Homozygous)	

Table 2.6 Estimated sizes of the repeat expansions (number of pentanucleotide repeats). In OGM, repeat size is indicated as mean ± standard deviation of the Gaussian. Highlighted in grey are the patients where OGM, unlike SB, could better discriminate the size of the two expanded alleles. Adapted from Facchini*, Dominik* *et al*., 2023.

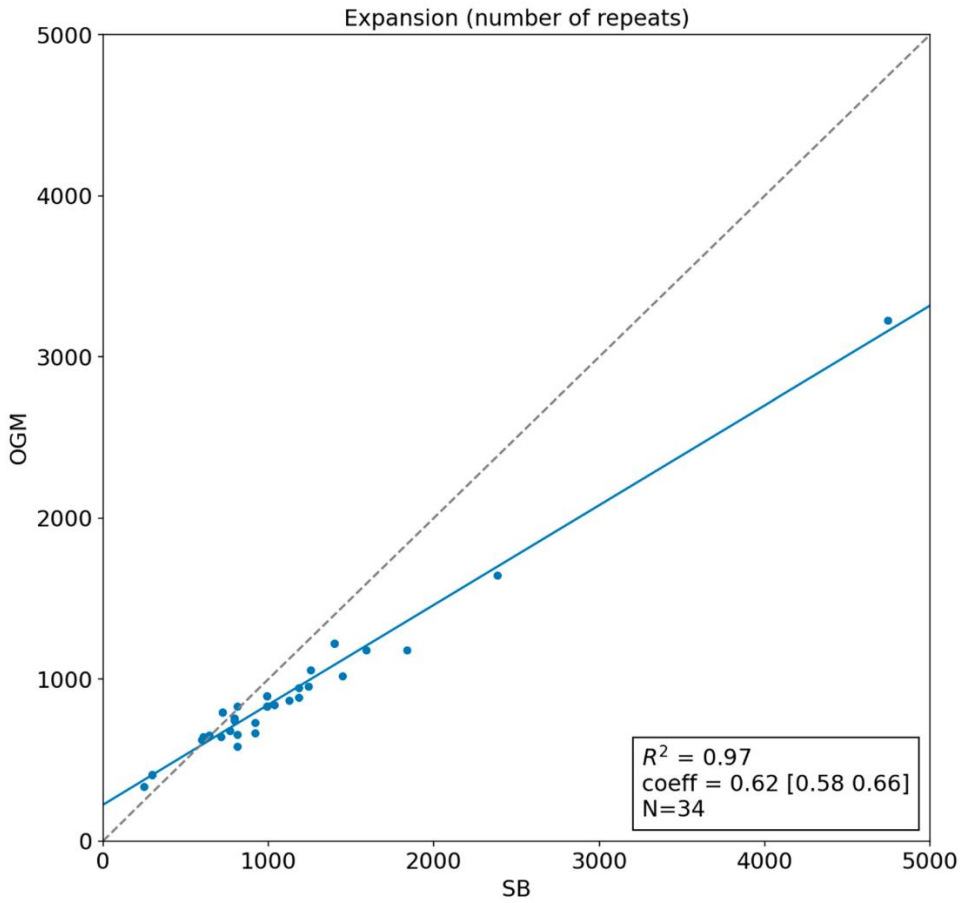


Figure 2.17 Expanded allele size comparison between Bionano and Southern blotting. Overall satisfactory comparison was observed between the two methods although Southern blotting tended to overestimate the size of very large alleles.

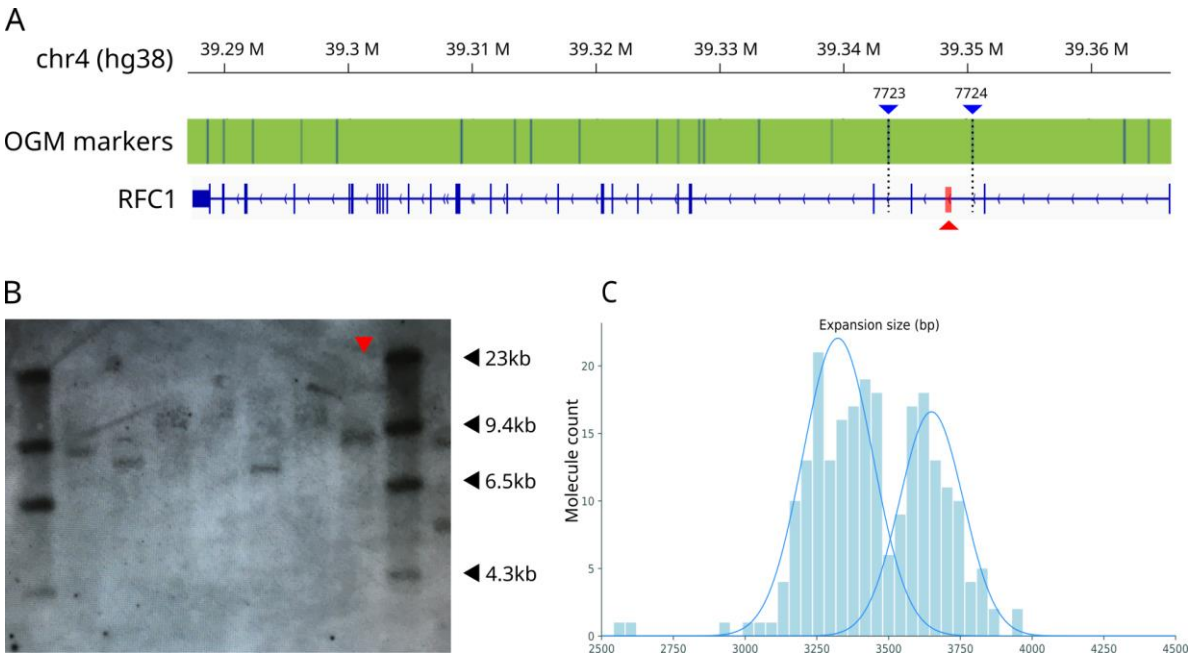


Figure 2.18 (**A**) Bionano OGM markers used for the analysis. The red triangle area indicates the position of the repeat expansion inside the second intron of *RFC1* (blue arrows point in the coding direction of the gene). The blue triangles indicate the position of the markers flanking the repeat (markers 7723 and 7724) (**B**) Optical genome mapping for Pt 6. Two alleles are observed as Gaussian components of size 664 and 730 repeats (3322bp and 3648bp, respectively) (**C**) Representative example of Southern Blotting plot. For Pt 6 (indicated by the red triangle), only one band is visible, corresponding to an expansion of 917 repeats (4585 bp). Adapted from Facchini*, Dominik* *et al*., 2023.

Visual comparison of allele size for both methods is available in figure 2.19. OGM sizes are represented as molecule size distribution with the estimated Gaussian components. The corresponding Southern blotting are presented for each patient and their alleles marked with arrows.

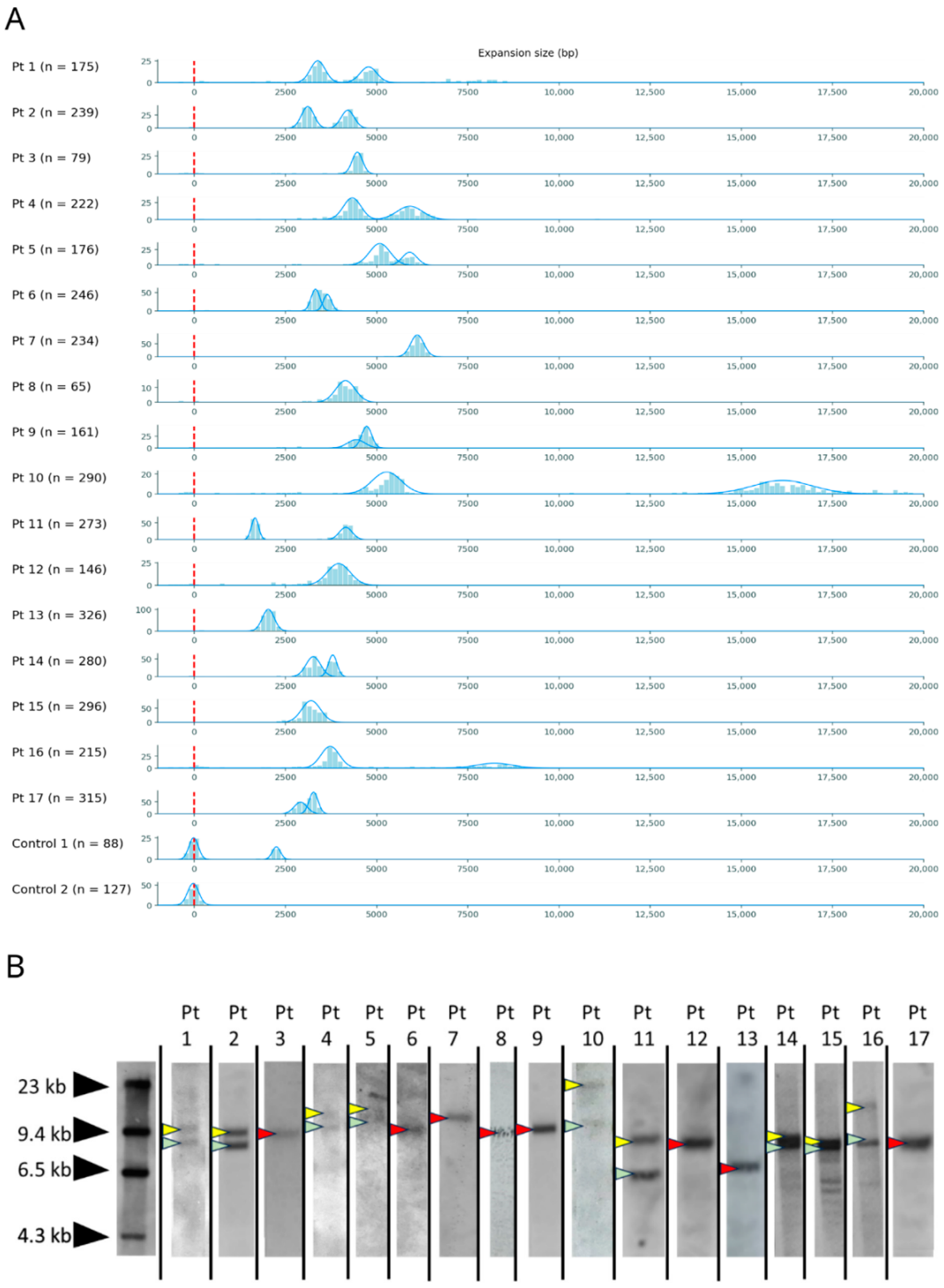


Figure 2.19 (A) OGM molecule size distribution for all samples, with estimated Gaussian components. On the vertical axis the molecule count is reported. The vertical dotted red line corresponds to a non-expanded allele. For each sample, we report the total number of observed molecules in parenthesis. (B) SB images for all patients. Arrows point to the alleles visible on Southern Blotting; yellow and green when two alleles of distinct sizes are seen and red when two alleles of the same size are seen. Adapted from Facchini*, Dominik* *et al.*, 2023.

2.3.8 Technical Considerations of Southern Blotting and Optical Genome Mapping

SB relies on large quantities (5µg) of high-quality and purity DNA. SB is compatible with most DNA extraction methods, thus facilitating sample processing and shipping of extracted DNA from collaborators across the globe. In comparison, OGM can only be performed on ultra-high molecular weight DNA fragments (>150 Kbp), which requires a bespoke extraction method using the Bionano extraction kit from fresh or snap frozen blood or cell pellets. Hands-on processing time at the bench is 4 working days for SB and 2 working days for OGM, followed by Saphyr imaging and automatic data collection.

SB size estimation relies on comparison to a ladder tract. OGM relies on fluorescent labels which bind to specific 6 bp DNA motifs (CTTAAG) present in the genome at an average of 20 times per 100 Kbp.

In addition to good technical skills, necessary for both methods, OGM requires computer literacy for size estimation in the online Bionano Access analysis platform, or to perform custom analysis (Facchini*, Dominik* *et al*., 2023)

2.4 Discussion

Biallelic *RFC1* expansions represent a common cause of late-onset ataxia and sensory neuropathy. However, implementing a diagnostic test remains challenging due to the disease heterogeneity, complexity of the molecular methods and the need for specific laboratory set up especially for Southern blotting (SB) confirmation which is often not available in diagnostic laboratories. Indeed, SB is a cumbersome technique and my work contributed to the optimisation of the protocol which allowed for better DNA transfer onto the positively charged nylon membrane and in turn better visualisation of the expanded alleles in general. This impacted positively the clinical studies based on allele sizing with Southern blotting such as the age of onset study, described in this chapter, where I performed Southern blotting for 315 patients; identification of novel pathogenic configurations confirmed with Southern blotting which I talk about in the next chapter; and various collaborations relying on Southern blotting (Scriba *et al*., 2020; Curro *et al*., 2021, Ronco *et al*., 2021, Dominik *et al*., 2023, Curro *et al*., 2024); and importantly, it allowed for the patients to receive a research diagnosis of their disease. Moreover, in 2023, the PCR screening procedures used for *RFC1* have been adapted and implemented by the diagnostics laboratories in the UK adding additional translational value to the research of Dr Cortese's group that I am a part of.

2.4.1 *RFC1* repeat expansion size predicting age of onset, disease progression and clinical variables

In this chapter, I described a collaborative work that I contributed to by screening 1531 patients with PCR techniques and 315 patients with Southern blotting, liaising with colleagues for standardised data collection and measuring allele sizes.

This multicentre study with an international cohort of patients is the largest study on *RFC1* repeat expansions to date, and it leveraged the data on expansion size of the AAGGG repeat expansion in patients with biallelic expansions and allowed to assess the impact of the allele sizes on age of onset of the RFC1 disease, progression and clinical phenotype.

CANVAS is a complex disease which can manifest with a typical triad of symptoms involving cerebellum, sensory neuron and the vestibular system. In our large cohort of patients, we confirmed that all patients with the expansions suffer from sensory neuropathy which is in line with previous findings (Curro *et al*., 2021, Cortese *et al*., 2020) and no patients have an isolated cerebellar ataxia. We were able to group the patients into three subgrups depending on their clinical presentation and they included full blown CANVAS with 50% patients, 33% of the patients had complex neuropathy with ataxia and 14% had sensory neuropathy at the last

available examination. These findings together with the repeat expansion sizes measured with Southern blotting, allowed for correlation of the size with disease severity. We found that patients with isolated sensory neuropathy had smaller expansions compared to the other subgroups. There was no significant difference of expansion sizes between patients with complex neuropathy and full-blown CANVAS. These findings suggests that the repeat expansion size can act as a modifier of the disease phenotype and sensory neurons are probably more susceptible to the AAGGG repeat expansion than other tissues.

The repeat expansion size in *RFC1* also influences the onset of neurological symptoms. Patients with larger expansions tended to have younger age of onset and the most common complaint at the disease onset in our cohort was unsteadiness and sensory symptoms. It is important to note that recollection bias may have influenced patients to report symptoms later in disease progression, for example, sensory symptoms such as pins-and-needles may be overlooked at first and only considered when progressed and more severe. Interestingly, as many as 70% of patients suffer from chronic cough which often precedes other symptoms by as many as 20 years, and it was the presenting symptom in half of our cohort, however, it was not considered for the purpose of the correlation.

In addition, the data from 27 families with affected and unaffected individuals tested and sized by Southern blotting demonstrated that AAGGG repeat expansion in *RFC1* appears meiotically stable and no large expansions or contractions are observed in vertical transmission of the allele. Moreover, we specifically interrogated the affected cerebellum and unaffected tissues and the data obtained from patient bulk tissues do not support the existence of significant instability of the repeat size in different tissues, however, a variation of the repeat size at single cell level cannot be excluded. Repeat expansion disorders often show a degree of germline instability, where an expanded allele may become larger when passed onto an offspring, and somatic instability, where the expanded allele may be larger in the affected tissues compared to unaffected tissues of the same patient, these however appear not to be the case in RFC1 disease. This is the first large multi-centre study of *RFC1* repeat expansions and their influence on disease onset and the progression. Indeed, the findings are in line with other neurodegenerative repeat expansion diseases where larger repeat expansions are shown to cause more severe disease with earlier age of onset such as in recessive FRDA or dominant *C9orf72* or DM1 (Filla *et al*., 1996: Santoro *et al*., 2000: van Blitterswijk *et al*., 2013; Peric *et al*., 2021). The findings have several implications which include better evidence for counselling the patients with biallelic *RFC1* expansions and ability to better prognose the disease

progression as well as possibility to identify patients with a higher risk of developing more complex and debilitating symptoms.

Indeed, this work further added to the ever-evolving picture of the *RFC1* disease spectrum by identifying more expanded alleles with smallest pathogenic AAGGG seen so far at 250 repeats (the smallest previously reported being 400 (Cortese *et al*., 2019)) and so far, no patient carrying AAGGG expansions in biallelic form that are not pathogenic – there are no premutation alleles and AAGGG is fully penetrant as opposed to some other repeat expansion disorders shown in fig.2.20

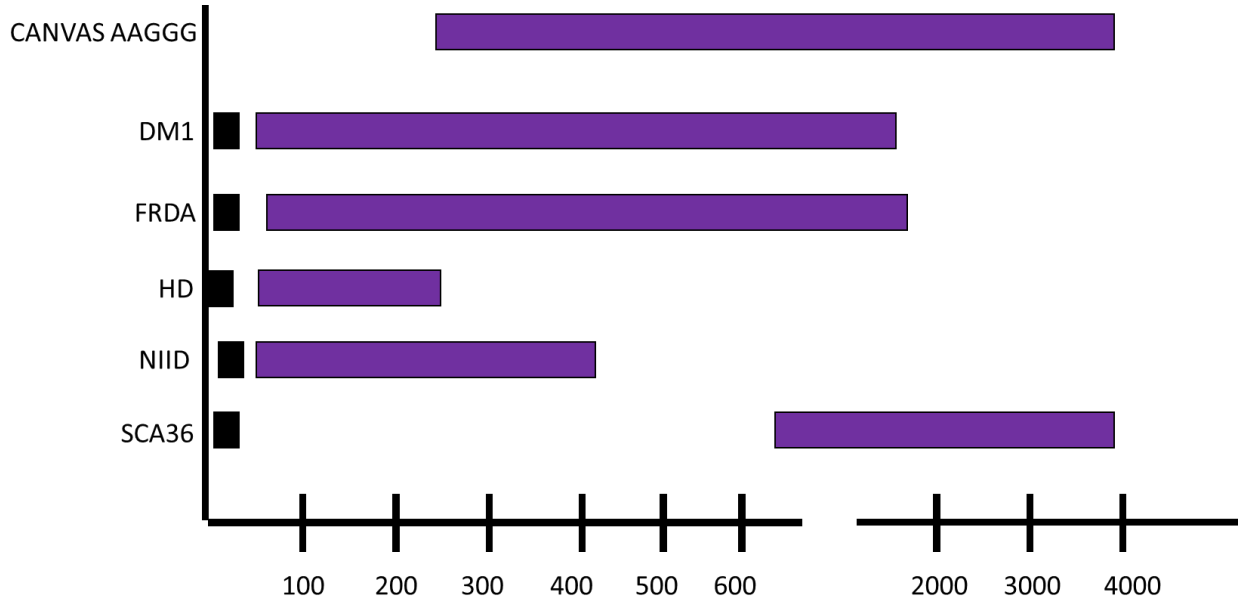


Figure 2.20 A visual representation of repeat expansion sizes in AAGGG CANVAS and five other neurological conditions caused by repeat expansions. Purple indicates pathogenic expansions and black indicates pre-mutation expansions. To date, the smallest observed AAGGG repeat expansion in CANVAS is 250 repeats and no premutation alleles are observed. Other neurological conditions such as DM1, FRDA, HD, NIDD and SCA36 have a premutation allele where expansions are present but are not yet sufficiently expanded to cause disease. There permutations may become expanded to pathogenic sizes when transmitted vertically to offspring.

2.4.2 Repeat expansion sizing in *RFC1*

The need for sizing of the alleles is also highlighted in this chapter. Whilst screening with PCR methods may indicate presence of biallelic expansions, those methods have many limitations, inability to size the repeat being one of the most important ones. Southern blotting has been a gold standard technique for sizing the expansions and used in this study, however, it also has its limitations. I was a part of the team who established optical genome mapping at Institute of Neurology, University College London and participated in validating the method as an alternative to Southern blotting.

Optical genome mapping is a new technology allowing for simultaneous detection of structural variants withing the entire patient genome. I processed 10 patient samples with optical genome mapping and altogether, we validated the technology on 17 blood samples with known biallelic *RFC1* expansions that have been sized with Southern blotting.

We compared the repeat sizing between SB and OGM and showed a very good linear correlation of the two techniques. We noticed a deviation from the expected identity function in the regression, which is accounted by a systematic error either in the SB or in the OGM method, particularly for the expanded alleles over ~1000 repeats. This could either be due to overestimation of repeat size with SB or underestimation with OGM. SB relies on gel electrophoresis to resolve large fragments of genomic DNA. Possible formation of secondary structures by the repeats, slowing down the migration during electrophoresis, could lead to an overestimation of the repeat lengths. Moreover, due to the necessity of a visual comparison with a logarithmic scale, estimation of the allele size is increasingly imprecise for larger fragments, and it often cannot resolve similarly sized alleles resulting in a single band (Facchini*, Dominik* *et al*., 2023).

On the other hand, OGM may underestimate expansion size by taking into account kinked DNA molecules during imaging, leading to the underestimation of expansion size.

Moreover, OGM, unlike SB, was able to distinguish two alleles of similar size in 3 out of the 17 patients (Pt 6, Pt 9, Pt 17) while in one case (Pt 15), the presence of two distinct alleles was suggested via SB, but only one component was detected with OGM (table 2.6). Overall, OGM improved the allele sizing resolution in 4/17 (24%) samples.

An additional advantage of OGM is the possibility to screen for SV as well as large expansions (>500 nt) in the entire patient's genome in parallel to *RFC1* testing.

Both techniques require good technical skills, specific laboratory setups and special sample storage and transport considerations. However, advantages of OGM include a short response time (in ideal conditions, approximately 10hours hands-on time for DNA isolation and DNA labelling, overnight homogenisation of ultra-high molecular DNA, 8hours of run time at 100X coverage and 24hours for automated data collection), higher accuracy and high-throughput output (Facchini*, Dominik* *et al*., 2023).

A known limitation of both OGM and SB is that they do not provide any information on the repeat sequence and need to be complemented with PCR, short or long-read sequencing. This

is particularly true in cases with typical CANVAS symptoms but only heterozygous expansion where a truncating variant could be present in trans with the expansion, or in cases with suspected configuration motifs different to canonical pathogenic AAGGG (both described in chapter 3).

2.4.3 Beyond CANVAS

Since the discovery of *RFC1* expansions as causative of CANVAS in 2019 much evidence has been gained that the disease has been underdiagnosed. Increasing numbers of patients being tested for the expansions have been adding to the knowledge of *RFC1* repeat expansions and our work shows that these expansions can cause a variety of phenotypes ranging from isolated sensory neuropathy, sensory neuropathy and ataxia and full-blown CANVAS. Chronic cough and dysautonomia were described in the first cohorts studied, however, in the recent years it is becoming increasingly apparent that other features, beyond CANVAS, may be present in patients with *RFC1* repeat expansions. For example, a patient initially diagnosed with Sjögren syndrome was found to carry biallelic pathogenic *RFC1* expansions (Kumar *et al*., 2020). Conversely, to date we have not encountered a patient with pure cerebellar ataxia harbouring biallelic AAGGG expansions and this was further shown in an independent patient cohort where 54 patients with idiopathic ataxia were tested (Hadjivassiliou *et al*., 2024).

With such a broad spectrum of signs and symptoms, it may be challenging to decide whether a patient should be tested for *RFC1* expansions, however, anyone with an unexplained sensory neuropathy may benefit from *RFC1* screening after exclusion of acquired causes. The table below presents typical and atypical RFC1 spectrum disorder features described thus far.

Typical features	Atypical features (RFC1 spectrum disorder unlikely)
Progressive sensory neuropathy Chronic cough Altered vestibular ocular reflex and visually enhanced vestibulo-ocular reflex Gaze-evoked nystagmus, broken pursuits Dysarthria and dysphagia (more advanced stages of the disease)	Absence of sensory neuropathy Presence of motor involvement Early age of onset Rapid progression Prominent dysautonomia

Table 2.7 Typical clinical features in patients with RFC1 spectrum disorder and atypical features where diagnosis of RFC1 spectrum disorder is unlikely Adapted from Cortese *et al*., 2022

2.4.4 Limitations and future horizons

The main limitations of the work described in this chapter include: retrospective nature, where the recollection bias of onset of neurological symptoms in patients might have resulted in reporting an older age than when the symptoms actually developed; large multi-centre nature of the study resulted in differences of collected information and variable clinical investigations between centres; large amounts of good quality patient samples needed, which resulted in some patients not being included in the Southern blotting sizing analysis; unavailability of WGS data for most of the patients which could potentially uncover pathogenic variants in other genes; and finally but importantly the reliance on time consuming and somewhat imprecise testing methods where a Southern blotting takes four days of work and might be unsuccessful and the sizing relies on comparison to a ladder tract and a logarithmic scale that may add some degree of human error.

The availability of more modern technologies for repeat expansions is improving with optical genome mapping described in this chapter and long read sequencing technologies further explored in the next chapter, however, there is still work necessary for those technologies to be more accessible, precise and cost effective.

2.5 Conclusion

In conclusion, this study explored the relationship of size of the AAGGG expansions in *RFC1* locus with the age of onset of the disease, the disease progression, and its severity. The data showed an inverse coloration of AAGGG repeat expansion size and disease severity, younger age of onset of neurological symptoms in patients with larger expansions.

In addition, exploring optical genome mapping technology and traditional southern blotting for measurement of repeat expansion sizes, this study showed that OGM appears as a valid alternative to SB for the detection and sizing of *RFC1* expansions, along with genome-wide assessment of structural variants and other large repeat expansions, which could support its use in a diagnostic setting.

CHAPTER 3. Genetic heterogeneity of RFC1 disease spectrum and its implications on laboratory testing

3.1 Introduction

3.1.1 Heterogeneity of RFC1 disease spectrum

Around 82%-97% of individuals with clinical CANVAS have biallelic AAGGG expansion however, few patients with full CANVAS phenotype do not test positive for the biallelic AAGGG repeat expansion and in some cases carry only one AAGGG expanded allele (Cortese *et al*., 2022; Dominik *et al*., 2023). At the discovery of *RFC1* repeat expansions in 2019, Cortese *et al*. had already shown that the locus is polymorphic with AAAAG, AAAGG and AAGGG expansions identified that differ from normal allele of $AAAAG_{11}$. Also in 2019, using a bioinformatic approach Akcimen *et al*., further added to the then, short list of pentanucleotide motifs found in *RFC1* locus by identifying AAGAG and AGAGG, conformations of unknown pathogenicity.

However, majority of confirmed cases have been of European ancestry and various studies of the locus in different populations have added to the knowledge of *RFC1* expansions. Indeed, first CANVAS testing in populations of New Zealand Māori and Cook Island Māori revealed that those patients suffering with CANVAS had a novel, mixed allele of $(AAAGG)_{10–25}(AAGGG)_n$ and their phenotype did not differ from the described European CANVAS sufferers (Beecroft *et al*., 2020). Further, an Asian-Pacific cohort was screened for *RFC1* expansions by Scriba *et al*., and novel pathogenic ACAGG motif was seen in 3 individuals. In collaboration with the group, we performed Southern blotting for one of the individuals where ACAGG motif has been discovered and saw large, expanded alleles.

This evidence suggested that further genetic heterogeneity may be present in CANVAS and the disease spectrum and a list of likely non-pathogenic motifs as well as motifs of uncertain pathogenicity available when we embarked on our study is shown in fig.3.1.

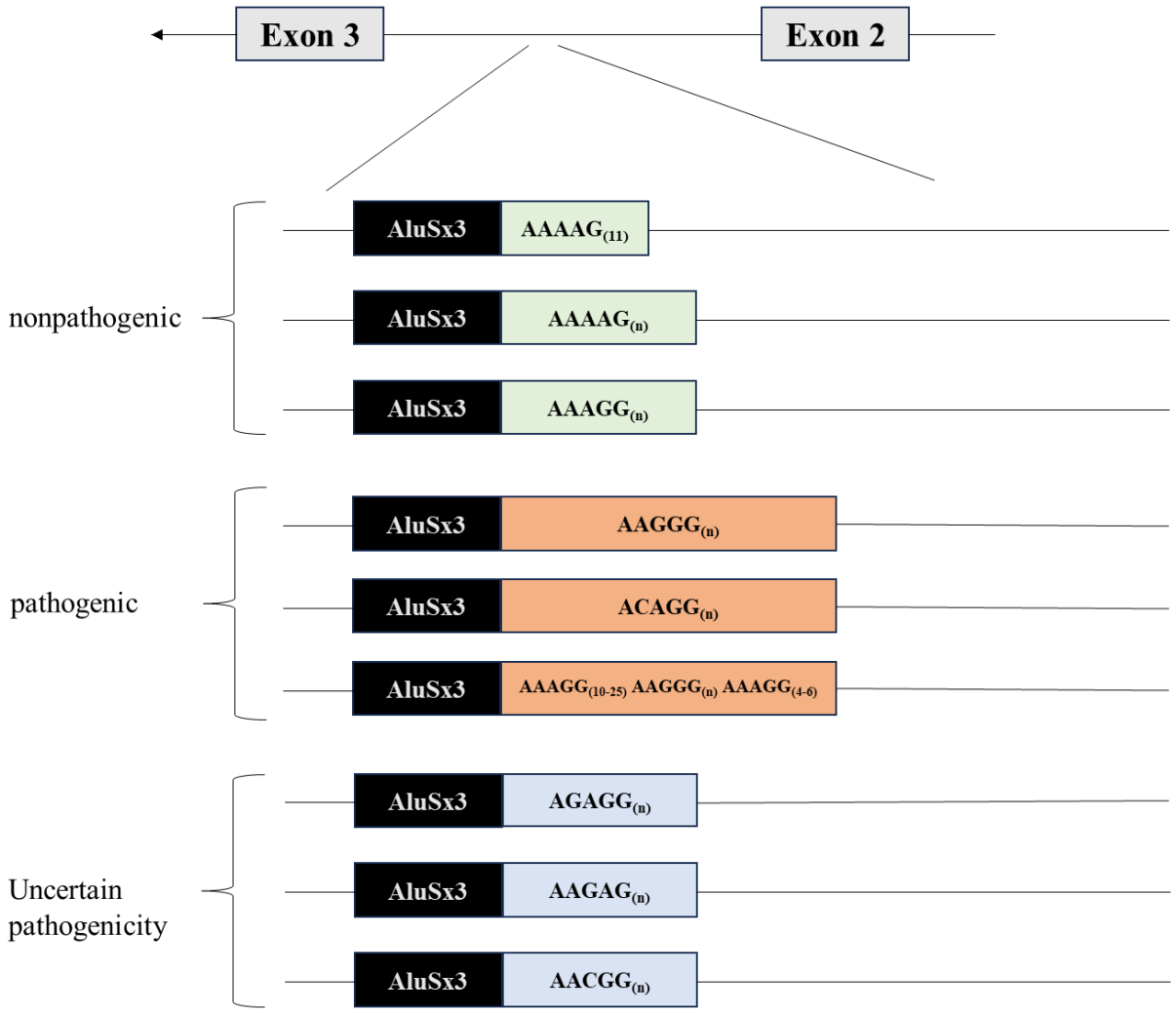


Figure 3.1. Genetic heterogeneity of *RFC1* repeat locus described up to 2020. Non-pathogenic expansions are shown in green and include the normal, reference allele $AAAAG_{11}$. The pathogenic alleles shown in orange include the common expansion of $AAGGG_{exp}$. Expansions of uncertain pathogenicity are shown in blue. Adapted from Davies *et al*., 2022

3.1.2 Sequencing methods in repeat expansion research

Sequencing of larger expansions can be challenging, prone to errors and might even not be possible using methods such as Sanger or short read sequencing (Cortese *et al*., 2019; Efthymiou *et al*., 2016; Dominik *et al*., 2023).

Read length limitation of short read whole genome sequencing (WGS) and their susceptibility to errors mean that we and other groups are increasingly using targeted long read technologies for looking into expansion disorders which can resolve large and complex repeat expansions and not only identify novel repeat expansion sequences but also recognise sequence interruptions. Undeniably, long read sequencing is not without errors and challenges. Samples for long read sequencing platforms have to be prepared using specific protocols that allow for retrieval of unfragmented, high molecular weight DNA (Leitao *et al*., 2024; Dominik *et al*.,

2023). A commonly used indication of molecule quality is N50, which is the length of the molecule that is the shortest of the population containing at least 50% of all the bases. Another consideration for long read sequencing is its read depth. If a molecule only passes once through the sequencing machinery (as for example in Nanopore technologies) it may contain lower signal and more noise resulting in more artefacts and errors compared to next generation technologies where the signal is a consensus of hundreds if not thousands of copies of a molecule. However, the technology is improving and is increasingly allowing for resolving short tandem repeat sequences (fig.3.2). This is of course of high importance to diagnostic testing, but also is invaluable in research by allowing for streamlined sequencing of control populations to find which sequences might be pathogenic or which might be a polymorphism.

Moreover, genome optical mapping is capable of detecting structural variants in most of the patients' genomes opposed to specific targeted repeat expansion in Southern blotting and it will replace Southern blotting in the future.

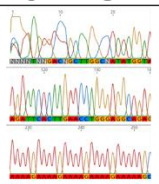
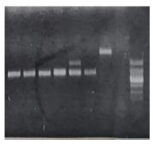
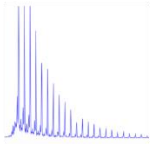
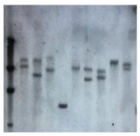
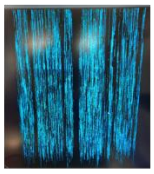
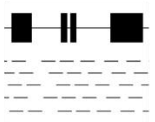
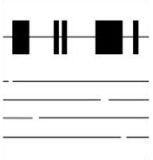
	Sanger sequencing	**Flanking PCR**	**RP-PCR**	**Southern Blot**	**Bionano OGM**	**Short read sequencing**	**Long read sequencing**
Schematic							
Advantages	• Quick • Easy • Indicate reference allele • Might show a repeat motif	• Quick • Easy • Indicate reference allele or small expansion	• Quick • Can indicate a presence of specific repeat motif	• Can accurately size expansions • Sequence independent	• Can accurately size expansions • Whole genome mapped	• Can detect a motif • Can detect a presence of an expansion	• Can indicate a sequence of entire repeat • Can size an expansion
Disadvantages	• Less than 1000bp • Beginning of sequence only • Often does not work	• Might give false no amplification result • Does not show large expansions	• Can indicate specific sequence only • Cannot size the repeat • Might show false negative/positive result	• Cumbersome • Needs large quantities of good quality DNA • Requires specific laboratory set up • Locus specific probe	• Not easily accessible • Special set up needed • Bioinformatics expertise needed • Does not show sequence	• Cannot size the expansion • Requires validation • Prone to errors and artifacts	• Prone to errors and artifacts • Low map rate • Expensive

Figure 3.2 Techniques used in RFC1 repeat expansion testing with their advantages and disadvantages.

In this study we leveraged short read WGS from the Genomics England sequencing project to investigate the normal and pathologic variation of the *RFC1* repeat expansion and to identify additional pathogenic repeat configurations in *RFC1* causing CANVAS and disease spectrum. The full sequencing of novel pathogenic repeats was further analysed by targeted long read whole genome sequencing and sizes of the expansions were measured by optical mapping and/or Southern blotting.

We identified 3 novel pathogenic repeat configurations AAGGC, AGGGC and AGAGG in homozygous or compound homozygous state with AAGGG.

Patients with novel pathogenic configurations mostly showed similar features to biallelic $(AAGGG)_n$ repeat expansion carriers, although in some cases the disease was more complex and the disease course more severe.

3.2 Materials and methods

Main contributors to the methods (and results) in this chapter are listed in table 3.1

Contribution table	
Methods used	**Contributors**
CANVAS Screening (PCR & Southern blotting)	**Natalia Dominik**, Stefania Magri, Riccardo Curro
Sanger sequencing	**Natalia Dominik**
Bionano Optical Genome Mapping	**Natalia Dominik**, Stephanie Efthymiou
Clinical Examinations	Andrea Cortese, Riccardo Curro, Elena Abati, Henry Houlden, RFC1 repeat expansion study group
Bioinformatic analyses	Stefano Facchini, Arianna Tucci, Valentina Pirota
Long read sequencing	Ira Deveson, Hannah MacPherson, **Natalia Dominik**

Table 3.1. Main contributors to the methods (and results) in this chapter

3.2.1 Whole genome sequencing data analysis

The 100,000 Genomes Project, run by Genomics England (GEL), was established to sequence whole genomes of UK National Health Service (NHS) patients affected by rare diseases and cancer (GPP Investigators *et al*., 2021). In this study, we leveraged GEL WGS data and screened for the presence of pentanucleotide expansions in *RFC1* in 893 samples from patients diagnosed with ataxia and 8107 controls, all aged 30 years or older since *RFC1* spectrum disease has a late age of onset. Repeat expansions were detected using ExpansionHunterDeNovo (EHDN) v0.9.0. We considered all motifs composed of five or six nucleotides at the *RFC1* locus. Repeat motifs present in the homozygous or compound heterozygous state with the AAGGG expansion in ataxia cases, but absent or significantly less frequent in controls, were considered to be possibly pathogenic and were further assessed.

Predicted genetic ancestries for samples from GEL were based on a principal component analysis (PCA), using the five macro-ethnicities of the 1000 Genomes project (European, African, South Asian, East Asian, American) as reference populations. Samples in which none of the components reached 95% were classified as ‘Mixed’.

3.2.2 RP-PCR

Samples identified to carry novel pathogenic repeat motifs with EHDN were tested using repeat-primed (RP)-PCR. In addition, we screened a cohort of 540 patients, of which I screened 414 for the 3 novel configurations, with genetically confirmed RFC1 CANVAS, as defined by the presence of a positive RP-PCR for the AAGGG expansion and the absence of an amplifiable PCR product from the flanking PCR, to look for expansions of different repeat motifs on the second allele. RP-PCR for AAAAG, AAAGG and AAGGG expansions was performed as described in chapter 2 together with the rationale of RP-PCR primers used.

Primers for the RP-PCR for the novel configurations were designed to detect the specific motif as in figure 3.3 and are available in the table 3.2 together with the PCR conditions for AGGGC and AAGGC modified to 30 s denaturation per cycle as opposed to 10 s for all other configurations.

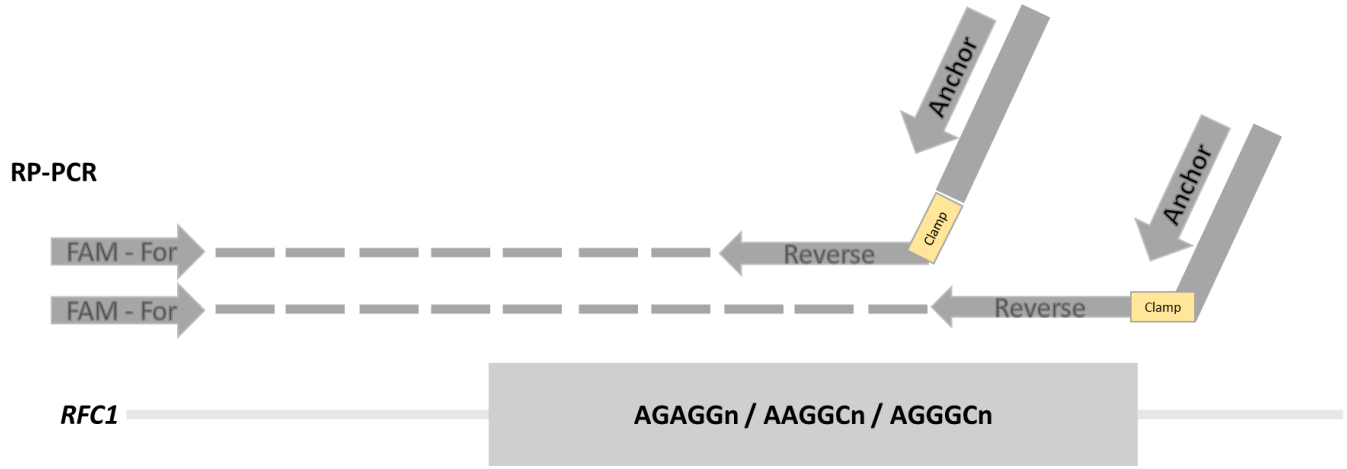


Figure 3.3 A schematic of binding of RFC1 primers used. RP-PCR uses a fluorescently labelled forward primer and two reverse primers –a reverse primer specific to the motif being investigated and an anchor which is complementary to the reverse primer and aids PCR amplification when the reverse primer becomes depleted. The reverse primer contains a stretch of DNA sequence termed 'a clamp' which is complementary to the DNA sequence immediately after the repeat expansion sequence.

	Sequence	Reagents	PCR conditions
Long range flanking PCR	Fw: TCAAGTGATACTCCAGCTACACCGTTGC Rv: GTGGGAGACAGGCCAATCACTTCAG	25µl Phusion Flash High-Fidelity PCR Master Mix 2X (Thermo-Fisher) 25ul H2O 0.4ul forward primer 10 µm 0.4ul reverse primer 10 µm 3% DMSO 1ul DNA (around 50ng/µl)	98°C 5 mins; [98°C 15 secs 65°C 20 secs 72°C 3 mins] x8; [98°C 15 secs 65°C 20 secs each cycle - 0.6°C 72°C 3 mins] x16; [98°C 15 secs 65°C 20 secs 72°C 3 mins] x16; 72°C 5 mins
RP-PCR	Fw FAM: TCAAGTGATACTCCAGCTACACCGT Anchor: CAGGAAACAGCTATGACC RFC1-AGAGG CAGGAAACAGCTATGACCAACAGAGCAAGACTCTG TTTCAAAAAAGAGGAGAGGAGAGGAGAGGA RFC1-AAGGC CAGGAAACAGCTATGACCAACAGAGCAAGACTCTG TTTCAAAAAAGGCAAGGCAAGGCAAGGCAA RFC1-AGGGC CAGGAAACAGCTATGACCAACAGAGCAAGACTCTG TTTCAAAAAAGGGCAGGGCAGGGCAGGGCA	Phusion Flash High-Fidelity PCR Master Mix 2X (Thermo-Fisher) Primers 0.5µM gDNA 50ng	98°C 3 mins; [98°C 30 secs 70°C (M)/ 65°C (V)/ 55°C (W) 50 secs 72°C 2 mins] x35; 72°C 2 min

Table 3.2 RP-PCR primer sequences and cycler conditions

3.2.3 Sanger sequencing

Any patients with no amplifiable PCR product and negative RP-PCRs for AAGGG, AAAGG and AAAAG or where further sequencing information was needed, were subjected to Sanger sequencing as described in chapter 2. Although Sanger sequencing cannot show the entire repeat, it is useful to indicate presence of nucleotide changes within the *RFC1* expansion region in ~1000bps amplified by PCR.

3.2.4 Targeted *RFC1* long-read sequencing

In collaboration with Ira Deveson in Australia, we performed long-read sequencing to establish the precise repeat sequence in patients carrying a novel, likely pathogenic, expansion of *RFC1*. Given the technical hurdle of sequencing large repeat expansions, samples were sequenced on different platforms, including those from Oxford Nanopore and Pacific Biosciences (PacBio). Target enrichment was performed with either a clustered regularly interspaced short palindromic repeats (CRISPR)-associated protein-9 nuclease (Cas9) system or ReadUntil programmable selective sequencing.

3.2.4.1 Single molecule real time sequencing

Single molecule real time sequencing (SMRT) technology is provided by Pacific Biosciences and relies on creating closed circles of DNA that are sequenced multiple times to generate a consensus read. This is accomplished by ligating the source DNA with two hairpin adapters for covalent closing of the DNA molecules. The information on sequence is collected with immobilised polymerase that extends DNA molecules with fluorescently labelled dNTPs who each have their signals detected when excitation at incorporation occurs (fig.3.4).

Errors in sequencing may occur due to noise such as excitation of not yet incorporated dNTPs or nucleotides with no fluorophore. These errors are random and will decrease with more passes of polymerase through the DNA molecule.

Pacbio relies on a large machine for its sequencing and is therefore not easily accessible nor affordable.

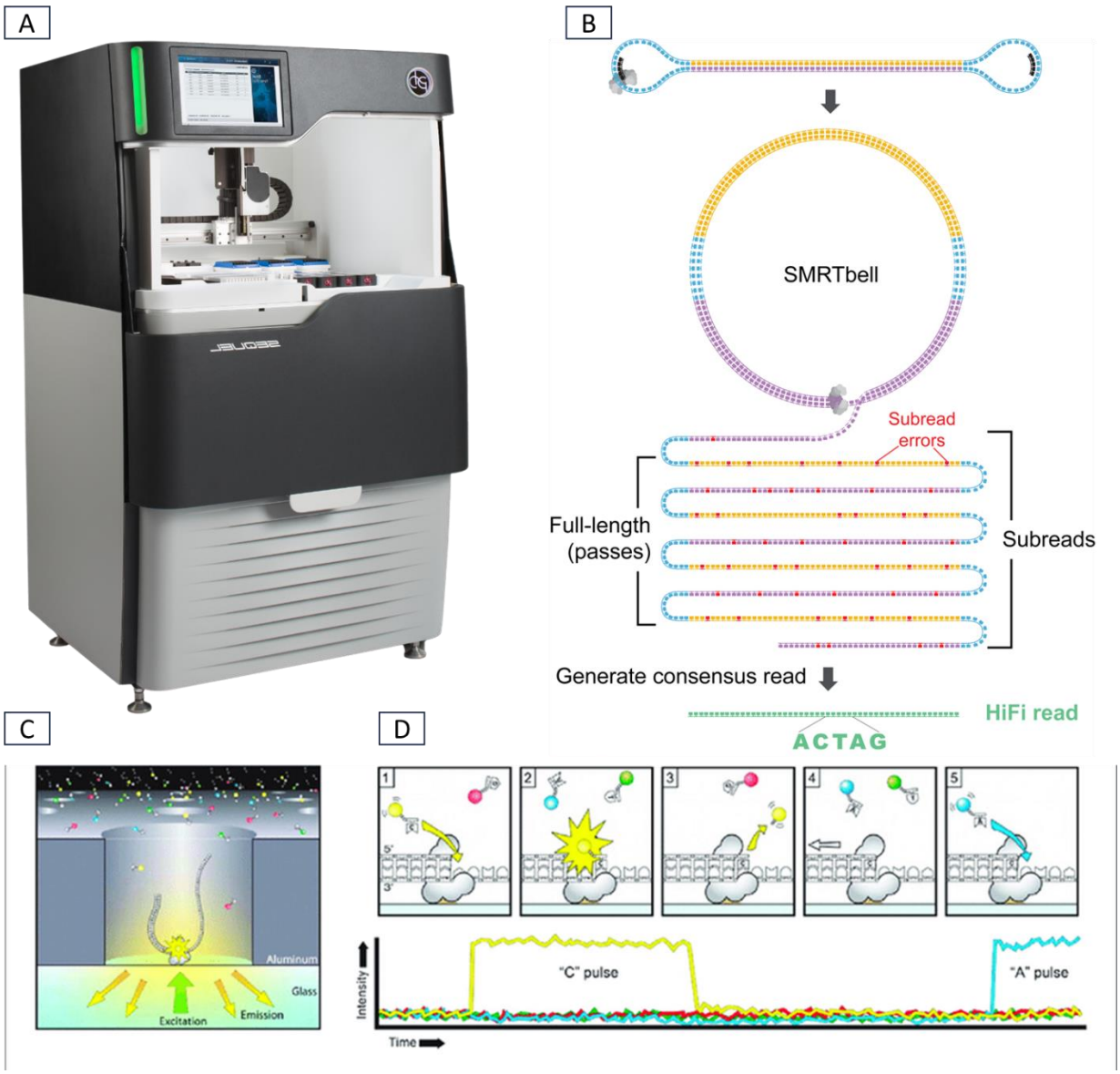


Figure 3.4. Pacbio machine and SMRT bell sequencing. A) An example of Pacbio machine which is a large machine and not easily affordable B) SMRTbell sequencing provides subreads from which consensus read is generated C) Excitation of different fluorescently labelled DNA bases results in different emission for each base (D) which is read by the machine. Adapted from https://www.pacb.com/

3.2.4.2 Nanopore sequencing

Nanopore sequencing is a technology commercially provided by Oxford Nanopore Technologies (ONT). These sequencing methods allow DNA molecules to pass through protein nanopores which allow for sequencing in real time by monitoring the electrical current intensity for each base in DNA as it passes through.

This technology is easily accessible with the smallest machine being portable and connecting to any computer by USB (fig.3.5). However, the technology has still a high error rate of 8-12%

which arises due to electrical current being similar for different bases. In addition, each DNA molecule only passes once through the pore adding to possible low depth of coverage.

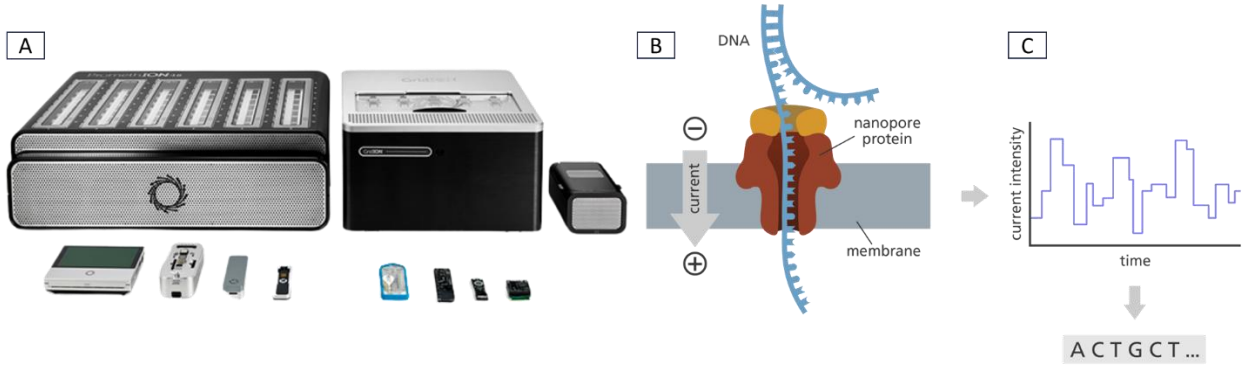


Figure 3.5 Oxford Nanopore technology and sequencing. A) Nanopore machines are of various sizes and the smallest MinION is portable and can be used wherever there is an access to a computer. B) the sequencing technology relies on a single strand of DNA passing a nanopore protein which results in D) different electrical current intensity for each base. Adapted from https://nanoporetech.com/

3.2.4.3 DNA extraction from blood

DNA samples for long read sequencing were extracted from blood using an extraction method that allows for retrieval of high molecular weight DNA which is more suitable for sequencing of long stretches of DNA than column-based extraction methods which mechanically shear DNA and produce shorter molecules. Therefore, I extracted DNA with Qiagen MagAttract HMW DNA kit and an overview of the extraction method is available in fig.3.6.

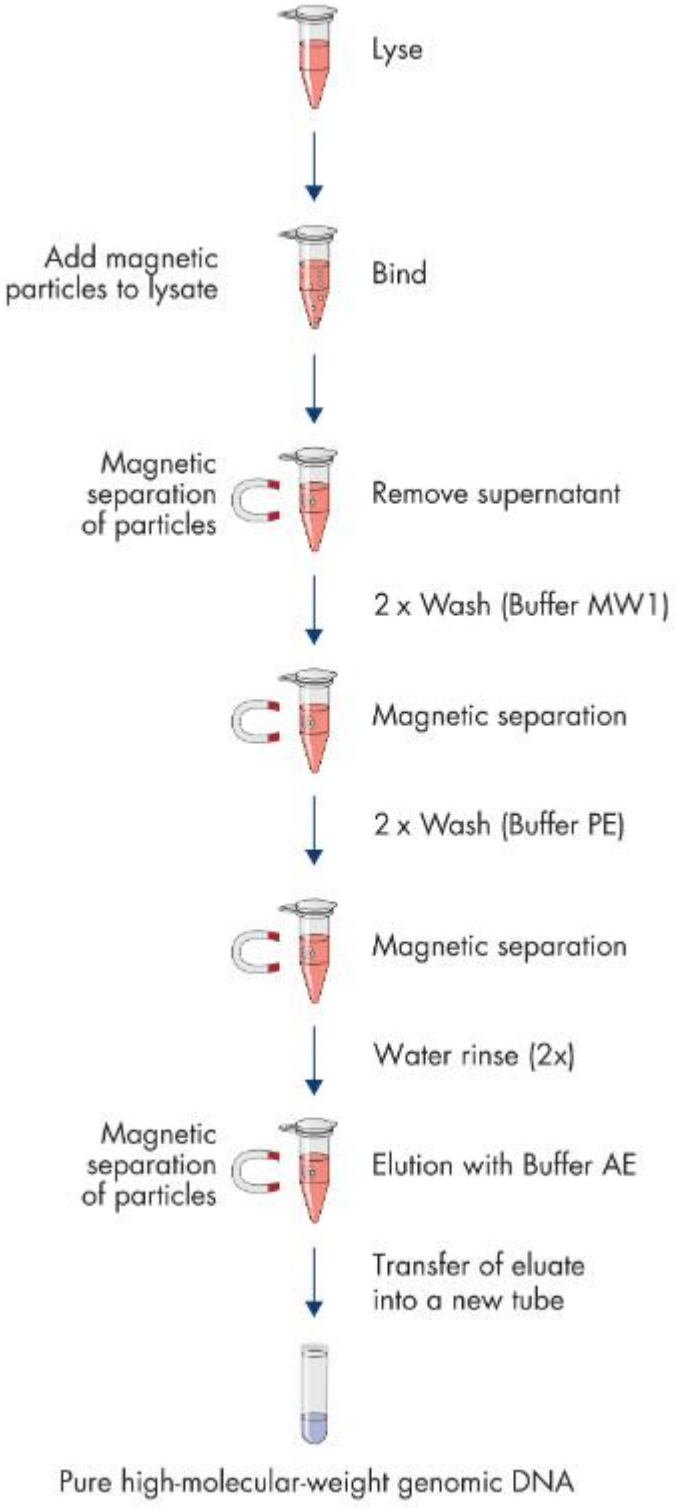


Figure 3.6. Qiagen MagAttract HMW DNA extraction method. This extraction method relies on DNA fragments binding to magnetic beads. Impurities are washed off with buffers MW1, PE and water. Figure from www.qiagen.com

For the extraction, 20µl of Proteinase K was pipetted into an Eppendorf tube and 200µl of blood was added and gently mixed by tapping. Subsequently, 4µl of RNase A solution and 150µl of Buffer AL were added and vortexed. The mix was incubated at room temperature for 30 minutes and 15µl of MagAttract Suspension G was added to the sample after the incubation. After addition of 280µl Buffer MB, the sample was incubated in a mixer at room temperature for 3 minutes at 1400rpm. The sample was placed on a magnetic rack and beads allowed to separate. Without disrupting the beads, the supernatant was collected and disposed of and 700µl Buffer MW1 was added, and the sample was incubated for 1 min at 1400 rpm. The wash step was repeated twice. After removing supernatant 700µl Buffer PE was added and incubation for 1 min at 1400 rpm followed. This step was repeated. After removing supernatant, 70µl of Buffer AE was added to the beads and incubated 3 min at 1400 rpm. To elute the DNA, the sample was placed on the magnetic rack and the supernatant collected into a fresh Eppendorf tube without disrupting the beads.

This DNA was used in long read sequencing collaboratively with Ira Deveson and Colleagues, Sydney, Australia, who also provided the methodology below.

3.2.4.4 Long read sequencing

For CRISPR/Cas9-targeted sequencing, fragment lengths were assessed using the Agilent Femto Pulse Genomic DNA 165 kb kit, and only samples in which the majority of the fragments were over 25 kb were used. Libraries were prepared from 5 µg of input DNA for each sample for both the PacBio No-Amp targeted sequencing utilizing the CRISPR-Cas9 system protocol (Version 09) and the Oxford Nanopore ligation sequencing gDNA Cas9 enrichment (SQK-LSK109) protocol (Version: ENR_9084_v109_revT_04Dec2018). Libraries were sequenced on the Oxford Nanopore PromethION or MinION platforms or the PacBio Sequel IIe, respectively. For the Oxford Nanopore ligation sequencing gDNA Cas9 enrichment, we used four CRISPR-Cas9 guides from Nakamura *et al*.,

RFC1-F1: 5′-GACAGTAACTGTACCACAATGGG-3′,

RFC1-R1: 5′-CTATATTCGTGGAACTATCTTGG-3′,

RFC1-F2: 5′-ACACTCTTTGAAGGAATAACAGG-3′ and

RFC1-R2: 5′-TGAGGTATGAATCATCCTGAGGG-3′, except for Cases IV-1, XI-1 and XII-1, for which only two, RFC1-F2 and RFC1-R2, were used. The guides RFC1-F3: 5′-GAAACTAAATAGAACCAGCC-3′ and RFC1-R3: 5′-GACTATGGCTTACCTGAGTG-3′,

designed in-house, were used for PacBio No-Amp targeted sequencing, and up to 10 samples were multiplexed using PacBio barcoded adapters. Libraries loaded onto the PromethION and MinION were run for 72 hours with standard loading protocols. Sequel IIe libraries were run for a movie time of 30 hours with an immobilization time of 4 hours. All libraries were loaded neat.

Programmable targeted sequencing was performed as described previously (Stevanovski *et al*., 2022). HMW DNA was sheared to fragment sizes of ∼20 kb using Covaris G-tubes. Sequencing libraries were prepared from ∼3–5µg of HMW DNA using a native library prep kit SQK-LSK110, according to the manufacturer's instructions. Each library was loaded onto a FLO-MIN106D (R9.4.1) flow cell and run on an Oxford Nanopore MinION device with live target selection/rejection executed by the ReadFish software package (Payne *et al*., 2021). Detailed descriptions of the software and hardware configurations used for the ReadFish experiments are provided in a recent publication that demonstrates the suitability of this approach for profiling tandem repeats (Stevanovski *et al*., 2022) The target used in this study was the *RFC1* gene locus ±50 kb. Samples were run for a maximum duration of 72 hours, with nuclease flushes and library reloading performed at approximately 24 and 48 hours' time-points for targeted sequencing runs, to maximize sequencing yield.

3.2.4.5 Amplicon long read sequencing using Oxford Nanopore Flongle

Standard testing of *RFC1* repeat region with RP-PCR allows for detecting a presence or absence of a known repeat expansion motif due to the reverse primers carrying that specific motif's sequence to bind along the repeat track. This inadvertently may result in false negative testing if the most common pathogenic AAGGG expansion is not present in the patient.

In order to devise a quick and easy method of discerning possible *RFC1* expansion motifs in genomic DNA, I used Oxford Nanopore long read sequencing with a Flongle flow cell on an amplified PCR product from a known biallelic AAGGG CANVAS sample. PCR amplification was carried out on 50ng/µl gDNA with primers flanking the repeat and with the long-range PCR protocol as in table 3.1.

The library was prepared with Ligation sequencing DNA V14 (SQK-LSK114) kit from Oxford Nanopore. 100fmol of the PCR product was diluted to 23.5µl with PCR grade water and 0.5µl of DCS, 1.75µl of NebNext FFPE DNA repair buffer, 1µl of NEBNext FFPE DNA Repair Mix, 1.75µl of Ultra II End-Prep Reaction Buffer and 1.5µl Ultra II End Prep Enzyme Mix were added and mixed gently by pipetting. Using a thermal cycler, the mixture was incubated for 5

minutes at 20ºC and 65ºC each. The sample was transferred to a DNA LoBind tube and 30µl of AMPure XP beads were added and incubated for 5 minutes at room temperature in hoola mixer. The tube was placed on a magnetic rack and the supernatant collected and disposed without disrupting the beads. The beads were washed twice with 80% ethanol and 31µl of nuclease free water was added. The eluate was collected to 1.5ml Eppendorf DNA LoBind tube and 1µl used for quantification with Qubit fluorometer.

12.5µl of Ligation Buffer, 5µl NEBNext Quick T4 DNA Ligase and 2.5µl of Ligation Adapter were added to the 30µl of DNA sample and incubated for 10 minutes at room temperature. 50µl of AMPure XP beads were added to the sample and mixed by flicking the tube. The sample was incubated on hoolamixer for 5 minutes at room temperature and placed on magnetic rack for 2 washes with short fragment buffer. The sample was eluted with 7µl of elution buffer and 1µl was used for concentration quantification using Qubit fluorometer. 10fmol of the prepared library was injected onto Oxford Nanopore Flongle flow cell and the sample was sequenced for 24 hours using fast model base-calling.

3.2.4.6 Bioinformatic analysis

Bioinformatic analyses were performed by Dr Stefano Facchini, the below methodology was provided by Dr Facchini.

Alignment to the hg38 reference of Nanopore reads, PacBio CCS and PacBio subreads was done using minimap228 with additional options '-r 10000 -g 20000 -E 4,0'. For PacBio sequences, the recommended step of generating circular consensus sequencing (CCS) maps from subreads was not always possible because of the low depth of the sequencing data. The only CCS map we could obtain was for the AAGGG allele in Case V-1. After alignment, we used PacBio scripts (https://github.com/PacificBiosciences/apps-scripts) to extract the repeat region (extractRegions.py) and obtain waterfall plots (waterfall.py) for the following motifs: AAGGG, AGAGG, AGGGC, AAGGC and AAAGG.

For programmable targeted sequencing, raw ONT sequencing data were converted to BLOW5 format using slow5tools (v0.3.0)29 then base-called using Guppy (v6). The resulting FASTQ files were aligned to the hg38 reference genome using minimap2 (v2.14-r883). The short-tandem repeat (STR) site within the *RFC1* locus was genotyped using a validated process (Payne *et al*., 2021). This method involves the local haplotype-aware assembly of ONT reads

spanning a given STR site and annotation of the STR size, motif and other summary statistics using Tandem Repeats Finder (4.09), followed by manual inspection and motif counting.

3.2.5 Haplotype analysis

We used SHAPEITv430 with default parameters to phase a 2 Mb region (chr4:38020000–40550000) encompassing the *RFC1* gene. To maximize available haplotype information, the entire Rare Diseases panel in Genomics England (78195 samples from patients affected by rare diseases) were jointly phased. The input data format was an aggregate VCF file with a total of 551795 variants.

The estimation of haplotype age was based on the online application Genetic Mutation Age Estimator (https://shiny.wehi.edu.au/rafehi.h/mutation-dating/) (Gandolfo *et al.*, 2014). The method required as input a list of ancestral segments for sampled individuals. We used the five individuals with pathogenic expansions: AAGGG hom, ACAGG hom, Case VII-1, Case I-1 and Case III-3.

3.2.6 Optical genome mapping and southern blotting

Patients for whom whole blood was available were subjected to BioNano optical genome mapping (OGM) as described in chapter 2.

3.2.7 In-silico prediction of G-quadruplexes formation

G-quadruplexes (G4s) are stable structures formed by nucleic acids (DNA and RNA) in regions that are rich in guanine. G4s can form multimers and therefore higher order structures, which can further stack together or connect by short loops (fig.3.7). The state of folding and unfolding of G-quadruplexes can affect numerous cellular processes such as genome replication, transcription, and translation (Frasson *et al.*, 2022).

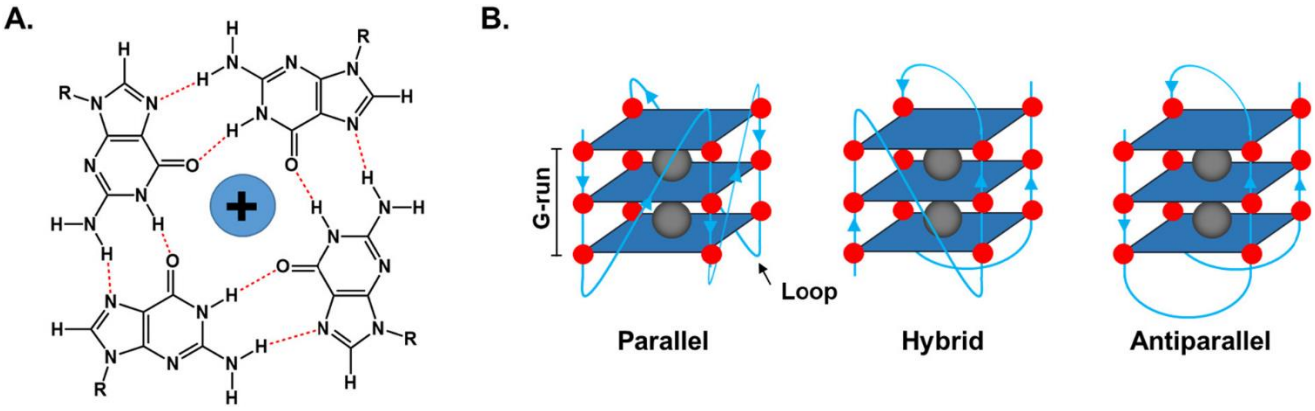


Figure 3.7 G-quadruplex A) G quadruplex is formed through hydrogen bonding in molecules rich in guanine which further B) forms into higher order structures. Adapted from Frasson *et al.*, 2022.

Multimeric G4s have been implicated in diseases such as: ALS/FTD in negative transcription regulation, generation of DNA:RNA hybrids and others (Haeusler *et al*., 2014); Fragile X syndrome in mRNA inefficient translation (Ofer *et al*., 2009).

Collaboratively with Valentina Pirota, Pavia, the propensity of the different repeat configurations in *RFC1* to form G-quadruplexes (G4s) (Frasson *et al*., 2022) was predicted using the Quadruplex forming G-Rich Sequences (QGRS) Mapper (Kikin *et al*., 2006) and G4-Hunter software (Bedrat *et al*., 2016) through which the likelihood to form a stable G4 is rated in terms of G-score values. Putative G4s were identified according to the following parameters for QGRS: a maximum sequence length of 30 nucleotides, minimum number of two G-tetrads in a G4, loop lengths in the range of 0–36 nucleotides and G-score values > 15. The G4-Hunter threshold was 1.5 with a window size of 20 nucleotides.

3.3 Results

3.3.1 Novel pathogenic repeat motifs in *RFC1* in patients from the 100,000 Genome project

Of 893 cases diagnosed with adult-onset ataxia (over the age of 30 years) recruited as part of the 100,000 Genome project, 124 cases harboured at least one AAGGG repeat expansion and 48 had biallelic AAGGG repeat expansions, thus confirming a diagnosis of CANVAS/spectrum disorder.

To identify additional likely pathogenic repeat motifs in *RFC1*, we specifically looked for rare repeat configurations present in patients diagnosed with adult-onset ataxia (over the age of 30 years) or in a compound heterozygous state with the known pathogenic AAGGG repeat expansion but absent or significantly less frequent in controls under the same conditions (fig.3.8).

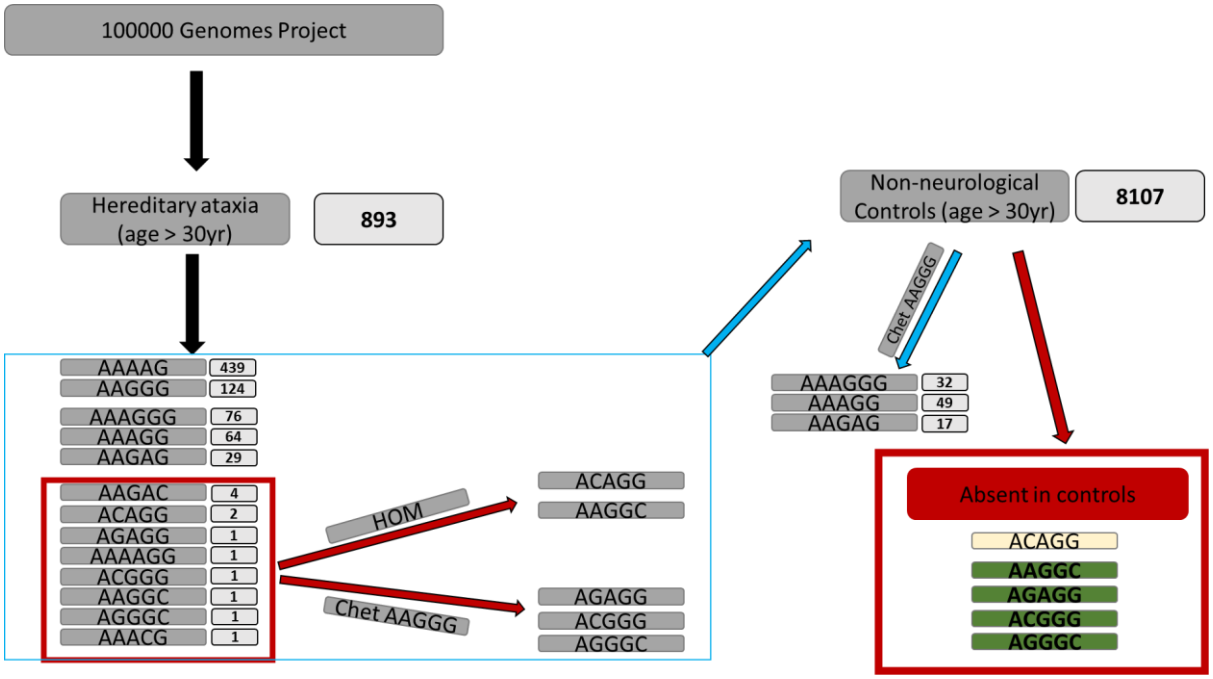


Figure 3.8 100000 genome project screening for novel configurations in *RFC1*. 893 patients over 30-year-old with hereditary ataxia were screened and homozygous and compound heterozygous motifs with AAGGG were noted. Non-neurological controls were also screened, and the 5 expansion motifs that were only found in the ataxia cohort are shown in the red box. The pentanucleotide motif ACAGG in yellow has previously been described (Scriba *et al.*, 2020) and the four pentanucleotide motifs in green are novel motifs described in our study.

Various pentanucleotide and two hexanucleotide motifs were identified and they include, three cases carrying repeat expansions AAGGC (Case I-1), AGGGC (Case II-1) or AGAGG (Case VII-1) repeat motifs, which were absent in non-neurological controls. AAGGC was present in the homozygous state, while AGGGC and AGAGG were in the compound heterozygous state with the AAGGG expansion. One additional case with self-reported Asian ancestry carried the previously reported rare pathogenic ACAGG repeat expansion in the homozygous state.

AAAAG, AAAGGG and AAGAG expansions were found at similar frequencies in patients and controls (table 3.3), supporting their non-pathogenic significance, while there was a higher percentage of compound heterozygous AAGGG/AAAGG carriers in ataxia cases (P = 0.05).

Patients carrying AAGGC (Case I-1) and AGGGC (Case II-1) expansions were of predicted South Asian and mixed ethnicity, respectively; an ACAGG expansion carrier was confirmed to be East Asian based on the predicted genetic ancestry, while other repeat configurations were mostly identified in individuals of European or mixed ethnicity.

We did not identify any loss-of-function variant or structural variant in the *RFC1* gene in individuals carrying heterozygous AAGGG repeat expansions.

	Hereditary ataxia (N=893)	Non neurological controls (N=8107)	P values
Rare homozygous (<1%) repeat expansions present in ataxia cases and absent in controls			
ACAGG/ACAGG	1 (0.01%)	0 (0 %)	-
AAGGC/AAGGC	1 (0.01%)	0 (0 %)	-
Repeat expansion found in compound heterozygous state with AAGGG expansions (allele 1/allele2)			
AAGGG/AAAAG	21 (2.3%)	248 (3%)	Ns
AAGGG/AAAGGG	5 (0.6%)	32 (0.4%)	Ns
AAGGG/AAGAG	3 (0.3%)	16 (0.2%)	Ns
AAGGG/AAAGG	10 (1.1%)	47 (0.6%)	0.05
AAGGG/ACGGG*	1 (0.01%)	0 (0%)	-
AAGGG/AGAGG	1 (0.01%)	0 (0 %)	-
AAGGG/AGGGC	1 (0.01%)	0 (0 %)	-

Table 3.3 Frequency of different *RFC1* biallelic expansions in 100000 genome projects' ataxia patients and non-neurological controls over the age of 30. Rare homozygous (<1%) and compound heterozygous with known pathogenic AAGGG motifs are shown. Orange circles highlight the AAGGG/AAAGG alleles that are significantly enriched in the hereditary ataxia cohort as opposed to the non-neurological controls. Ns=non-significant. *ACGGG is found in small non-pathogenic repeat expansion range. Adapted from Dominik *et al.*, 2023.

3.3.2 Genetic screening for validation of novel motifs

The presence of AGGGC, AAGGC or AGAGG repeat expansions was confirmed by RP-PCR in all three cases, and the AAGGC repeat segregated with the disease in Family I, as it was also present in the affected sister Case I-2 (fig. 3.9).

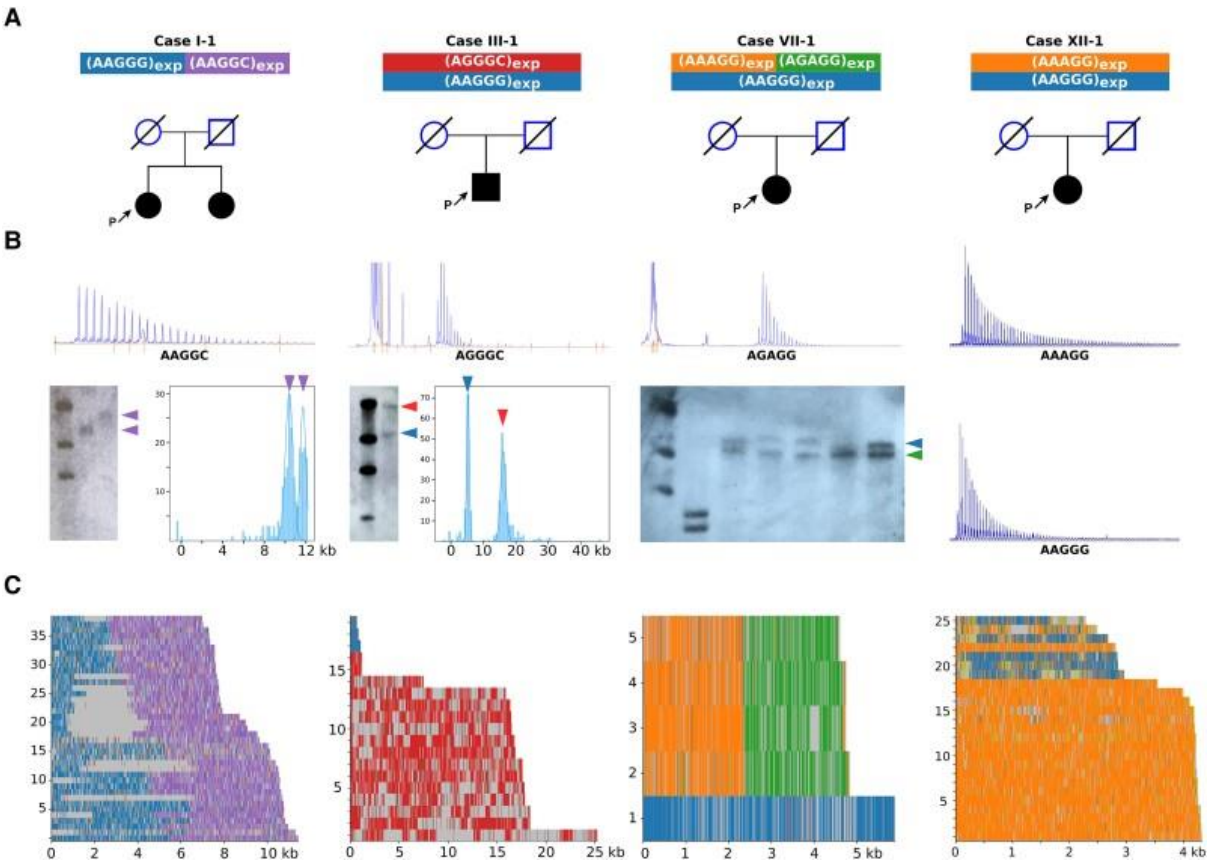


Figure 3.9 Long-read sequencing defines the precise sequence of the novel pathogenic *RFC1* motifs. (A) Pedigrees. P = proband. (B) RP-PCR plots and, where available, Southern blotting images and optical genome mapping plots. (C) Long-read sequencing results of representative patients with AAGGC, AGGGC, AGAGG and AAAGG expansions (Cases I-1, III-1, VII-1 and XII-1). In Case III-1, only partial reads, which did not span the entire *RFC1* repeat locus, could be obtained from the AAGGG allele. Adapted from Dominik *et al*., 2023.

Additionally, one case with isolated cerebellar ataxia carried the AAGGG expansion along with an ACGGG repeat, which was absent in the controls. However, Sanger sequencing showed that the ACGGG expansion was only 50 repeats, which is considerably below the lower limit of pathogenicity (250 repeats) for the pathogenic AAGGG motifs and was therefore considered likely to be non-pathogenic in this case (fig.3.10). Notably, the patient exhibited isolated cerebellar ataxia but no neuropathy, which is unusual in *RFC1* disease.

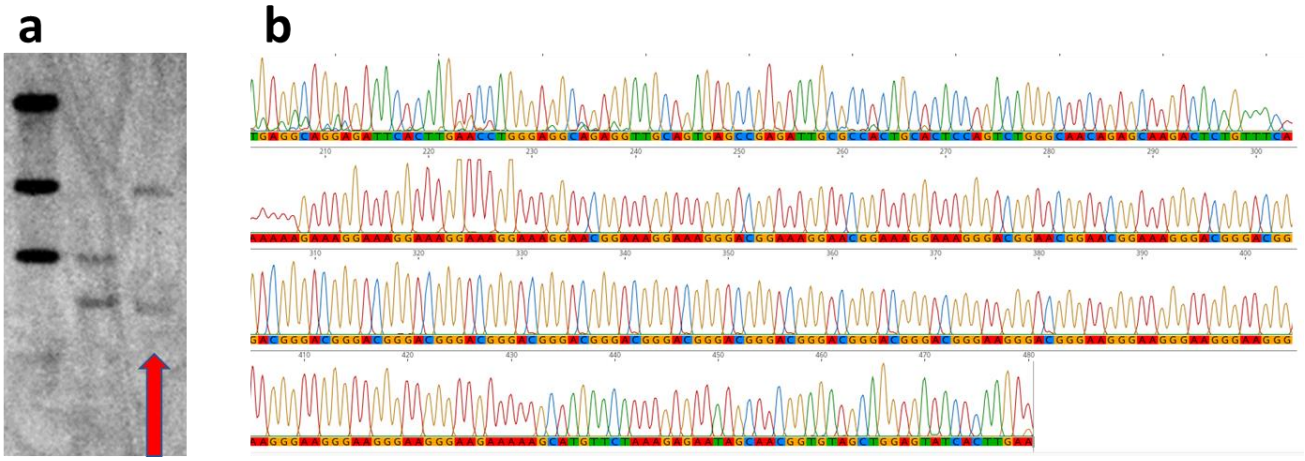


Figure 3.10 ACGGG repeat motif. a) Southern blotting shows one allele with small expansion and one expanded allele. B) Sanger sequencing shows ACGGG motif and suggests a very small expansion of about 50 pentanucleotide repeats which is within the non-pathogenic repeat expansion range of up to 220 repeats.

Next, we used RP-PCR to screen an internal cohort of 540 DNA samples from cases with sensory neuropathy, ataxia or CANVAS and identified five additional cases carrying an AGGGC expansion (Cases III-1, IV-1, V-1, V-2 and VI-1) and three cases carrying AAAGG expansions on the second allele (Cases X-1, XI-1 and XII-1). We did not identify additional AGAGG or AAGGC repeat expansion carriers. All cases were of self-reported Caucasian ethnicity.

Based on Southern blotting, OGM or long-read sequencing (fig. 3.9 B and C) when available, we observed that the sizes of the rare AGGGC, AAGGC and AGAGG repeat expansions were >600 repeats in all cases [mean ± standard deviation (SD), 892 ± 247 repeat units] (fig. 3.11 A). Furthermore, enough DNA for Southern blotting was available from five patients with CANVAS/spectrum disorder (Cases VI–X), as defined by the presence of sensory neuropathy and at least one of the additional features of the full syndrome (cerebellar dysfunction, vestibular areflexia, cough), and eight controls carrying compound heterozygous AAGGG/AAAGG expansions (fig.3.11 B).

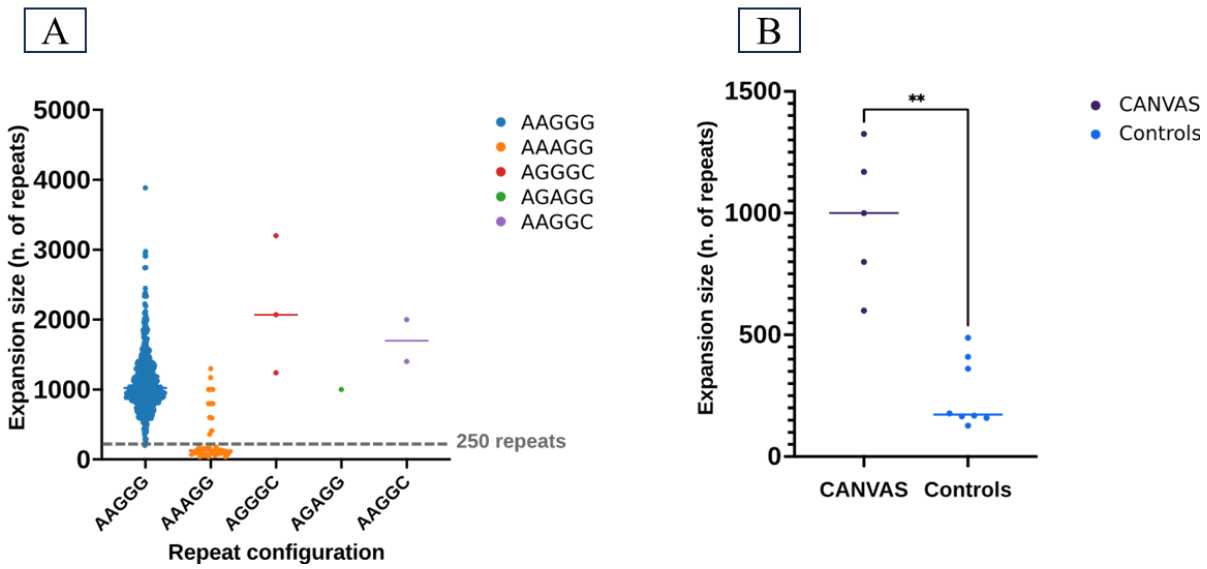


Figure 3.11 *RFC1* repeat expansion sizes. A) Expansion sizes of common pathogenic AAGGG, previously thought non-pathogenic AAAGG and three novel configuration motifs. The dotted lines refer to the smallest pathogenic expansion of 250 AAGGG repeats identified so far B) Expansion sizes of AAAGG compared in CANVAS patients and non-neurological controls. CANVAS patients show significantly ($p<0.01$) larger AAAGG expansions than non-neurological controls. The AAAGG expansions are in compound heterozygous state with the pathogenic AAGGG expansion. Adapted from Dominik *et al.*, 2023.

3.3.3 Long-read sequencing confirms the sequence of the expanded repeats

To gain further insight into the exact sequence of the novel pathogenic motifs, we performed targeted long-read sequencing (fig. 3.9D). We confirmed the presence of uninterrupted $AGGGC_{1240}$ in Case II-1 and $AGGGC_{3200}$ in Case III-1. Moreover, long-read sequencing enabled us to define the exact repeat composition of the AGAGG and AAGGC expansions, which revealed the presence of mixed repeat motifs $(AAGGC)_{900}(AAGGG)_{940}$ and $(AGAGG)_{470}(AAAGG)_{470}$ in Cases I-1 and VII-1, respectively. Long-read sequencing was also performed in five cases carrying large AAAGG expansions and showed the presence of uninterrupted AAAGG motifs in three (Cases X-1, XI-1 and XII-1), with sizes of 980, 800 and 600 repeat units, respectively, while two probands (Cases VIII-1 and IX-1) carried complex $(AAAGG)_{610}(AAGGG)_{390}$ and $(AAAGG)_{700}(AAGGG)_{200}$ repeats.

3.3.4 All pathogenic repeat configurations share an ancestral haplotype

A haplotype is a combination of different single nucleotide polymorphism along the same allele that tend to be inherited together (Greenspan and Geiger 2004).

Haplotypes are important tools in investigating disease-causing loci in both family and population-based studies and can provide information on recombination events, population mutation events or distant events such as founder effects.

We looked at the inferred haplotypes associated with the novel pathogenic repeat motifs. A region of 66 kb (fig. 3.12, between Markers B and C, chr4:39302305–39366034, hg38) was shared among all pathogenic alleles. It is worth noting that a larger region of 207 kb (between Markers A and C) containing the *WDR19* and *RFC1* genes was shared among all the pathogenic alleles, except one (Case III-1), where the haplotype became the same as the wild-type allele. This suggested a more recent recombination event at Marker B in Case III-1. The larger shared region identified in carriers of the novel pathogenic configurations, as well as in AAGGG and AAAGG carriers, supports the existence of an ancestral haplotype that gave rise to these expanded alleles. Notably, non-pathogenic $AAAAG_{(9–11)}$ and expanded AAAAG repeats originated from a different haplotype.

We estimated that the ancestral haplotype that gave rise to different pathogenic repeat configurations in *RFC1* likely dates to 56 100 years ago (95% confidence interval: 27 680–115 580 years).

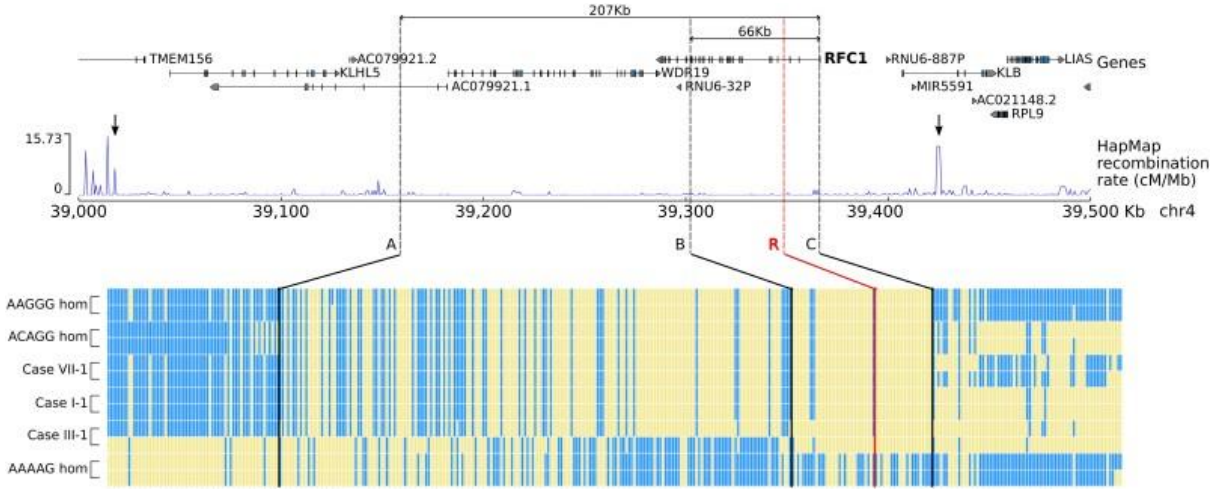


Figure 3.12 A shared ancestral haplotype in patients with pathogenic *RFC1* motifs. Graphical representation of the haplotypes associated with AAGGG, ACAGG and novel pathogenic repeat motifs identified in this study. For each single nucleotide polymorphism, the reference allele is represented in blue, while the alternative allele is represented in yellow. The repeat expansion locus is marked with a red line (R). There is a shared region (B–C, -rs2066782-rs6851075, chr4:39302305–39366034, hg38) of 66 kb for all novel configurations. A larger region of 207 kb (A–C, rs148316325- rs6851075, chr4:39158847–39366034, hg38), which is flanked by two recombination hotspots (arrows), is also shared among all but one allele for Case III-1, suggesting a recombination event at B (rs2066782) in this family. The shared haplotype lies in a region of low recombination rate (HapMap data) and is delimited by small peaks at A and C. A smaller increase in the recombination rate is also visible at B. hom = homozygous. Adapted from Dominik *et al*., 2023.

3.3.5 Clinical features of patients carrying novel pathogenic repeat configurations in *RFC1*

We found 14 patients from 12 families carrying novel pathogenic *RFC1* repeat configurations. The demographic and clinical characteristics of patients are available in appendix 1. All patients were Europeans, apart from Cases I-1 and I-2, who were from India, and Case X-1, who was from Australia. The mean age-of-onset was 51.5 ± 13.7 (24–73) years, and mean disease duration at examination was 17.2 years ± 8.7 (3–34) years. Six patients had isolated sensory neuropathy, which was associated with cough in four of them; one patient had sensory neuropathy and vestibular dysfunction; while seven cases had full CANVAS. Additional features were observed in some cases, including early onset and rapid progression (Case I-1), cognitive impairment (Cases III-1 and VI-1), muscle cramps (Cases I-1, II-1, III-1 and IV-1) and REM sleep behaviour disorder with positive dopamine transporter scan (DatScan) (Case IX-1). Autonomic dysfunction was observed in six cases, and in two of them (Cases II-1 and III-1), who both carried AGGGC expansions, it was severe and led to syncopal episodes.

3.3.6 Pathogenic configurations in *RFC1* are predicted to form G-quadruplexes

As repetitive G-rich sequences are known to form G4s, secondary DNA structures which act as transcriptional regulators by impeding transcription factor binding to duplex-DNA or stalling the progression of RNA polymerase, we set out to evaluate the propensity of the different repeat configurations in *RFC1* to form G4s.

All pathogenic repeat configurations showed high G4 scores, which were in the range observed for the well-known G4-forming regions of the *cMYC37* and *HRAS138* genes, as predicted by QGRS-Mapper and G4Hunter, in contrast to the non-pathogenic AAAAG (Table 3.4).

Gene: analysed sequences	QGRS-Mapper score	G4Hunter score
RFC1: $(AGGGC)_{10}$	42	1.83
RFC1: $(AAGGG)_{10}$	42	2
RFC1: $(AAGGC)_{10}$	21	1.82
RFC1: $(AAAGG)_{10}$	21	0.94
RFC1: $(AGAGG)_{10}$	21	1.12
RFC1: $(AAAAG)_{10}$	No putative G4 identified	
c-MYC: TGGGGAGGTGGGGAGGGTGGGGAAGG	41	2.59
HRAS-1: TCGGGTTGCGGCGCAGGCACGGGCG	41	1.19

Table 3.4. G quadruplex formation prediction by various *RFC1* pentanucleotide motifs compared to well-known G4 forming sequences of c-MYC and HRAS-1. All the pathogenic pentanucleotides found in *RFC1* repeat locus (AGGGC, AAGGG, AAGGC, AAAGG and AGAGG) show high G quadruplex scores comparable to the well-known G4-forming regions of the *cMYC37* and *HRAS138* genes. The non-pathogenic AAAAG expansion is shown not to form G4. Adapted from Dominik *et al*., 2023.

3.3.7 Motif detection using long read sequencing on amplified PCR product

The currently used Sanger sequencing can only sequence the beginning of the repeat up to around 1000bp and often does not provide a good quality sequence. Similarly, and as discussed RP-PCR only provides indication of presence or absence of a specific motif tested. Whole exome or whole genome short read sequencing may enable to infer a motif at the beginning or end of the repeat sequence directly flanking the repeat region, however, this sequence would be short (around 100bp) and may incorrectly map to the region due to short read sequencing limitations. In addition, targeted long read sequencing using CRISPR/Cas9 guides proved challenging in the case of *RFC1* repeat expansions, and it is still expensive therefore difficult to scale up.

Therefore, a known biallelic AAGGG sample has been PCR amplified across the repeat region and I performed library preparation and long read sequencing with Oxford Nanopore Flongle sequencing to inspect whether a repeat motif can be inferred with this method thus warranting a possibility of easier, more streamlined and scaled-up *RFC1* motif screening without the need for expensive and challenging CRISPR/Cas9 target enrichment.

The Flongle sequencing of PCR product (fig.3.13) shows AAGGG sequence at the beginning of the repeat only followed by regions of high variability of bases and presence of thymines

and cytosines further downstream. These are errors likely to have arisen due to PCR amplification across the repeat region and relatively high error rate in nanopore basecalling. In addition, the sequencing indicates a presence of non-expanded allele (top panel, fig.3.13), contrary to flanking sequencing where no reference band was detected and Southern blotting where no allele of normal size was present. This is likely caused by a contamination of the sample at the DNA amplification stage.

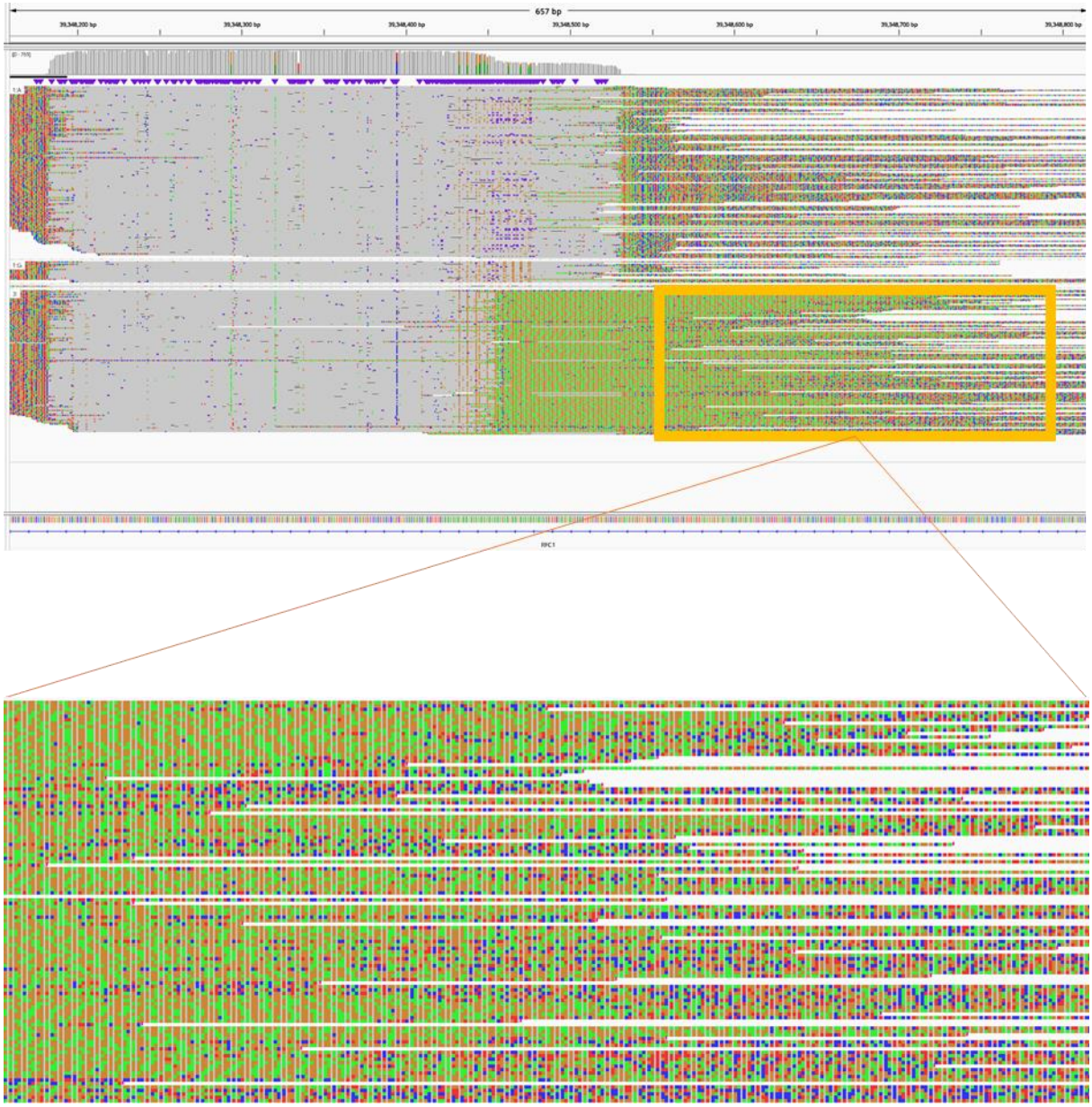


Figure 3.13 Oxford Nanopore Flongle sequencing. Barcoded primers targeting *RFC1* repeat were used to visualise expansion motif present in a DNA sample known to contain biallelic AAGGG expansions for validation of using Flongle for *RFC1* repeat expansion motif detection. Top panel indicates an unexpanded allele which might have resulted from contamination at PCR amplification stage, bottom panel indicates a repeat expansion and yellow box is used to zoom in on the repeat expansion which shows high error rate and presence of various bases other than expected in AAGGG expansions (green and orange).

3.4 Discussion

We leveraged WGS data from nearly 10 000 individuals recruited to the Genomics England sequencing project to investigate the normal and pathogenic variation of the *RFC1* repeat. We identified three novel repeat configurations associated with CANVAS/spectrum disorder, including AGGGC, AAGGC and AGAGG. Notably, we also showed a pathogenic role for large uninterrupted or interrupted AAAGG expansions, that had previously been thought not pathogenic. AAAAG, AAGAG and AAAGGG expansions are likely always to be benign (fig.3.14)

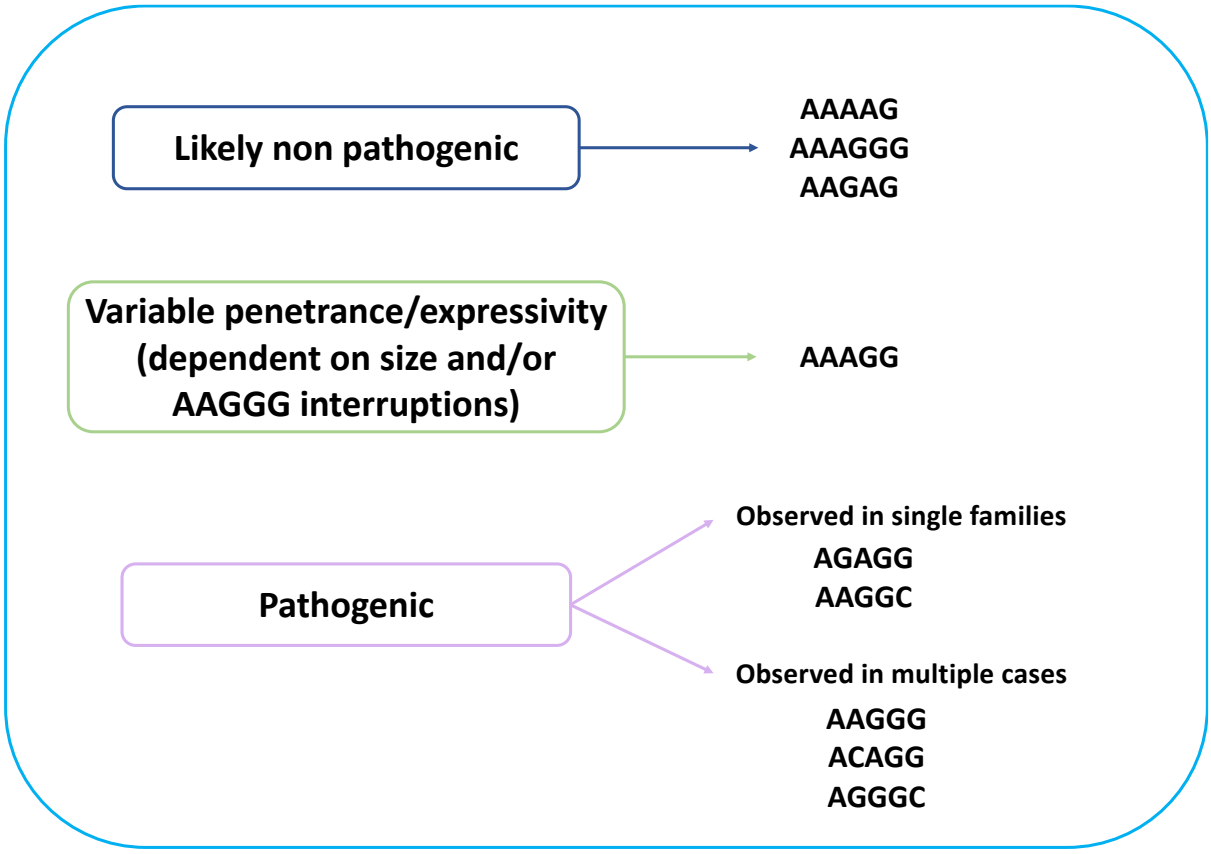


Figure 3.14 Normal and pathogenic significance of repeat expansion motifs at the *RFC1* locus. Likely non-pathogenic repeat motifs found in *RFC1* found in our cohort are shown as well as the pathogenic motifs observed either in single families or in multiple cases. AAAGG repeat expansion motif is found to be either non-pathogenic or pathogenic when sufficiently expanded and in compound heterozygous state with known pathogenic AAGGG. Adapted from Dominik *et al*., 2023.

Most pathogenic repeat expansions were found in individuals of Caucasian ancestry; however, ACAGG seemed to be common in East Asians, while AAGGC was identified in a family of South Asian ancestry. Interestingly, most pathogenic repeats seem to have arisen from a shared region of 207 kb, supporting their origin from a common ancestor who lived ~50 000 years ago. Rafehi *et al*. previously identified a larger ancestral haplotype in Australian patients affected by CANVAS of 360 kb and estimated that the most recent common ancestor lived

approximately 25 880 (confidence interval: 14 080–48 020) years ago (Rafehi *et al*., 2019). In our study, the inclusion of additional pathogenic repeat configurations and multiple ethnicities allowed the identification of a smaller core haplotype and has extended further back in time the origin of the common ancestor carrying a pathogenic repeat in *RFC1*. It is reasonable to believe that the occurrence of subsequent A–G transitions and A–G or G-C transversions in the poly-A tail of the AluxSx3 element on the ancestral haplotype favoured the further expansion of GC-rich motifs over the millennia. Since the most significant recent wave out of Africa is estimated to have taken place about 70 000–50 000 years ago, we can speculate that the repeat-containing haplotype spread with the migration of early modern humans from Africa through the Near East and to the rest of the world.

Patients showed clinical features undistinguishable from those of patients carrying biallelic AAGGG expansions. In some cases, however, the disease appeared to be more severe due to symptomatic dysautonomia, early cerebellar involvement or disabling gait disturbance.

The identification of these motifs has direct clinical implications. Given their frequency, RP-PCR for AAAGG and AGGGC should be considered in all cases. Particular attention should be paid to carriers of compound AAGGG/AAAGG expansions and accurate sizing, and full sequencing of the satellite through long-read sequencing is recommended to establish its possible pathogenicity. In addition, depending on availability, Southern blotting, genome optical mapping or long-read sequencing are warranted in patients with a suggestive clinical phenotype but inconclusive screening, such as in cases with absence of a PCR-amplifiable product on flanking PCR but negative RP-PCR for AAGGG expansion.

In addition, during the course of this thesis, Dr Cortese's group uncovered 7 patients with clinical CANVAS phenotype but with AAGGG expansion on one allele only. We tested 15 individuals with standard screening methods of flanking PCR and RP-PCR and Southern blotting if sufficient DNA was present and further, they were submitted for whole genome or whole exome sequencing to test for presence of a second coding variant in *RFC1* in trans with the AAGGG repeat expansion. 7 patients from 5 unrelated families were found to carry a point mutation in *RFC1* in trans with the AAGGG repeat expansion and patient fibroblasts were found to have reduced *RFC1* transcript and protein (Ronco *et al*., 2023).

These findings add complexity to CANVAS genetic testing and highlight the importance of a full characterisation of *RFC1* expansions sequence and size to provide the patients with correct diagnosis (3.15).

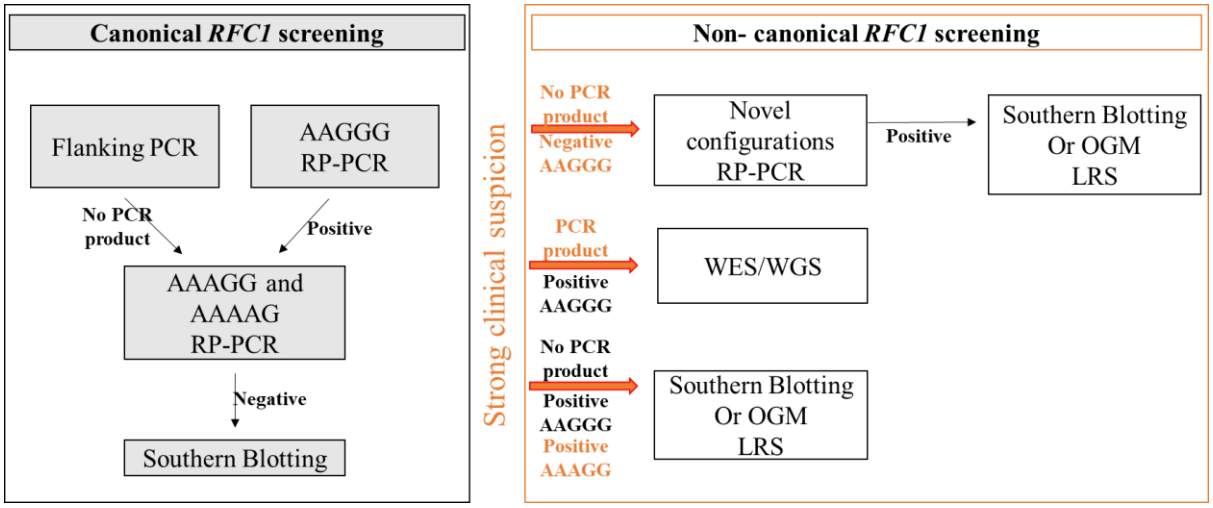


Figure 3.15. Proposed algorithms for screening of *RFC1* mutations. Canonical *RFC1* screening includes RP-PCR for most common pathogenic AAGGG repeat expansion and sizing confirmation with Southern blotting. If a patient's phenotype warrants strong clinical suspicion of CANVAS in absence of clearly biallelic AAGGG screening results, non-canonical *RFC1* screening should be employed which can include expansion sizing via SB or OGM or LRS or WGS/WES for detection of a second pathogenic variant in trans with the pathogenic expansion.

The findings of this study highlight the genetic complexity of RFC1 related disease and lend support to the hypothesis that the size and GC-content of the pathogenic repeat is more important than the exact repeat motif. Consistently, all pathogenic repeat configurations are rich in G-content and are predicted to form highly stable G4s, which have previously been demonstrated to affect gene transcription in other pathogenic conditions (Varshney *et al*., 2022; Wang *et al*., 2021).

Both Nanopore or PacBio sequencing platforms and either the targeted CRISPR/Cas9 or adaptive selection approach were used to increase the accuracy of the sequencing of the *RFC1* repeat locus. Despite several attempts and similarly to other large satellites, long-read sequencing of the *RFC1* repeat remained challenging and, depending on the specific configurations, size and DNA quality, only a few reads were available for analysis in some cases. Notably, uneven coverage at the *RFC1* locus across samples was also observed in a recent study of *RFC1* repeat composition using Nanopore sequencing (Erdman *et al*., 2023). The authors attributed the variability to variable degrees of DNA fragmentation depending on the delay between blood sampling and DNA extraction.

3.4.1 The future of repeat expansion testing

In my thesis, I have used a variety of repeat expansion sizing and sequencing techniques. My work expanded the knowledge of the heterogeneity of the *RFC1* expansion locus and explained the importance of not only detecting the correct repeat expansion motif but also of sizing the repeat. I showed that PCR techniques are informative in detection of likely *RFC1* RE positive

samples, however, these cannot size the repeats and might lead to false negative results if the common expansion motif is not found. We validated a new method of sizing repeat expansions, Optical Genome Mapping. This method is advantageous to Southern blotting as it allows for analysing structural variants across the entire genome of a patients whilst Southern blotting relies on locus-specific probes for repeat detection. Although diagnostic laboratories in the UK have already adopted PCR-based screening for *RFC1*, Southern blotting is not routinely used for sizing or confirming results. OGM may be advantageous in diagnostic settings for diseases such as FSHD (Efthymiou et al., 2023), however, for *RFC1* disease it may prove too expensive for purpose of confirmation of PCR screening results and it does not provide DNA sequence information which is key in RFC1 disease.

Should long-read sequencing technologies overcome current obstacles, such as low read count and relatively low accuracy as we encountered in RFC1 sequencing, they hold the potential to deliver optimal outcomes for diagnostic applications—enabling comprehensive genome screening, accurate expansion sizing and phasing, and precise detection of specific repeat motifs. Recently, PacBio have released Puretarget which is a new method of characterising repeat expansions using expansion specific panels for high coverage of high molecular weight DNA. This method can resolve repeat expansions to single base accuracy, size the expansions but also detect methylation patterns. From a diagnostic perspective, where fresh blood or tissue of interest can be extracted from the patient and processed almost immediately, LRS may be advantageous, however, it does come with challenges, such as the need for substantial data storage and skilled bioinformaticians for rapid data analysis.

From a research perspective, LRS technologies are still expensive and not readily available in many research centres. The inability to use fresh tissues, especially if only archived DNA is available, may hinder discoveries using these technologies. Despite these hurdles, structural variations remain a crucial area of research, as they are linked to many hereditary diseases. There are likely to be further discoveries as these technologies evolve. Hopefully, constant advancements in long-read sequencing platforms and a decrease in cost ((currently ~£1100 per sample and more expensive if PureTarget is applied)) will soon translate into increased accessibility to this technology and higher levels of accuracy while allowing for concurrent sizing of the repeat expansions and the precise detection of the repeat expansion motifs.

3.4.2 *RFC1* repeat motifs in literature

Since the discovery of *RFC1* repeat expansions causing CANVAS disease spectrum and the heterogeneity of the locus, the need to screen various populations and individuals suffering from clinical CANVAS has been increasingly met. Much remains elucidated about the disease mechanism and other pathogenic configuration motifs may be discovered in the near future. To date, studies of *RFC1* repeat motifs, have been carried out in Australia, Japan, Canada and Brazil (Scriba *et al*., 2020; Tsuchiya *et al*., 2020; Nakamura *et al*., 2020; Rafehi *et al*., 2019). Perhaps the study most similar to our approach was this of an Australasian cohort published by Scriba *et al*, in 2023. There, the group investigate the *RFC1* expansion motifs within a cohort of 242 patients with neurological disease and they use the approach of flanking and RP-PCR screening. Targeted long read sequencing was employed where gaps in expanded alleles were observed using standard methods. 3 repeat motifs were reported, and they included AGGGG, AAGAC and AAAGGG. Similarly, to our data, the group observed high percentage of cases with AAAGGG motif, in our data we observed the motif in both disease and control cohort; and AAGAC which was present in our disease cohort, however, never in homozygous or compound heterozygous state with AAGGG. We both argue the importance of *RFC1* repeat expansion screening in diagnostic settings, however, while we argue the importance of the repeat expansion size as exampled by AAAGG motif, Scriba *et al*., theorise that repeat motif present is more important determinant of pathogenicity.

Interestingly, heterogeneity of *RFC1* and limitations of short read sequencing in tandem repeats are increasingly recognised and novel approaches to bioinformatic analysis of long read sequencing are being developed such as Tandem Repeat Genotyping Tool (TRGT, Dolzhenko *et al*., 2024). This method allows for exact genotyping of HiFi reads from PacBio sequencing which includes exact sizing of repeat expansions in *RFC1* and repeat motif calling. This was evidenced in 100 samples analysed with the software where five different motifs (AAAAG, AAGAG, AAAGGG, AAGGG and AAAGG) were called and sized. Interestingly, TRGT is now used as a tool in PacBio's new method of long-read sequencing, Puretarget. Whilst PacBio whole genome sequencing is still expensive, together with high quality bioinformatic tools, in the future, it may replace current PCR and Southern blotting workflows to offer more streamlined, single test for interpreting challenging loci such as *RFC1*.

3.4.3 Exploring RFC1 Pathomechanisms: Insights from Genetic Studies

The RFC1 disease causing mechanism is still unknown. Repeat expansions can cause disease through loss-of-function or gain-of-function mechanisms and both mechanisms are being investigated in the context of RFC1 pathomechanism.

To date, there has been very limited evidence for gain-of-function mechanism in RFC1 disease. For example, RNA foci formation has been studied in CANVAS patient brains, and indeed, separate studies seem to come to different conclusions with some not detecting foci in patients' brains (Cortese et al. 2019) and some detecting relatively small foci with which could only be seen with super resolution microscopy (Wada et al., 2023) or in small number of post-mortem samples (Maltby et al., 2024). These studies however are limited to few post-mortem samples and further research is needed to lead to unequivocal answer.

Interestingly, despite the recessive mode of inheritance in RFC1 disease spectrum, the protein and RFC1 transcripts are unchanged in the bulk tissues of the patients with biallelic AAGGG expansions, thus challenging the loss-of-function hypothesis.

One of the main limitations of examining post-mortem brains stems from the neuronal degeneration. It is highly probable that the most affected neurons of interest have already degenerated, thereby limiting our ability to study their original structure, function and indeed any possible accumulation of toxic RNA or protein species, or conversely change in RFC1 expression. Moreover, to date only bulk tissue samples from RFC1 patients have been analysed, which carries the risk of masking gene expression changes which could potentially be only visible at the single-cell level.

Consequently, these challenges can lead to an incomplete understanding of the disease-causing mechanisms. However, much can be learned about the possible pathomechanism from the genetic heterogeneity underlying this condition and I have directly contributed to these studies.

More specifically, genetic studies conducted in this thesis seem to support a loss-of-function hypothesis of RFC1 expansion disease. Firstly, the most relevant clue to the pathomechanism comes from the identification of patients with compound heterozygous AAGGG expansions and truncating variants in *RFC1*, who have decreased expression at both RNA and protein levels (from the non-expanded allele). Importantly, other recessively inherited conditions have been associated with compound heterozygous pathogenic variants/repeat expansion genotypes. For example, approximately 1–4% of individuals with Friedreich ataxia have a pathogenic

GAA expansion on 1 FXN allele and another pathogenic point mutation on the other allele; notably, aloss-of-function pathomechanisms with reduced frataxin level is well established (Cook and Giunti, 2017).

Secondly, I described a large genetic heterogeneity at the *RFC1* locus with 3 novel pathogenic repeat expansion motifs, AGGGC, AAGGC and AGAGG, discovered in the course of this thesis. The discovery of pathogenic expansions other than AAGGG adds important information to our knowledge of *RFC1* pathogenesis as, together with the truncating variants, it clearly shows that different mutations can cause the same phenotype; we proposed that the GC content of the repeat unit, rather than exact motif, might drive the pathogenicity of the repeat expansions at *RFC1* locus. Indeed, from the literature, repetitive sequences rich in guanine content are known to form G quadruplexes, secondary DNA structures which act as transcriptional regulators by impeding transcription factor binding to duplex-DNA or stalling the progression of RNA polymerase. Here, we showed that all the pathogenic repeat expansions, regardless of the exact motif, are predicted in silico to form stable structures and in particular G-quadruplexes. In contrast, the non-pathogenic, reference pentanucleotide motif AAAAG, is not predicted to form G-quadruplexes. G-quadruplexes have been implicated in negative transcription regulation in ALS/FTD, generation of DNA:RNA hybrids (Haeusler *et al*., 2014) and in mRNA inefficient translation Fragile X syndrome (Ofer *et al*., 2009). Additionally, the wide variety of pathogenic motifs makes it improbable that they all generate toxic species.

While the mechanism of *RFC1* has still not been concluded, the above genetic evidence point towards the loss-of-funciton, in addition, interrogation of public datasets shows that *RFC1* seems intolerant to loss of function, as demonstrated by the absence of biallelic truncating variants, suggesting this would be incompatible with life, and the presence of only 11 *RFC1* truncating variants of 251,000 alleles present on gnomAD v2.1.1 (allele frequency = 0.00002) with an observed/expected ratio lower than 0.35 (o/e = 0.18, 90% CI = 0.12–0.3) and a very high probability of being loss-of-function intolerant (pLI = 0.97). Moreover, as described in Chapter 5 of this thesis, *Drosophila melanogaster* model of global knock-down of RFC1 fly ortholog by use of RNA interference resulted in no viable animal which further supports that *RFC1* is intolerant to loss of function.

However, more functional studies are needed to fully elucidate the *RFC1* pathomechanism, including further examinations of patient post-mortem tissues, iPSC neurons and animal models.

3.4.4 Limitations and future horizons

The main limitations of the work described in this chapter include predominantly European ancestry of the individuals recruited in the 100000 genomes project which could result in any other possible expansion motifs not being represented, only one family with AGAGG motif being found which limits the evidence of the motif segregation within disease, and technology limitations in targeted long read sequencing of the *RFC1* locus meant that we were able to obtain limited information on the whole sequence of the satellite. Future lines of investigations may include functional assessments on how these novel repeat expansions contribute to the mechanism of disease.

3.5 Conclusion

In conclusion, this study expanded the genetic heterogeneity underlying *RFC1* CANVAS/spectrum disorder and identified three additional pathogenic AAGGC, AGGGC and AGAGG repeat motifs. In addition, we report novel likely non-pathogenic expansion motifs (fig.3.16). We also demonstrated a pathogenic role for large uninterrupted or interrupted AAAGG expansions, thereby highlighting the importance of sizing and, if possible, full sequencing of the *RFC1* satellite expansion in clinically selected cases, to correctly diagnose and counsel patients and their families.

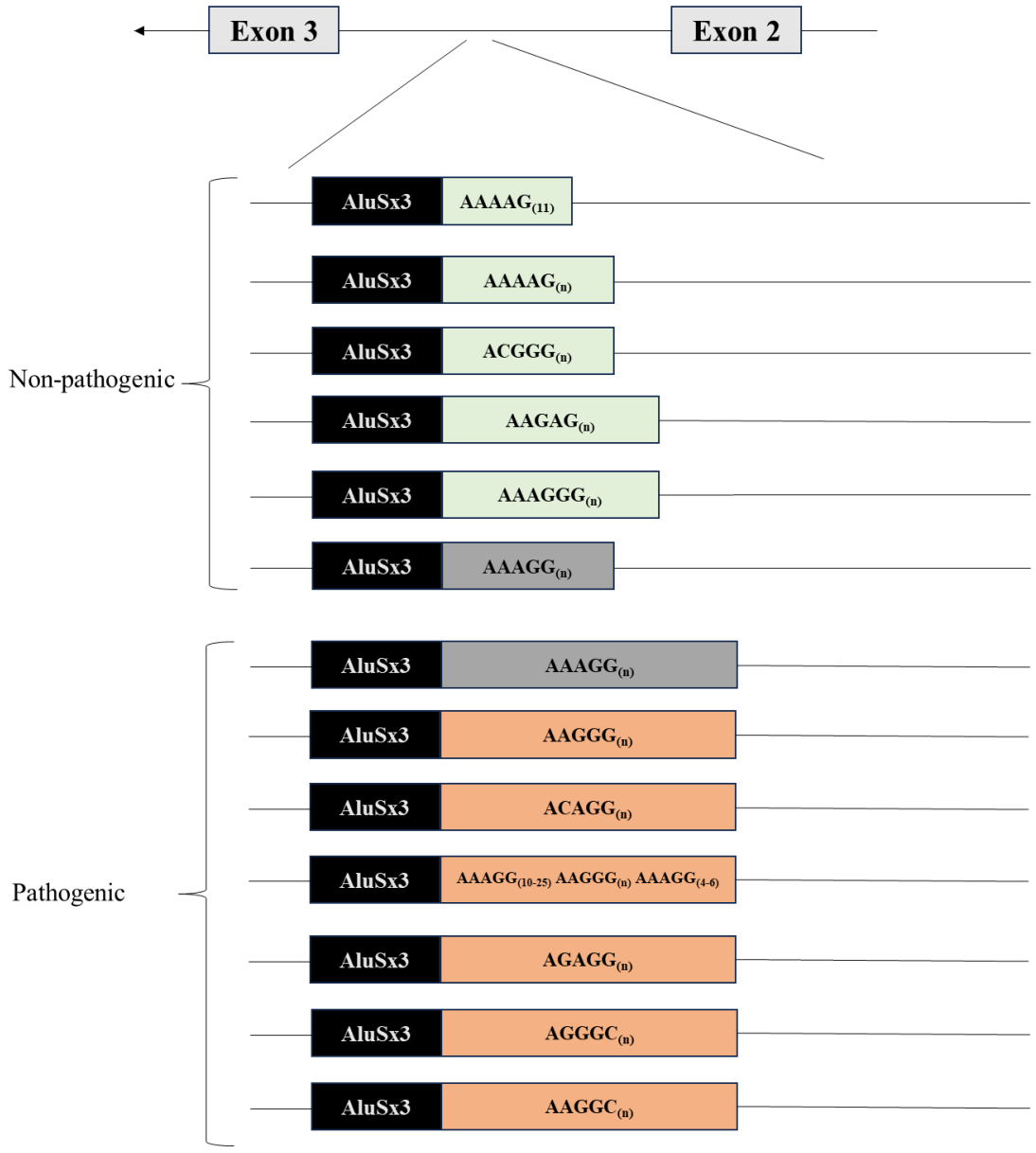


Figure 3.16. Current known genetic heterogeneity of *RFC1* locus. Shown in green are currently known non-pathogenic repeat expansion motifs in RFC1. In grey is a non-pathogenic $AAAGG_{exp}$ which can become pathogenic when sufficiently expanded and in compound heterozygous state with the pathogenic $AAGGG_{exp}$ motif. In orange are currently known pathogenic expansion motifs.

CHAPTER 4. Biallelic variants in *ARHGAP19* cause a motor-predominant neuropathy with asymmetry and conduction slowing

4.1 Introduction

Charcot-Marie-Tooth Disease (CMT) is a heterogeneous group of disorders with over 100 causative genes identified to date. Despite the progress in identification of genetic causes of CMTs, approximately a quarter of patients remain without genetic diagnosis (Record *et al.*, 2024; Pisciotta and Shy 2023). Here, I report on a novel gene, *ARHGAP19*, in which recessive mutations cause a motor predominant neuropathy with asymmetry and conduction slowing.

4.1.1 Rho/ROCK pathway

The Rho family of GTPases are small (~21 kDa) G proteins and the family consists of about 20 members but most studied include RhoA, Rac1, CDC42 (Huang *et al.*, 2024). They act as molecular switches by cycling between inactive guanine nucleotide diphosphate (GDP) bound state and active, triphosphate (GTP) bound state and are involved in signalling pathways that control cell adhesion, cell cycle control, migration and others (fig.4.1). The activity of Rho GTPases is very tightly regulated by guanine exchange factors (GEFs) whose role is to facilitate the exchange from GDP to GTP; Guanine Nucleotide Dissociation Inhibitors (GDIs) that negatively regulate Rho to keep it in inactive state; GDI displacement factor (GDF) that help GDI release Rho; and GTPase activating proteins (GAPs) which stimulate intrinsic low GTPase activity of Rho (Niftullayev and Lamarche-Vane, 2019)

4.1.1.1 RhoA/ROCK and downstream effectors

When in its active, GTP bound state, RhoA acts on serine-threonine specific Rho-associated protein kinase (ROCK) family, ROCK1 and ROCK2. Numerous downstream targets have been identified for ROCK that can be regulated by phosphorylation events. Many of them result in regulation of the cell shape and motility and some can influence the cell cycle and survival pathways. Some known downstream effectors of ROCK include (fig.4.1):

- myosin light chain phosphatase targets regulatory subunit (MYPT1) which can inhibit the phosphatase and result in increased actomyosin assembly and contractility. Phosphorylation events in this branch of the pathway have been shown to cause slow relaxation of pulmonary arteries in rats (Oh *et al.*, 2024).

Moreover, in rats, after spinal injury inhibition of ROCK can result in improved outcomes by regulating the phosphorylation of MYPT1 (Chiang and Cao, 2021).

- LIM Kinase (LIMK) can phosphorylate cofilin which can stabilise actin filaments. LIMK has been shown to play an important role in rat hippocampus for learning and memory and LIMK knock out mice have lower levels of phosphorylated cofilin and show dendritic spines abnormalities with memory deficits (Meng *et al*., 2004; Lunardi *et al*., 2018)
- phosphatase and tensin homolog (PTEN) has an inhibitory role on proliferation. Interestingly, in a study of human patients' ductus arteriosus tissue, another GAP protein, ARHGAP26, was found downregulated and the expression levels were confirmed in knockdown mice tissues. Further, the expression of ARHGAP26 had consequences on phosphorylation of PTEN and thus proliferation of the tissues (Li *et al*., 2019).
- phosphorylation state of vimentin can control cytokinesis. Importantly, *ARHGAP19* has been shown to act as GAP to RhoA and via ROCK2 mediate phosphorylation of vimentin in T-lymphocytes (David *et al*., 2014).

  Vimentin is also implicated in myelination of peripheral nerve and expressed not only in Schwann Cells but also in neurons (Triolo *et al*., 2012; Chen *et al*., 2023) thus placing it as an interesting target to explore in ARHGAP19 CMT disease.
- tau is a well described microtubule stabilising protein. Tau knockdown in a malignant gliomas cell line (U87) showed mislocalisation of another GAP protein, ARHGAP35 which inhibits Rho-ROCK signalling (Breuzard *et al*., 2019).
- FHOD1 is a formin family member and when phosphorylated by ROCK it can result in intermediate filament disruption. FHOD1 is highly expressed in aorta and smooth muscle (Status *et al*., 2011).

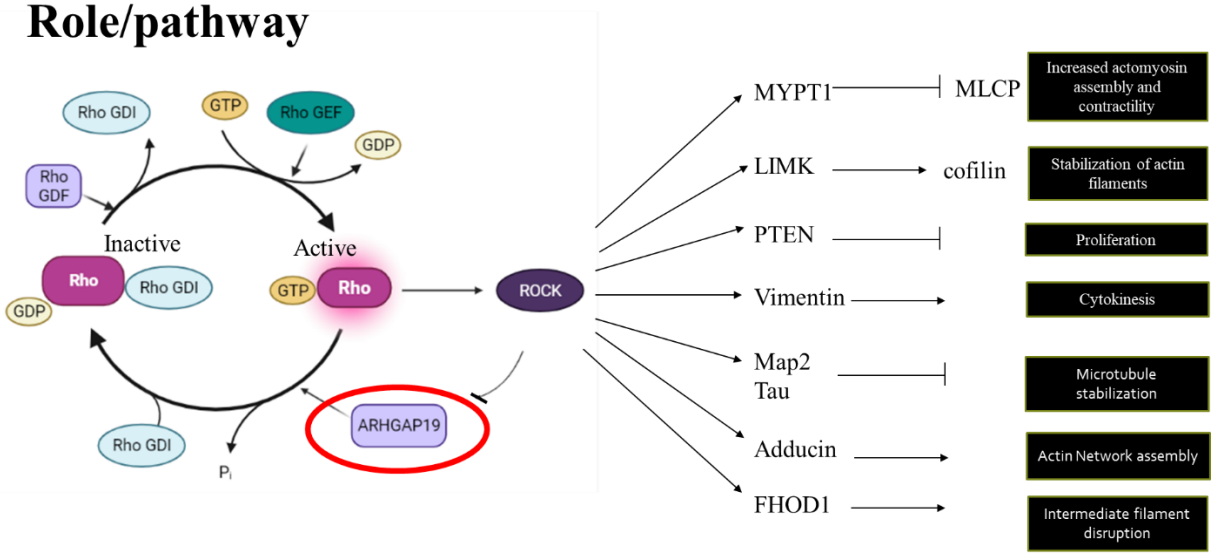


Figure 4.1 Rho/ROCK pathway. Rho cycles between inactive GDP bound state and active GTP bound state. The pathway is tightly regulated by regulatory proteins – GAPs GEFs and GDIs. ARHGAP19 is a GAP protein. When in its active state Rho activates ROCK which can further activate many downstream effectors. The effectors in this figure are speculative for the purpose of the functional investigations of *ARHGAP19*

4.1.1.2 Rho GTPase activating proteins

RhoGTPase activating proteins stimulate the low intrinsic GTPase activity of Rho therefore allowing Rho to dissociate from GTP and turn from its active state, therefore negatively regulating Rho/ROCK pathway (Huang *et al*., 2017). p50RhoGAP was the first GAP protein identified some 30 years ago (Garrett *et al*. 1989). Later, RhoGAP domain, was described to contain approximately 170 ammino acids that were required for the protein GAP activity.

In eukaryotes, over 66 RhoGAPs (and 82 RhoGEFs) have been identified for 20 members of Rho family that they can act on (DeGeer and Lamarche-Vane, 2013), indicating that each RhoGAP may have a specialised role, their activity may be spatially and temporally regulated, there may be a complex interplay between Rho GTPases GAPs and GEFs and/or some of the GAP may be redundant. The specific GTPases that the GAPs interact with cannot be predicted by their sequence and rather they have to be experimentally determined (David *et al*., 2014; DeGeer and Lamarche-Vane, 2013). In addition to their GAP domain, many RhoGAPs may have other functional domains which can interact with different proteins and therefore implicate those RhoGAPs in other pathways. However, I focus on GAP domain and will only briefly mention any other domains if necessary.

4.1.1.3 GAPs in neuronal development

A well-known role of RhoGAPs is regulation of cytoskeleton organisation during neuron development and therefore mediation of neuronal morphology. Here, I briefly describe the roles of better studied RhoGAPs in neuronal development as they have been published in literature.

- Slit-Robo (sr) GAPs, srGAP 1, 2 and 3, are highly expressed in central nervous system during development and they are structurally similar. They contain not only the rhoGAP domain but also Fes-Cip4 homology Bin/Amphiphysin/Rvs (F-BAR) domain and an SH3 domain. Srgap1 acts towards RhoA, and if bound to the intracellular site of Robo through its SH3 domain, it can display its GAP activity towards Cdc42. Through its GAP activity, srGAP1 can disrupt the actin structure and inhibit migration of neurons. srGAP2 acts towards Rac1 and through its F-BAR domain it can increase neurite growth and branching and decrease neuronal migration. srGAP3 acts towards Rac1 and induces neurite outgrowth (Carlson *et al*., 2011; Fosatti *et al*., 2016).
- ARHGAP39 (also knowns as Vilse) acts towards Rac1 dowstream of Slit-Robot, to inactivate Rac1 activity in axon repulsion. Vilse can also mediate dendritic spine formation and normal dendritic spine morphology in hippocampus (Nowak, 2018).
- Synapse defective protein 1 (SYD1) is a GAP protein located in presynaptic terminals and it is reported to negatively regulate Rac1 GTPase activity to promote axon guidance (Holbrook *et al*., 2012).
- Breakpoint cluster region protein (BCR) has GAP activity towards Rac1 and is regulated by phosphorylation and dephosphorylation events. It has been reported to have functions in spine formation restriction and actin polymerisation in excitatory synapses. In addition, BRC has a key role in cerebellar development (Kaartinen *et al*., 2011).
- P250GAP which acts towards Cdc42, Rac1 and RhoA is highly expressed in brain and is involved in mediation of axonal growth. Through its interactions with beta cathenin, N-cadherin, NMDA receptors and PSD95, it is involved in dendritic spines reorganisation (Chagnon *et al*., 2010; Nakamura *et al*., 2016).
- TCGAP, also known as ARHGAP33, which acts towards Cdc42 and is highly expressed in CNS, is responsible for spine formation and maturation. It also supresses axonal outgrowth (Nakazawa *et al*., 2016).

- Oligopherin-1 has been shown to act towards RhoA and is expressed in the brain. It can mediate spine length and through stabilisation of AMPA receptors it is involved in maturation and plasticity of synapses (Khelfaoui., 2007).
- Chimaerins act towards Rac1 and are expressed in the brain. They mediate spine formation and are necessary for axon guidance, retraction and normal pruning.
- P190rhoGAPs are encoded by ARHGAP5 and ARHGAP35 and act towards RhoA. They are highly expressed in CNS and are involved in axon guidance and outgrowth and in dendritic spine stability and maturation (Heraud *et al*., 2019).
- ARHGAP17 (also known as Nadrin) is a neuron specific GAP which has been shown to act towards all major Rho protein members in vivo, its main function is inhibition of neurite outgrowth (Beck *et al*., 2014; Beck *et al*., 2015).

4.1.1.4 GAPs (and GEFs) in neurological conditions

Given the pivotal role of RhoGAPs in the human nervous system, it is perhaps no surprise that many are implicated in neurological disease**.**

Of GAPs listed above, oligophrenin-1 and srGAP3 have been linked to intellectual disability. Moreover, a private mutation in *ARHGAP4* was found to cause intellectual disability in a Chinese family (Liu *et al*., 2016). *ARHGAP35* has been implicated with progressive non-fluent aphasia in a study of gene-based association in frontotemporal dementia (Mishra *et al*., 2017). *ARHGAP26* autoimmunity was associated with subacute inflammatory cerebellar ataxia in a patient with severe limb and gait ataxia, dysarthria and diplopia when screened with protein microarray (Jarius *et al*., 2010). Recently, copy number variations in *ARHGAP10* have been associated with schizophrenia in a Japanese patient and a mouse model recapitulates the disease (Sekiguchi *et al*., 2020; Tanaka *et al*., 2023). Another recently identified gene with a GAP domain, *MY09B* is implicated in CMT2 with optic atrophy (Cipriani *et al*., 2023).

Of GEF proteins, two have been associated with CMT- *ARHGEF10* is the first association with CMT disease and polyneuropathy where a family with autosomal recessive mutations in this gene presented with thinly myelinated axons (Verhoeven *et al*., 2003); and *PLEKHG5* is another GEF gene, involved in axon terminal autophagy of synaptic vesicles, associated with CMT (Chen *et al*., 2020).

4.1.2 *ARHGAP19* in literature

ARHGAP19 is a Rho-GTPase activating protein with GAP activity on RhoA acting in RhoA/ROCK pathway, as negative regulator of ROCK. It can be phosphorylated by kinases

such as ROCK and CDK1 with evidence that its phosphorylation state can control the proteins localisation during cell cycle progression by changing its ability to interact with other proteins. *ARHGAP19* is predominantly expressed in hematopoietic cells and to date it has been shown to have an essential role in T lymphocyte division (David *et al*., 2014). Interestingly, previous in vitro research by David M.D. 2014 *et al*., showed that not only does *ARHGAP19* have GAP activity towards RhoA but not Rac and CDC42, but also that this activity is lost when a mutation of R143A, within GAP domain, was introduced.

Recently, a Korean cohort of individuals with high cardiovascular risk but normal coronary arteries were genetically screened for variants associated with the phenotype. ARHGAP19 expression levels in tissues such as coronary artery, whole blood and adipose tissue were identified as potentially of biological impact (Kim *et al*., 2023).

4.1.3 Emerging role of *ARHGAP19*

ARHGAP19 is not very well studied especially in the context of neurological disease.

ARHGAP19 is located on the reverse strand of chromosome 10 (Chromosome 10: 97,222,173-97,292,673). According to ensembl (www.ensembl.org), the gene has 7 transcripts, of which 4 are protein coding, one undergoes nonsense mediated decay, one has a retained intron, and one is still undefined (fig.4.2 A), however, uniprot (www.uniprot.org) suggest 7 potential protein isoforms formed by alternative splicing.

An isoform is a different form of a protein, often with differences in amino acid sequences, resulting from the same gene but arisen through alternative splicing, alternative promotor usage or post-translational modifications. Different protein isoforms may act in different tissues, at different time points or respond to different signals. *ARHGAP19* transcripts and isoforms are not yet well characterised, however, ENST00000358531.9 is considered a canonical transcript resulting in ARHGAP19 protein of 494 amino acids.

Structurally (fig.4.2 B), ARHGAP19 has the main functional GAP domain between amino acids 102 and 308. In addition, it carries 3 nuclear localising residues in its C-terminus: Threonine 404 and 467 which can be phosphorylated by CDK1 and Serine 422 that can be phosphorylated by ROCK (Marceaux *et al*., 2018).

According to The Genotype-Tissue Expression (GTEx) project (https://gtexportal.org/home/) bulk tissue expression of ARHGAP19 is highest in Tibial Nerve (fig.4.2 C) and single cell RNA

sequencing in the portal shows high expression of the RNA in skeletal muscle (as well as oesophagus, prostate and heart).

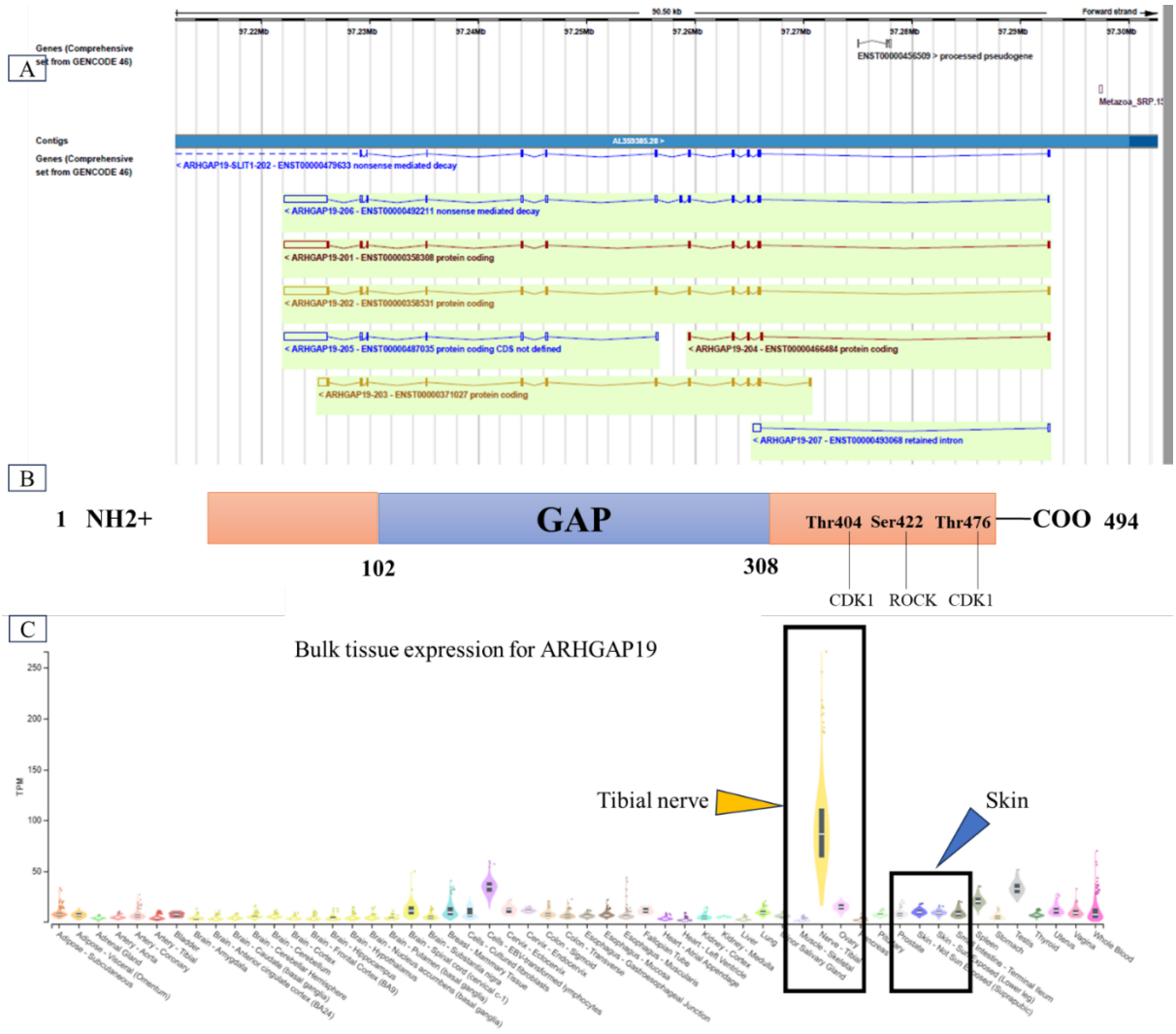


Figure 4.2 A) *ARHGAP19* transcripts from ensmbl B) ARHGAP19 protein structure with known functional GAP domain (between AA 102 and 308) and three phosphorylation sites at the C-terminus of the protein – Thr404 and Thr476 can be phosphorylated by CDK1 and Ser422 can be phosphorylated by ROCK. C) Bulk tissue expression of ARHGAP19 from GTEx Project. ARHGAP19 is highly expressed in tibial nerve (yellow arrow). Low expression is found throughout different tissues in the body including skin (blue arrow).

Here, I report on a large international collaboration that allowed for the identification and characterisation of *ARHGAP19* mutations in individuals presenting with a motor predominant CMT. I describe the clinical phenotypes associated with biallelic, autosomal recessive mutations in *ARHGAP19* in 25 affected individuals from 20 families. Collaboratively, I provide strong genetic evidence for these rare mutations and analyse their impact through the use of individual cell lines, in-vitro GAP assay, in-vivo *Drosophila melanogaster* (described separately in chapter 5) and *Danio rerio* knockdown and in-silico molecular modelling.

4.2 Materials and methods

Main contributors to the methods (and results) in this chapter are listed in table 4.1

Contribution table	
Methods used	**Contributors**
Sanger sequencing	**Natalia Dominik**, Stephanie Efthymiou, ARHGAP19 study group
Clinical assessment	Christopher Record, Henry Houlden, Mary Reilly, ARHGAP19 study group
RNA extractions, qPCR	**Natalia Dominik**
Western blotting	**Natalia Dominik,** Abigail Wilson
In silico modelling	Jevin Parmar
GAP activity assays	Xinyu Miao, Nathalie Lamarche-Vane
Danio rerio assays	Renee Lin, Daniel Osborn
Fibroblast assays	Yi He, Annarita Scardamaglia
Motor neuron differentiation	Riccardo Curro, Ricardo P Schnekenberg
Immunohistochemistry	Kristina Zhelchenska, **Natalia Dominik**
Bioinformatic analyses	Shahryar Alavi, Stefanie Efthymiou, **Natalia Dominik**

Table 4.1. Main contributors to the methods (and results) in this chapter

4.2.1 Study Participants

All participants provided informed consent for participation. Individuals were recruited via an international collaborative network of research and diagnostic sequencing laboratories through partnerships such as SYNAPS, ICGNMD and Gene Matcher. Samples and clinical information were obtained, with informed consent, from each institution using local institutional review board (IRB) ethics for functional analysis of human DNA, fibroblasts and biomaterial.

For genetic analyses, DNA was extracted from peripheral blood. For functional analyses skin biopsy was taken and used for culturing fibroblasts. Where consent was given, pictures of foot deformities and videos of ambulation were taken. All patients underwent examination by a neurologist which, where possible, included nerve conduction studies (NCS) or Electromyography (EMG). Exome sequencing was performed on gDNA at Macrogen or in

participating centres, and raw data was annotated by bioinformaticians. Using in silico predictions of pathogenicity and populational allele frequency we found candidate variants in the gene that were further confirmed by Sanger sequencing. I re-screened our internal database Koios and found an additional family with a biallelic variant in the gene - Leu228His.

4.2.2 Next generation sequencing

Genomic DNA was extracted from peripheral blood samples of subjects and parents according to standard procedures of the centre where blood specimen was collected. Where DNA samples were sent to Professor Houlden laboratory, these were exome sequenced in Macrogen, Korea. Briefly, target enrichment was performed with 2µg genomic DNA using the SureSelectXT Human All Exon Kit version 6 (Agilent) to generate barcoded whole-exome sequencing libraries (fig.4.3). Libraries were sequenced on the HiSeqX platform (Illumina) with 50x coverage. Quality assessment of the sequence reads was performed by generating QC statistics with FastQC.

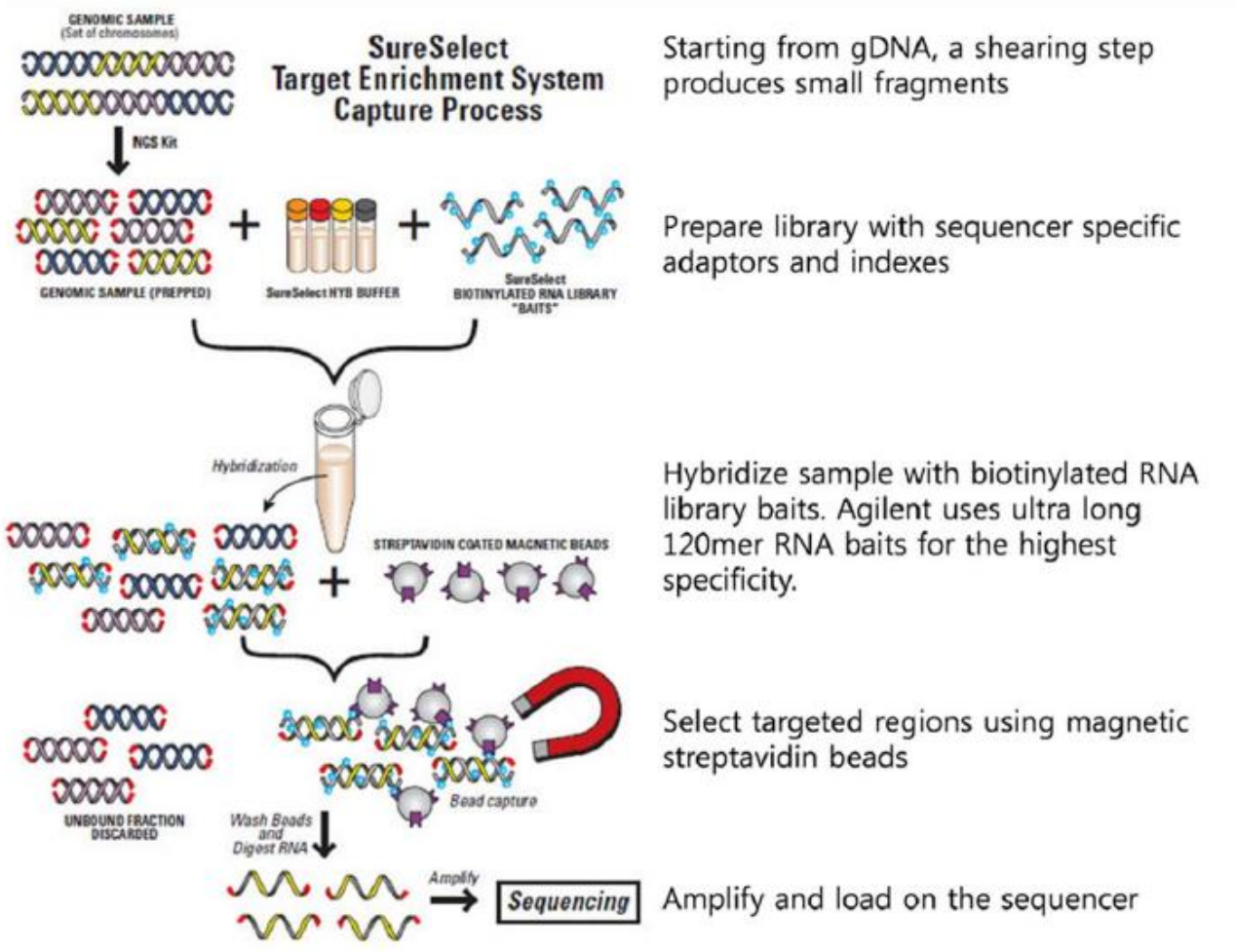


Figure 4.3 DNA library preparation and target enrichment for sequencing with an Illumina sequencing platform. Figure from Macrogen Inc.; https://dna.macrogen.com

4.2.2.1 Genetic analysis

Paired end, 100bp or 150bp reads protocol was used and analysed on Illumina HiSeq4000 or Hiseq X platform. The raw data was then processed by the Bioinformatics team at UCL (Hallgeir Jonvik and David Murphy) using a standardised bioinformatic pipeline with softwares such as bwa-0.7.12, GATKv3.4.0 or SnpEff, which allow the generation of all sequence reads with quality score assessment values for each read, alignment to the reference genome as well as annotated variants.

Typically obtained files from NGS are the following:

- Fastq: Native sequence/raw data file with forward and reverse sequence per sample.
- SAM: Sequences mapped to the reference sequence
- BAM: Binary version of SAM files which can be visualised with the Integrative Genomics Viewer (http://software.broadinstitute.org/software/igv/)
- VCF: Variant calling format is a tab delimited format for storing variant calls and individual genotypes such as SNPs and indels
- Text files: After annotation with ANNOVAR the output can be retrieved in an CSV file that can be filtered with a text editor or also loaded onto Excel.

4.2.2.2 Variant prioritisation

The above pipeline generates a large list of 20000-250000 variants (fig.4.4) which are further filtered according to family pedigree (inheritance), patient phenotype and the genes of interest. Further, pathogenic variants were assessed using the following: confirmed to be conserved across species with score for Genetic evolutionary rate profiling (GERP) greater than 2; genes expressed in central nervous system; genes must be in regions of interest and the variants are predicted damaging with CADD (https://cadd.gs.washington.ed), SIFT (sift.bii.astar.edu.sg), Polyphen-2 (genetics.bwh.harvard.edu/pph2) and Mutation Taster (www.mutationtaster.org); they must have frequency of less than 1% on population databases such as Genome Aggregation Database (gnomAD), dbSNP and Exome Aggregation Consortium (ExAC). If the criteria are met, Sanger sequencing is used for variant confirmation in proband and segregation for allelic state of the variant is performed with the available affected and unaffected members of the family.

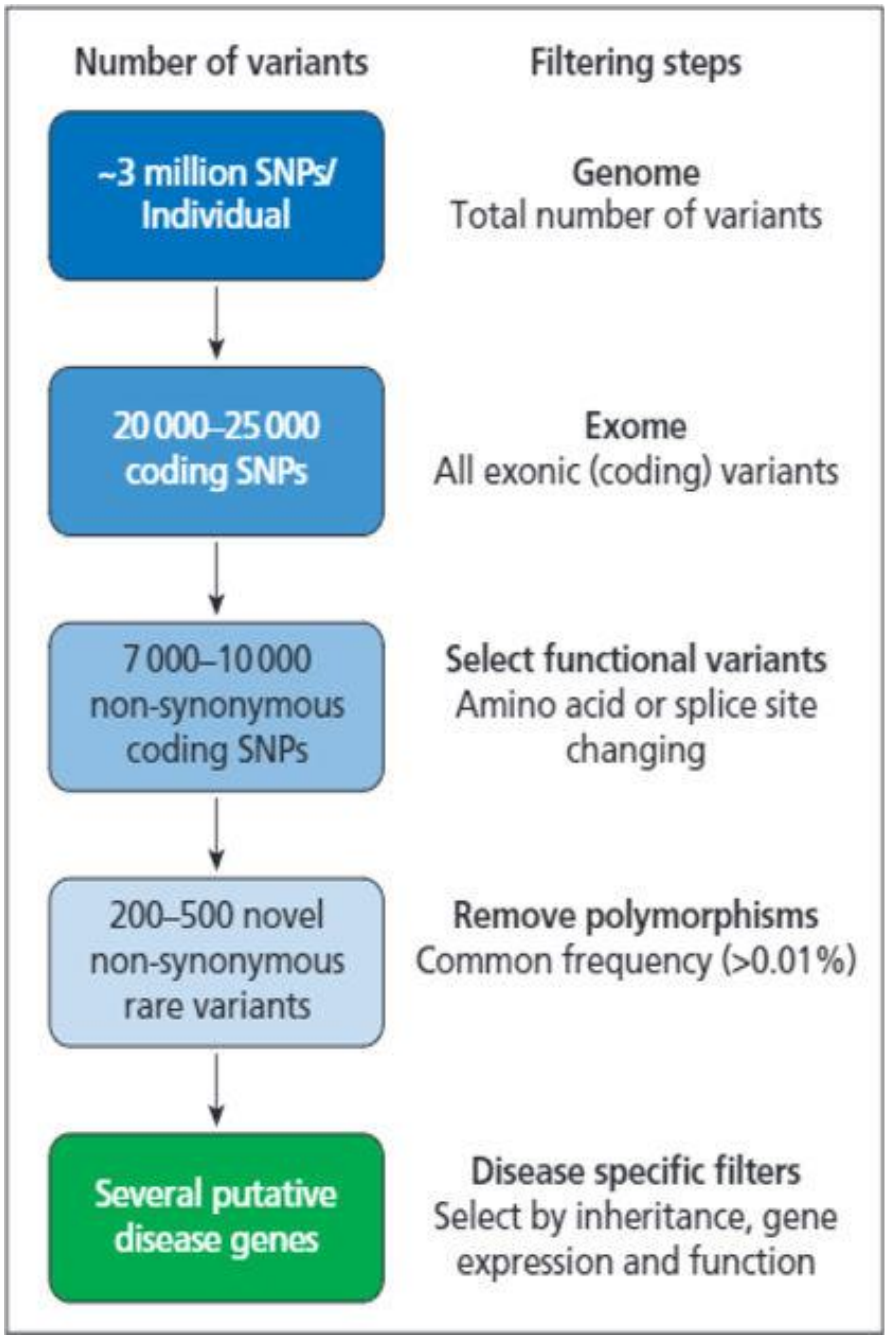


Figure 4.4 Filtering and functional analysis technique for whole exome sequencing. Around 3 million of individual variants can be found in a genome sequencing and 20-25 thousand in an exome sequencing. Through variant filtering strategies and algorithms, several putative disease genes can be found and further investigated for phenotype-genotype correlations. Adapted from Efthymiou *et al*., 2016

4.2.3 Sanger sequencing

Sanger sequencing was performed for the probands and any available family members to confirm the variant found through exome sequencing and to confirm the segregation within family either in house or by participating centres. In house, I performed the following:

4.2.3.1 Primer design

The DNA sequence was downloaded from Ensembl website (http://www.ensembl.org/index.html) either as the longest protein coding transcript or using the transcript ID provided with the variant identified through short read WES or WGS. The variant in question was marked along with approximately 250-500bp flanking regions. The sequence was input into Primer3 (http://primer3.ut.ee/) and the primers were designed by selecting approximately 50bp on each side of the variant by square brackets ([ ]) that must be amplified in PCR. The default parameters were used to generate left and right primers (forward and reverse). The optimum primer size is usually 18-20 bp, melting temperatures of 55 – 65, low self-complementarity and low GC content (less than 50%).

I used BLAT genome search tool on UCSC website (https://genome.ucsc.edu/cgi-bin/hgBlat) to check the specificity of the primers and ensure no other genome fragments will be amplified. Finally, I used Ensembl website to ensure no, or minimal amount of SNP lie in the primer complementary sequence. SNPs could cause issues with primer binding and result in phenomenon called allelic drop out, where only one allele is amplified resulting in false homozygous or wild type result if SNP is present in heterozygous state.

Primers were ordered from Sigma (https://www.sigmaaldrich.com/GB/en) diluted in double distilled MilliQ water at 100uM. The stock solutions were diluted 1:10 for final working concentration of 10uM.

4.2.3.2 Polymerase chain reaction

For each reaction, 7.5µl Fast-start Master Mix (Roche) was mixed with 5µl PCR water and 1µl of 10uM each forward and reverse primer was added. Then 1µl of gDNA of concentration 25-150 ng/µl was added and the reactions were loaded into a thermal cycler. A standard “touch-down” cycler program was used (table 4.2).

94 °C	10mins	
94 °C	30sec	
65 °C	30sec	8 cycles
72 °C	45sec	
94 °C	30sec	
65 °C * -0.7 °C per cycle	30sec	16 cycles
72 °C	45sec	
94 °C	30sec	
55 °C	30sec	16 cycles
72 °C	45sec	
72 °C	10min	
4 °C	Storage	

Table 4.2 Touch down conditions used for PCR reaction of Sanger sequencing

4.2.3.3 Agarose gel electrophoresis

The product was electrophoresed on 1.5% gel agarose to determine whether the PCR was successful (presence of bands of known size). A 50X stock of Tris-Acetate-EDTA (TAE) was diluted 50 times with water to 1-time concentrated working solution. 1.5 grams pf agarose was mixed with TAE and microwaved for approximately 2 minutes until dissolved. 8µl of Sybr Safe (Invitrogen) was added and the solution was poured into gel casting mould after adding the combs. After set, the tank was filled with TAE buffer, PCR products mixed with 6x sample dye and a 100bp ladder and the PCR products were pipetted into the wells. The gel was run at 100V for approximately 20 minutes and visualised in an UV transilluminator.

4.2.3.4 PCR purification and Sanger sequencing

7µl of the product was mixed with 3µl of Exo-Fast (50µl Exo I, 200µl Fast-AP, 750µl Water Roche) for enzymatic clean up and subjected to 37°C for 30min 80°C for 15min in a PCR cycler. The resulting purified PCR product was sent to Source Bioscience for sequencing on ABI 3730 DNA analyser and the data was analysed using Geneious Prime and the sequences were aligned and compared with the reference sequence.

4.2.4 In Silico modelling

In silico modelling was performed collaboratively with Jevin Parmar from Gianina Revenscroft laboratory, Perth, Australia.

4.2.4.1 Protein structure modelling and in silico mutagenesis

The predicted wild-type ARHGAP19 protein structure was retrieved from the AlphaFold Protein Structure Database1 using its UniProt accession code (Q14CB8). Three-dimensional protein structures were visualised using PyMol (v.2.5.2). To examine variant effects on three-dimensional protein structure, in silico mutagenesis for identified missense substitutions was performed through the PyMol 'mutagenesis' function.

For nonsense and frameshift variants, mutant protein structures were generated using the open-source AlphaFold v2.0 (AlphaFold2) pipeline, with the input being FASTA files of the mutant amino acid sequences. The resultant protein structures were then aligned to the wild-type ARHGAP19 protein structure within PyMol.

4.2.5 Homozygosity mapping

Homozygosity mapping was performed by our bioinformatician Dr Shahryar Alavi for the affected individuals that carry the same homozygous mutation and unaffected controls. The patients were as follows: p.H196Qfs*9 (patient P2 (F2-II:5) and P14 (F13-II:3)); p.Leu68Pro (P6 (F6-II:7) and P17 (F15-II:3)); p.Gln151Lys (P8 (F7-II:1) and P15 (F14-II:2)) and p.Leu228His (P13 (F12-II:1) andP21 (F19-II:1)) and the FASTQ files available after WES were merged and plotted with a colour banding of the flanking variants. Homozygosity mapping was performed on Automap (Quinodoz *et al*., 2021) and haplotype analysis was performed by comparing banding patterns of patients and controls.

In the case of founder variants, we estimated the age of the most recent common ancestor (MRCA) using the length of shared haplotypes between patients (Gandolfo *et al*., 2014)

4.2.6 Plasmid design for GAP activity assay

We chose 3 of the mutations found in patients for GAP activity assay Gly140Asp, Gln151Lys and His196*. These mutations were chosen as they lie within the functional GAP domain of the protein thus being most likely candidates to disrupt the protein's function. In addition, Arg143Ala mutation that lies within the GAP domain had previously been shown to disrupt GAP activity in vitro (David *et al*., 2014). *ARHGAP19* constructs were cDNA synthesized by Genescript into pGEX-6P-1 by replacing the CDS for these inserts via BamHI and XhoI (NEB) (fig.4.5), generating: pGEX-6P-1 (GAP_WT); pGEX-6P-1 GAP_G140D; pGEX-6P-1

GAP_G151L; pGEX-6P-1 GAP_His196fs. The plasmids were sent to our collaborators, Nathalie Lamarche-Vane and her Team, for the subsequent GAP activity assays as described below.

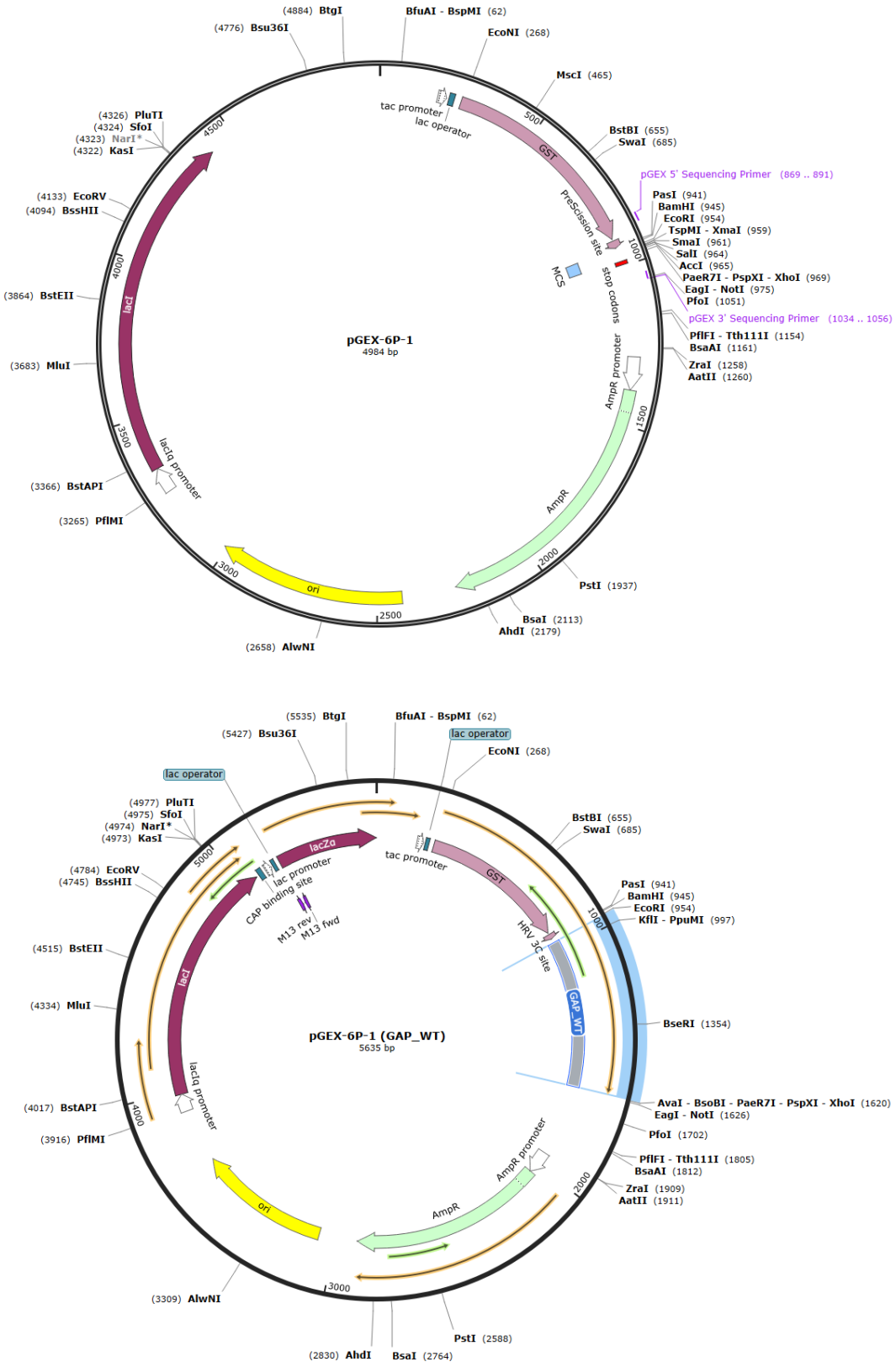


Figure 4.5 Plasmid maps of ARHGAP19 construct using Snapgene software. An empty pGEX-6P-1 plasmid (top) was used as template for all the GAP domains of ARHGAP19 plasmids. An example of wild type (WT) Gap domain is shown (bottom)

4.2.7 GAP activity assay

In vitro GAP assay tests the GAP activity of different mutants by measuring the amount of inorganic phosphate (Pi) that is produced as a result of G-protein dependent hydrolysis of GTP to GDP + Pi.

The synthesized plasmids were subcloned and fused with a Glutathione-S-transferase (GST) domain to generate a fusion protein. The GST tagged GAP domains were transformed into competent *Escherichia coli* cells for protein expression. Bacteria stocks were inoculated into 100ml LB culture with ampicillin and shaken overnight. The culture was diluted 1:10 to fresh LB culture with ampicillin and shaken for 1 hour at 37ºC. Induction with Isopropyl β- d-1-thiogalactopyranoside (IPTG) for 3-4 hours at 37°C, 250rpm followed and cells were harvested by centrifugation at 4°C, 4000x g for 20 mins. The pellet was resuspended with 50mM Tris-HCL pH 7.5, 50mM NaCl, 5mM MgCl2, 1mM DTT and sonicated on ice 15s ON/15s OFF for 6 cycles, the centrifuged at 4°C, 4000 x g for 20 mins. Supernatant was collected and 250µl Glutathione agarose beads for 5ml supernatant was added and incubated for 2 hours. Five washes were performed with 50mM Tris-HCL pH 7.5, 150mM NaCl, 1mM EDTA, 1mM DTT 10µl and PreScission cleavage protease (GenScript) was added into each tube, and all shaken overnight after which it was centrifuged at 4°C, 2000rpm to collect the supernatant. Eluted proteins were concentrated using Amicon Ultra-4 centrifugal filters (MilliporeSigma) and resolved by SDS-PAGE followed by Coomassie blue staining. Proteins were quantified using bovine serum albumin (BSA) as standard.

4.2.8 In vitro p190A GAP assay

The GAP activity of the wild type ARHGAP19 and protein mutants Gly140Asp, Gln151Lys and His196* was assessed using the RhoGAP assay biochem kit (Cytoskeleton) according to manufacturer's instructions and recommendations. P50RhoGAP was used as a positive control. 1.5µg of purified ARHGAP19 protein was mixed with His-RhoA protein and GTP for 20 min at 37°C. CytoPhos reagent was added to the reaction mixture for 10 min at room temperature before measuring the absorbance at 650nm with Infinite M200 Pro Microplate reader (TECAN).

4.2.9 Fibroblast cell culture

Primary skin fibroblast cultures were obtained from patients P5 (F5-II:2) with p.Gly140Asp variant, P6 (F6-II:7) with p.Leu68Pro variant and P11 (F10-II:1) with p.Asn29Asp who consented to a skin biopsy. Primary fibroblasts were cultured in Dulbecco's Modified Eagle's Medium (DMEM, Life Technologies) supplemented with 10% fetal bovine serum (FBS, Hyclone), 1% penicillin and streptomycin (PS, Life Technologies) incubated at 37°C with 5% CO2. LookOut Mycoplasma PCR Detection Kit (Sigma) is routinely used in our laboratory for screening for mycoplasma. Frozen cells are stored in our BioBank in liquid nitrogen, and they are available for further studies together with age matched healthy controls.

The following assays - cell proliferation, migration and wound healing scratch assay - were performed collaboratively in the laboratory of Professor Lamarche-Vane in Canada on fibroblasts of patients P5 (F5-II:2), P6 (F6-II:7) and P11 (F10-II:1).

4.2.9.1 Cell proliferation

Cell proliferation was performed as previously described using the cell growth determination MTT kit (3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) (Abcam) (Popat *et al*., 2005). Briefly, 1500 cells were seeded in triplicates in 96-well plates and grown over a period of 3 days. MTT solution was added to each well for the last 4 hours of treatment on each day as per manufacturer's protocol. Absorbance was measured at 590 nm.

4.2.9.2 Migration assay

Migration assays were performed and analysed as previously described (He *et al*., 2023). 100,000 cells were resuspended in serum-free medium and plated on the top chamber (24-well transwell insert, Falcon). Cells were incubated at 37°C for 24hours, which allowed migration towards the bottom chamber containing complete medium with 10% FBS. Cells on the bottom surface of the insert were fixed in 10% formalin (BioShop) and stained with a crystal violet solution. Ten images were taken for each transwell insert using a Nikon inverted microscope camera with a 10X objective lens (Nikon Eclipse TE300 Inverted microscope). Quantitative analysis was assessed using Image J software. Data represent the fold change relative to that of untreated control cells obtained from at least three independent experiments.

4.2.9.3 Wound healing scratch assay

15,000 cells were seeded in 100 µL of culture medium into sterile transparent 96-well plates and incubated for 24hours. A scratch wound was made in the confluent cell monolayer of each well using the IncuCyte 96-well WoundMaker from Essen Bioscience as described in

manufacturer's manual. After carefully removing the cellular debris, 100µl of culture medium was added to each well. Cell images were captured every two hours using IncuCyte Live-Cell Imaging Systems (Essen BioScience, USA). Images were analysed using the IncuCyte S3 software (2019A) to calculate cell confluency over time.

4.2.10 Motor neuron iPSCs

Induced pluripotent stem cells (iPSCs) were differentiated into motor neurons collaboratively with Dr Andrea Cortese laboratory by using a protocol previously published (Rizzo *et al*., 2023) from patient fibroblasts of P5 (F5-II:2) and P6 (F6-II:7). Briefly, iPSCs were dissociated with accutase and resuspended in a 10mm2 Petri dish to form embryoid bodies (EBs). Differentiation medium consisted of (1:1 DMEM/F12-Neurobasal media, supplemented with N2, B27, 2 mM L-glutamine, 1% Pen-Strep, 0.1 mM β-ME; all from ThermoFisher Scientific), with 10 µM Y-27632 (Tocris), 0.1 µM LDN 193189, 20 µM SB431542, and 3 µM CHIR-99021 (Cambridge Bioscience). Media was replaced every 2-3 days, with the addition of the following small molecules: 100 nM retinoic acid (RA, Sigma Aldrich) from day 2; 500 nM Smoothened Agonist (SAG, Sigma Aldrich) from day 4; 10 µM DAPT (Cambridge Bioscience) from day 9. LDN 193189, SB431542 and CHIR-99201 were discontinued on day 7. The neurotrophic factors BDNF, CNTF and GDNF (Peprotech) were added to the differentiation medium from day 11, at a concentration of 10 ng/mL. On day 14, EBs were dissociated, and post-mitotic neurons were seeded on poly-L-ornithine (Sigma Aldrich) and laminin (Biotechne) coated plates. Eleven days after seeding cells were fixed for immunocytochemistry and harvested for RNA and protein extraction.

4.2.11 Immunocytochemistry

Immunocytochemistry (ICC) was performed on the available fibroblasts of patients 5, 6 and 10 and controls to interrogate possible downstream RhoA effectors at the protein level. We were able to use commercially available ARHGAP19 antibody (Genetex Rabbit, 1:50 and Invitrogen Rabbit, 1:100), vimentin (Sigma Mouse, 1:100) for possible cytokinesis dysfunctions and beta-actin (Merc Mouse, 1:200) for possible actin cytoskeleton changes.

Forty-eight hours after seeding the cells onto coverslips the medium was aspirated and the cells washed three times in PBS and fixed with paraformaldehyde for 15 mins. Permeabilisation with triton (0.2%) in PBS for 1 hour. Three 5-minute washes with PBS followed and the cells were blocked with primary antibody overnight.

After the primary antibody incubation samples were washed 3 times for 5 minutes in PBS. Secondary antibody incubation followed for 2 hours. After three 5-minute washes with PBS, nuclear staining followed for 20 minutes with Hoechst (1:1000 dilution in PBS) and the samples were washed again 3 times with PBS. The coverslips were mounted into slides sing Antifade mountant and imaged with Zeiss LSM 710 confocal microscope.

4.2.12 RNA extraction

I extracted RNA from cultured fibroblasts and iPSCs derived motor neurons using Qiagen RNeasy extraction kit with the optional on column DNase digestion. General work through of the protocol is provided in figure 4.6.

10million cells were harvested as a cell pellet and buffer RLT was added and homogenised by vortexing. The sample was centrifuged for 3 minutes at maximum speed to collect debris as a pellet and the supernatant was transferred to a clean vial. 1 volume of 70% ethanol was added to the sample, mixed by pipetting and the resulting mixture was transferred onto RNeasy Mini spin column in a collection tube. Centrifugation for 15 seconds at >8000g speed followed and the flow through was discarded. 350µl buffer RW1 was added to the column and centrifuged for 15 seconds at >8000g and the flow through discarded. 80µl of DNase (10µl DNase I stock and 70µl buffer RDD) was added onto the column and incubated for 15 minutes at room temperature. The column was washed with buffer RW I and centrifuged for 15 seconds at >8000g. Two more washes followed with 500µl of buffer RPE, first for 15 seconds and second for 2 minutes. Following the washes, the column was centrifuged at full speed for 1 minute to discard any residual flow through. The column was then incubated with 30-50µl of DNase free water and the RNA eluted by centrifugation for 1 minute at maximum speed. RNA concentration and quality was checked on Nanodrop and RNA stored in -80^{o}C until used.

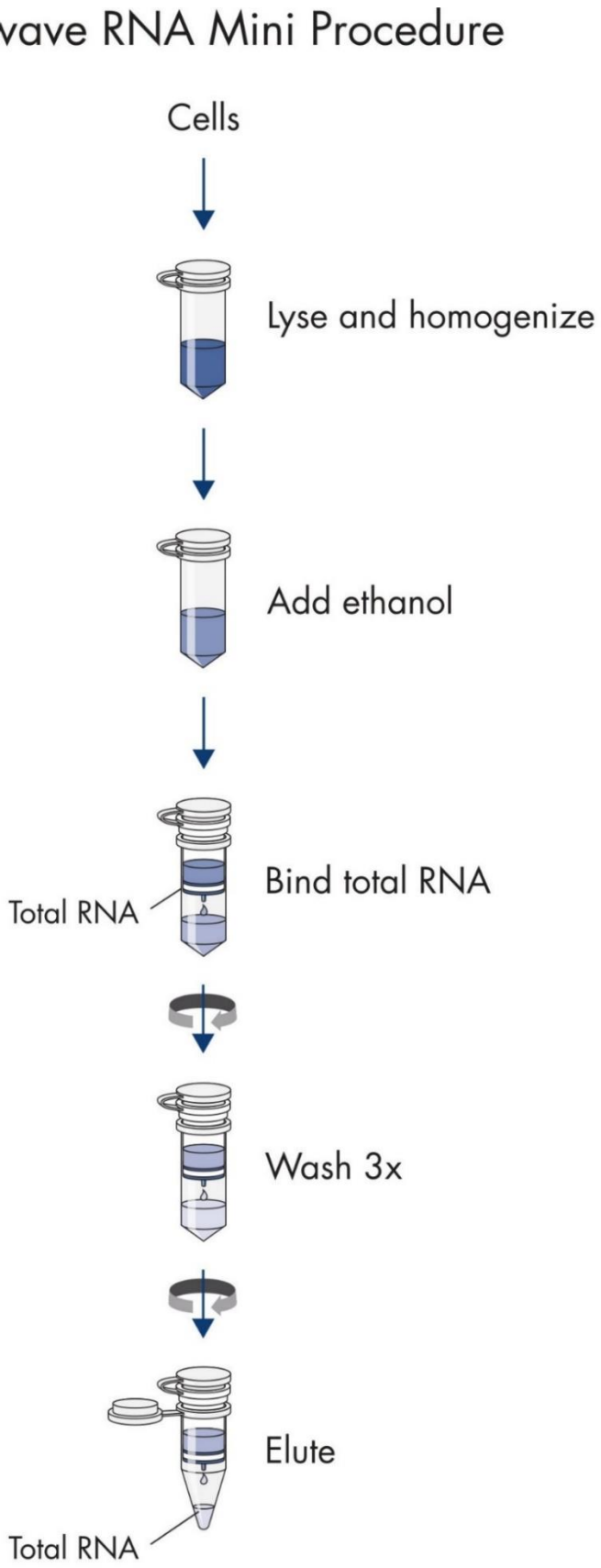


Fig.4.6 Qiagen RNeasy RNA extraction workflow. This procedure relies on lysing and homogenisation of blood cells and subsequent binding of RNA to the matrix in the column which can be washed and eluted and further quantified. Figure from www.qiagen.com

4.2.13 Reverse transcription (RT-PCR)

4.2.13.1 cDNA synthesis

To generate complimentary DNA (cDNA), I used SuperScript® III First Strand Synthesis System kit (Invitrogen). Depending on RNA yield after extractions, 250ng – 1µg of RNA was used for reverse transcription, however, the RNA input was always the same for each sample in the same batch of experiments to ensure the same levels of reverse transcription.

1µl of random hexamers (50ng/µl), 1µl of dNTPs and the correct quantity of RNA were added to a PCR tube and the volume adjusted to 10µl with RNase free water. The mixture was incubated at 65ºC for 5 minutes and placed on ice. Concurrently, cDNA synthesis mix was prepared by mixing 2µl of 10X RT buffer, 4µl of 25mM MgCl2, 2µl of 0.1 M DTT, 1µl of RNase OUT and 1µl of SuperScript. 10µl of the cDNA mixture was added to the RNA on ice and incubated in the following cycler conditions: 10 minutes at 25°C, 50 minutes at 50°C and 5 minutes at 85°C. Subsequently, 1µl RNase H was added and further 20 minutes incubation at 37°C degrees followed.

4.2.13.2 Real time quantitative PCR

Reaction volume of 15µl was prepared by mixing 7.5µl of Fast SYBR Green Master Mix (2x) with 5µl of DPEC water and 0.75µl each forward and reverse primer (10uM) and 1µl of DNA. Samples were plated on 96 well plate in technical duplicates and plate run in QuantStudio Real-Time PCR machine in standard conditions as shown in table 4.3. The primers used are listed in table 4.4.

qPCR cycler conditions	**95°C 20 secs;** **[95°C 1 secs** **60°C 20 secs] x40;** **Ramp rate increase 2.63 °C/sec** **95°C 15 secs** **60°C 1 min** **95°C 15 secs**

Table 4.3 Standard qPCR cycler conditions

Gene expression data analysis was performed in QuantStudio Real-Time PCR Software by comparing ΔΔCT of experimental samples to controls.

Primer name	Primer sequence
ARHGAP19 qPCR forward	GGCATCAAAGGATGACCTTG
ARHGAP19 qPCR reverse	CTTTGCTGACTCTGGCATATC
RPL18A qPCR forward	CCCACAACATGTACCGGGAA
RPL18A qPCR reverse	TCTTGGAGTCGTGGAACTGC

Table 4.4 Primer sequences used in qPCR. RPL18A was used as a housekeeping gene

4.2.14 Western blotting

I performed Western blotting for protein expression level quantification. Firstly, I extracted proteins from fibroblasts of patients P5 (F5-II:2), P6 (F6-II:7) and P11 (F10-II:1) and controls and iPSC motor neurons P5 (F5-II:2) and P6 (F6-II:7) and controls using RIPA buffer with Protease Inhibitor as follows:

Cell pellets were resuspended with 70µl of RIPA buffer with Protease Inhibitor and kept on ice for 1 hour with vortexing every 10-15 minutes. The samples were then centrifuged at 14000rpm for 15 minutes at 4°C and supernatant collected into a fresh tube.

4.2.14.1 Bicinchoninic acid Assay

Protein quantification followed with Bicinchoninic acid (BCA) assay kit (Thermo Scientific).

10µl of supplied protein standards were aliquoted onto 96 well plate in duplicates. The protein lysates were dispensed onto the plate in duplicates by diluting 2µl with 8µl in water. Reagent A and B were mixed at 50:1 dilution for working solution and 190µl was dispensed to protein standards and diluted proteins. The plate was covered in aluminium foil and incubated at 37°C for 30 minutes to be read in Omega plate reader.

4.2.14.2 Western blotting

30µg of protein for each sample and control were mixed with 10µl of MES, 4µl of 10X reducing agent and the mixtures were topped with nuclease free water up to 40µl total volume. The samples were boiled for 5 minutes at 95°C and together with protein ladder, immediately loaded onto western blotting gel (Invitrogen) positioned in a tank with 1X MES SDS running buffer (950ml distilled water, 50ml 20XMES SDS running buffer). The gel was run at 120V for approximately and hour. PVDF membrane was activated for 1 minute in MeOH and subsequently, the transfer was assembled with 1x bolt transfer buffer (100ml MeOH, 850ml distilled water, 50ml 20x BOLT buffer) in the following order: cathode core (-) sponge pad, filter paper, gel, PVDF membrane, filter paper, sponge pad and anode core (+). The transfer was inserted into the chamber and ran for approximately an hour at 20V.

The transfer was checked with Ponceau solution, and the membrane washed twice in TBS-T for 5 minutes. Then the membrane was blocked in 5% milk in TBS-T for an hour in room temperature. Incubation with primary antibody in 2.5% milk in TBS-T followed overnight at 4°C. Primary antibodies and dilutions used were mouse ARHGAP19 (Santa Cruz) at 1:100 and rabbit GAPDH (Abcam) at 1:10000.

After the overnight incubation, the membrane was washed 3 times with TBS-T for five minutes each and digitally imaged using a Machine iBright 750 (Invitrogen) after 1-minute incubation with ECL reagent. Exposure time was optimized to avoid saturation.

4.2.14.3 Protein expression quantification

The raw tif images of the membranes were used for protein expression quantification in Image J. Rectangles of the same size were drawn to encompass the visible bands and the analyse gel function was used to measure the density of the bands expressed as area under the curve. The values were calculated in excel as the ration of protein of interest to the housekeeping protein and normalised to the first band. Data was presented in GraphPad prism and student t-test used for statistical analysis.

4.2.15 *Drosophila melanogaster*

The ortholog of human *ARHGAP19* in the fly is *RhoGAP54D* and I am using the fly as a model organism to investigate the function of the gene. I am covering the methods and tools I used for this work as well as the results in the next chapter 5 titled "*Drosophila melanogaster as model organism for discerning function of RFC1 and ARHGAP19*"

4.2.16 *Danio rerio*

Functional analyses of *ARHGAP19* in zebrafish were performed collaboratively by Renee Lin in the laboratory of Dr Daniel Osborn.

4.2.16.1 Zebrafish Line and maintenance

Wild-type zebrafish, *Danio rerio*, were housed and bred within UCL Fish Facility at 28.5°C on a 14h day/10hour dark cycle. All experiments were conducted under licences awarded by the UK Animal (Scientific Procedures) Act 1986 implemented by the Home Office in England.

4.2.16.2 Whole-mount in situ hybridization

Zebrafish embryos were first fixed in 4% paraformaldehyde (PFA/PBS) overnight at 4°C, dechorionated, and dehydrated in methanol at -20°C. cDNA fragment of *arhgap19* was amplified from cDNA library at three different developmental stages: 24 hours post-fertilization (hpf), 48 hpf, and 5 days post-fertilization (dpf). Digoxigenin-labelled sense and

antisense probes were synthesized using DIG-RNA Labelling Kit (T7 polymerases, Roche). Whole-mount in situ hybridization (WISH) was carried out as previously described (Coutelle *et al*., 2001) on 48 hpf, 72 hpf, and 5 dpf larvae. For the colour reaction, NBT/BCIP Stock Solution (Roche) was used in the staining buffer. The whole embryos were mounted in methanol and imaged on light microscope.

4.2.16.3 sgRNA/ Cas9 mRNA Synthesis and Microinjection

The zebrafish *arhgap19* gene (ZDB-GENE-100922-157) is an ortholog of Homo sapiens *ARHGAP19*; there is 59.95 % nucleotides similarity and 55 % amino acid similarity between the zebrafish and human loci. No zebrafish paralogs corresponding to *ARHGAP19* exist. CRISPR/Cas9-mediated F0 biallelic knockout was performed as described previously (Kroll *et al*., 2021). Six crRNAs were designed according to their on-target and off-target scores. The six sgRNAs sequences were: 1) 5'-ACGCTCCTCAAGAGTTTCTT-3', 2) 5'-CAAGATGTCTGCTCACAACC-3', 3) 5'- CAAGAGTTTCTTGGGAGAAT-3', 4) 5'-GAACCCAAGACTCCCAACGC-3', 5) 5'-TGACTTTCATCCCAATGACG-3', 6) 5'-AGGCAACAAGACGAGTTTTC-3'. Each cRNA was annealed to tracrRNA and complexed to Alt-R S.p. HiFi Cas9 nuclease to form a ribonucleoprotein (RNP). A pool of three RNPs were co-injected into one to two-cell stage zebrafish embryos. For each RNP pool, 0.5nL and 1nL volumes were injected into the yolk of batches of embryos. At 5 dpf, three zebrafish larvae were randomly selected for genomic DNA extraction to determine the targeting efficiency by Illumina Miseq (Eurofins). PCR amplification targeting RNP cut sites were performed for each crRNA and products purified, pooled and sent for next generation sequencing. Primers included universal adaptors for MiSeq

4.2.16.4 Morpholinos and Microinjection

An antisense translational blocking Morpholino (MO, GeneTools.LLC) against the AUG-containing mRNA sequence (5′-GGCCATCTTTCATCTTCCGTTTGAA-3′) and a splice MO targeting the Exon1-Intron1 (E1I1) boundary (5′-ATAAATCTTCGTTACCTTCTGTCTC-3′) was designed to knockdown *arhgap19* function. MOs were diluted to the desired working concentrations (3 ng, 4 ng, 6 ng, and 8ng per embryo) before use. Microinjection was performed by injecting 0.5-1nL morpholino solution into one to two-cell stage embryos in the yolk. Dose dependent phenotyping was used to identify an appropriate concentration that balanced survival with specific phenotypic changes.

4.2.16.5 Behavioural assays

Zebrafish larvae at 5 dpf were transferred into individual wells of a 96-well plate. Baseline locomotor activity was recorded for a duration of 30 minutes and analysed using the DanioVisionTM monitoring chamber, which was integrated with the EthoVision XT 14 video tracking software (Noldus, Netherlands). Plots were analysed for distance travelled (in millimetres), total time spent and velocity.

4.2.16.6 Assessment of Muscle Integrity

For assessment of muscle integrity, zebrafish larvae at 5 dpf were fixed and their skeletal muscle was analysed for total birefringence using polarized light microscopy on a Nikon SMZ1000 stereo microscope.

4.2.16.7 Immunohistochemistry

Zebrafish larvae at 5 dpf were permeabilized with PK, followed by fixation in 4% PFA. Larvae were then blocked with goat serum (5% goat serum in PBT), and incubated with anti-tubulin mouse monoclonal antibody (1:500; Sigma) overnight at 4°C. After four washing steps, larvae were incubated with Alexa Fluor™ anti-mouse 568 secondary antibody (1:1000 dilution; ThermoFisher) in the dark at 4°C. For nucleus detection, larvae were incubated with DAPI (1:1000 dilution; ThermoFisher) under dark conditions at 4°C. Stained larvae were imaged using a Nikon confocal microscope to assess motor neuron morphology. Confocal fluorescent images were processed and adjusted with FIJI/ImageJ.

4.2.16.8 Statistical Analysis

Statistical evaluations were carried out using IBM SPSS Statistics. One-way ANOVA and Welch's t-test were used for statistical comparisons between groups with calculated standard deviations/errors of the mean. Significant differences were determined at a threshold of $p < 0.05$.

4.3 Results

Through international collaborations we have collected 25 affected individuals from 20 unrelated families with *16* biallelic variants in *ARHGAP19*, with four (4/16, 25%) variants occurring in more than one family.

4.3.1 Biallelic *ARHGAP19* variants

WES carried in the index patient PT3 (F3-II:1) revelled a biallelic C to T substitution at position c. 1243 resulting in introduction of premature stop codon ENST00000358531.9 c.1243 C>T p.Gln415*. This variant was prioritised based on the in-silico predictions with CADD score of 38 and Mutation Taster predicting nonsense mediated decay as well as very low frequency of the variant seen in populations according to publicly available databases <0.0001. Sanger sequencing analysis confirmed WES results and showed segregation within the family with both the parents having the variant in heterozygous state.

Through our network of collaborations including SYNAPS and ICGNMD as well as GeneMatcher, we identified 18 additional families with clinical phenotypes resembling that of the index patient and the variants identified together with their population frequencies and in silico prediction scores are shown in table 4.5.

FAMILY ID	Family 1	Family 2 & 13	Family 3	Family 4	Family 5
VARIANT DESCRIPTION					
Genomic Position Change (GRCh38/hg38)	chr10:97265921	chr10:97263448-97263450	chr10:97235258-97235258	chr10:97235282-97235282	chr10:97263614-97263614
Genomic Position Change (GRCh37/hg19)	chr10:99025678	chr10:99023205-99023207	chr10:98995015-98995015	chr10:98995039-98995039	chr10:99023371-99023371
Coding Sequence Change (NM_001136035)	c.261dup	c.585dupA	c.1243C>T	c.1219C>T	c.419G>A
Protein Sequence Change (NP_001129507)	p.Pro88Alafs*43	p.His196 Thrfs*10	p.Gln415*	p.Arg407*	p.Gly140Asp
Exon Number Position	2 of 12	4 of 12	9 of 12	9 of 12	4 of 12
Codon Change	C/CC	_/A	CAA/TAA	CGA/TGA	GGT/GAT
Consequence	Insertion	frameshift	stop gained	stop gained	missense
Zygosity	Hom	Hom	Hom	Hom	Hom
dbSNP ID	rs767899310	rs772718801	rs757781028	rs751754099	-
ALLELE FREQUENCY [Allele count/Total Allele Number]					
gnomAD v3.1.2 (Highest subpopulation if applicable) (76,156 genomes)	0.00001315 (0.0000294 Non-Finnish European) [2/152076]	0	0.00001315 (0.00003267 South Asian) [1/251376], (0.000008795 Non-Finnish European) [1/251376]	0.00003287 (0.0002621 Latino/Admixed American) [4/152091], (0.0000147 Non-Finnish European) [1/152091]	0
gnomAD v2.1.1 (Highest subpopulation if applicable) (15,708 genomes and 125,748 exomes)	0.00003181 (0.00006154 Non-Finnish European) [7/251464]	0.00002784 (0.0001848 Finnish) [4/251439], (0.00002638 Non-Finnish European) [3/251439]	0.00001591 (0.0004139 South Asian) [2/152140]	0.00003978 (0.0000868 Latino) [3/251344], (0.00006533 South Asian) [5/251344], (0.00004399 Non-Finnish European) [2/251344]	0
Ensembl (Highest frequency observed)	< 0.01	< 0.01	< 0.01	< 0.01	N/A
Iranome (~800 exomes)	0	0	0	0	0
GME Variome (~2,500 exomes)	0	0	0	0	0
UK Biobank (394,841 exomes)	0	0	0.00000271782	0	0
TOPMed (132,345 genomes)		0.000007964	0	0.000015928	0
54KJPN (~54,000 genomes)	0.000023891	0	0	0	0
QSG Database (23741 exomes)	0	0	1 hom	0	0
EDICTIONS AND CLASSIFICATION					
GERP	3.11	2.64	1.57	-1.56	5.87
CADD_Phred	-	34	39	35	26.3
Polyphen-2	-	-	-	-	PD 1
SIFT	-	-	-	-	D 0
PROVEAN	-	D-14.45	N -1.66	N -2.25	D -6.72
MutationTaster	-	DC 1 NMD	DC 0.99 NMD	DC 0.99 NMD	DC 0.99

FAMILY ID	Family 6 & 15	Family 7 & 14	Family 8	Family 9	Family 10
VARIANT DESCRIPTION					
Genomic Position Change (GRCh38/hg38)	chr10:97265979	0:97263582-97263	chr10:97292627	chr10:97263554	chr10:97266097
Genomic Position Change (GRCh37/hg19)	chr10:99025736	0:99023339-99023	chr10:99052384	chr10:99023311	chr10:99025854
Coding Sequence Change (NM_001136035)	c.203T>C	c.451C>A	c.1A>G	c.479del	c.85A>G
Protein Sequence Change (NP_001129507)	p.Leu68Pro	p.Gln151Lys	p.Met1?	p.Asn160Metfs*21	p.Asn29Asp
Exon Number Position	2 of 12	4 of 12	1 of 12	4 of 12	2 of 12
Codon Change	CTC/CCC	CAG/AAG	ATG/GTG	AAT/AT	AAT/GAT
Consequence	missense/NMD	missense	start lost	frameshift	missense
Zygosity	Hom	Hom	Hom	Hom	Hom
dbSNP ID	rs1026767404	rs1842859321	-	-	rs754312797
ALLELE FREQUENCY [Allele count/Total Allele Number]					
gnomAD v3.1.2 (Highest subpopulation if applicable) (76,156 genomes)	0.000006572 (0.00002414 African/African American) [1/152165]	0	0	0	0
gnomAD v2.1.1 (Highest subpopulation if applicable) (15,708 genomes and 125,748 exomes)	0.000003976 (0.00002891 Latino) [1/251491]	0	0	0	0
Ensembl (Highest frequency observed)	< 0.01	< 0.01	N/A	N/A	< 0.01
Iranome (~800 exomes)	0	0	0	0	0
GME Variome (~2,500 exomes)	0	0	0	0	0
UK Biobank (394,841 exomes)	0	0	0.0000298948	0	0.00002174138
TOPMed (132,345 genomes)	0.000015928	0	0	0	0
54KJPN (~54,000 genomes)	0	0	0	0	0
QSG Database (23741 exomes)	0	0	0	0	0
EDICTIONS AND CLASSIFICATION					
GERP	2.63	3.94	2	-	2.63
CADD_Phred	28	26.2	22.6		23.7
Polyphen-2	PD 0.987	PD 0.996	B 0	-	PD 0.57
SIFT	D 0.01	D 0	D 0	-	D 0.21
PROVEAN	D -2.75	D -3.58	N -0.68	N -1.15	N -0.28
MutationTaster	DC 0.99	DC 0.99	DC 1	DC 1 NMD	DC 0.503

FAMILY ID	Family 11	Family 12 & 19	Family 16	Family 17	Family 18	Family 20
VARIANT DESCRIPTION						
Genomic Position Change (GRCh38/hg38)	chr10:97246333	chr10:97259559	chr10:97263611	chr10:97259525	chr10:97263470	chr10:97259424
Genomic Position Change (GRCh37/hg19)	chr10:99006090	chr10:99019316	chr10:99023368	chr10:99019282	chr10:99023227	chr10:99019181
Coding Sequence Change (NM_001136035)	c.932C>G	c.683T>A	c.422T>G	c.717T>A	c.563del	c.818C>T
Protein Sequence Change (NP_001129507)	p.Pro311Arg	p.Leu228His	p.Leu141Trp	p.Asn239Lys	p.Pro188Argfs*5	p.Pro273Leu
Exon Number Position	7 of 12	5 of 12	4 of 12	Exon 5 of 12	4 of 12	5 of 12
Codon Change	CCT/CGT	CTC/CAC	TTG/TGG	AAT/AAA	CCT/CT	CCC/CTC
Consequence	missense	missense	missense	missense	missense	missense
Zygosity	Hom	Hom	Hom	Hom	Hom	Hom
dbSNP ID	-	rs562006800	rs755015441	-	rs760101189	rs1162465048
ALLELE FREQUENCY [Allele count/Total Allele Number]						
gnomAD v3.1.2 (Highest subpopulation if applicable) (76,156 genomes)	0	0.00002628 (0.001 South Asian) [4/152192]	0	0	0.000006576 (0.0002076 South Asian) [1/152065]	0
gnomAD v2.1.1 (Highest subpopulation if applicable) (15,708 genomes and 125,748 exomes)	0	0.00009543 (0.001 South Asian) [24/251464]	0.000003979 (0.00003267 South Asian) [1/251345]	0	0.00005568 (0.0004573 South Asian) [14/251438]	0
Ensembl (Highest frequency observed)	N/A	< 0.01	< 0.01	N/A	< 0.01	< 0.01
Iranome (~800 exomes)	0	0	0	0	0	0
GME Variome (~2,500 exomes)	0	0	0	0	0	0
UK Biobank (394,841 exomes)	0	0	0	0	0	0.0000013589
TOPMed (132,345 genomes)	0	0	0	0	0.000007964	
54KJPN (~54,000 genomes)	0	0	0	0	0	0.000007964
QSG Database (23741 exomes)	0	2 hom	0	0	0	0
EDICTIONS AND CLASSIFICATION						
GERP	5.66	2.46	3.94	2.97	3.94	5.11
CADD_Phred	27.5	27.6	29.2	-	27	27.1
Polyphen-2	PD 1	PD 0.998	PD 1	-	PD 0.998	-
SIFT	D 0	D 0	D 0	0.004	D 0	0.003
PROVEAN	D -7.95	D -6.07	D -5.33	D -4.93	-	D -9.23
MutationTaster	DC 0.99	DC 0.99	DC 0.99	DC 0.99	DC 0.99	DC 1

Table 4.5 ARHGAP19 variants, population frequency and in silico prediction scores. 16 variants that were found in our ARHGAP19 cohort are shown in the table, including their position in the genome and in the specific exon of ARHGAP19 they lie in. Further, the consequence of the variant is described as missense start loss or frameshift. Population databases were checked for the frequency of the variant and finally in silico prediction scores are shown for each variant where available: PD=possibly damaging, D=damaging, DC=disease causing.

Apart from one patient PT8 (F8-II:2) who had a p. Met? Variant in the non-canonical *ARHGAP19* transcript - ENST000000371027.1; all other patients harboured biallelic variants in the canonical ENST00000358531 (fig.4.7 A). Four mutations occurred in more than one

family p.H196Qfs*9, p.Leu68Pro, p.Gln151Lys, and p.Leu228His found in 2 independent Arab, 4 independent Turkish, and 2 independent Bangladeshi/Afghani families, respectively. Nine mutations occurred within the functional GAP domain: p.Gly140Asp, p.Leu141Trp, p.Gln151Lys, p.Asn160Metfs*21, p.Pro188Argfs*5, p.His196Glnfs*5, p.Leu228His, p.Asn239Lys and Pro273Leu. Seven mutations occurred outside the GAP domain: Four mutations in the N-terminus of the protein are p.Met?, p.Asn29Asp, p.Leu68Pro, Pro88Alafs*43 and three mutations are in the C-terminus: p. Pro311Arg, p. Arg407* and p. Gln415* (fig.4.7 B).

4.3.1.1 Species conservation

Protein sequence conservation can be assessed and compared to the human protein sequence if the given species has an ortholog of the protein in question. High inter-species conservation of amino acids indicates its maintenance throughout natural selection and importance in biological processes.

The conservation of ARHGAP19 (fig.4.7 C) was assessed using CLUSTAL Omega (1.2.4) multiple sequence alignment in human and gene orthologs in *Pan troglodytes* (chimpanzee), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Bos taurus* (bovine) *Gallus gallus* (chicken), *Xenopus laevis* (frog) and *Drosophila melanogaster* (fruit fly). Especially, the affected amino acids were interrogated (fig.4.7 C) and a high conservation was observed for majority of the variants. Only Asn160, Arg407 and Gln415 appear less-well conserved at approximately 50% within the interrogated species; these three amino acids are the positions where frameshift or nonsense mutations occurred in our cohort.

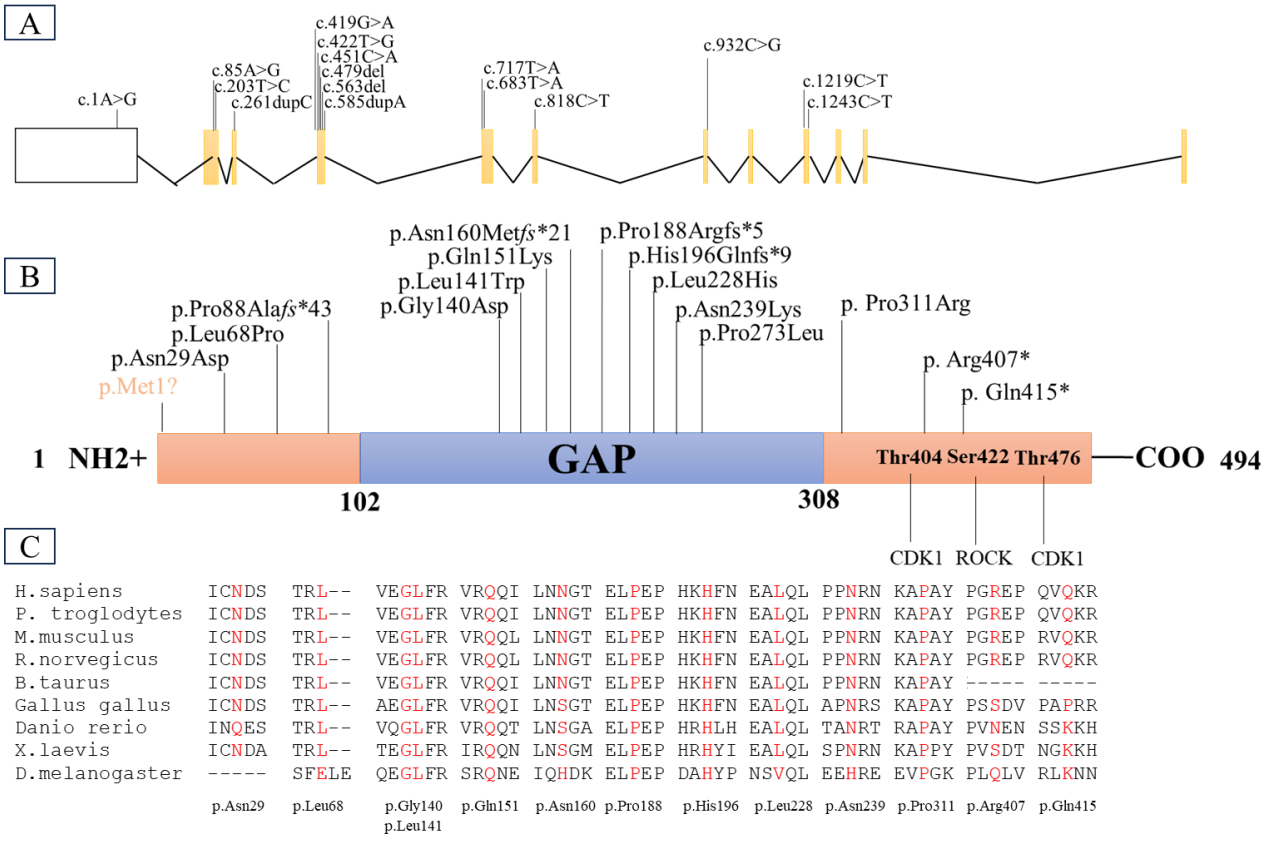


Figure 4.7 ARHGAP19 is a small G protein of less than 500 amino acids. It contains evolutionarily conserved GAP domain (purple) between 102 and 308 amino acids. Below are mutations found in biallelic state in our cohort. One mutation (in orange) p.Met? appears in a non-cannonical ARHGAP19 transcript. Four mutations: p.Leu68Pro, p.Gln151Lys, p.H196Qfs*9 and p.Leu228His appear in more than one family. 9 of the mutations lie within the GAP domain and 7 outside the domain.

4.3.2 Clinical and demographic data

Table 4.6 and figure 4.8 summarise the core phenotypic features of 25 individuals from 20 families included in our cohort. The affected persons were aged between 11 months and 30 years at the first examination. A total of 16 patients were female and 9 (36%) were male and originated from a wide range of ancestral backgrounds and countries of origin included Pakistan, Egypt, Turkey, Syria, Bangladesh, Spain, Italy, Iran, Dubai, Brazil and Afghanistan. However, patients from Turkey and Brazil were most represented.

Demographic	
M:F ratio	0.56:1
Consanguinity	80%
Age at onset	11mnths - 30y (median 10 years)
Deceased	8%
Symptoms	**At onset**
Areflexia	14/16 (88%)
Foot deformities	16/20 (80%)
Foot drop	9/15 (60%)
Sensory symptoms	14/22 (64%)
Motor deficit of the lower limbs	21/23 (91%)

Table 4.6 Demographic information and prevalence (as %) of primary clinical symptoms presented in patients in our cohort

For the majority of patients, the age of onset was in pre-teen to teen years (9-20 years), two patients (PT6 (F6-II:7) and PT13 (F12-II:1)) were examined in infancy and one patient PT11 (F10-II:1) reported a later age of onset at 30, however, a recollection bias could not be excluded. Consanguinity was reported in 80% (16/20) families.

The core complaint between the patients was distal weakness in lower limbs, progressive difficulties in walking and foot deformities as well as areflexia, apart from one patient who reported leg cramps as the first presenting symptom. Patients had various degrees of foot deformities, ranging from mild to well pronounced, which included pes cavus or flat feet and hammer toes. Stork leg appearance was also noted and was a result of leg muscle atrophy. 60% of patients had foot drop and 2 patients had upper limb involvement.

The clinical phenotype of *ARHGAP19* patients, was assessed by a colleague neurologist, Dr Christopher Record, who examined and combined clinical and neurophysiological data and observed the following: The mean (median) age at symptom onset (AAO) was 9.9 (10.0) years, and at assessment was 22.8 (16.0) years. The presenting symptom was a motor deficit of the lower limbs in 91% (21/23), and 64% (14/22) had some form of clinical sensory involvement (symptoms or signs), but as the disease progressed it typically remained either exclusively motor or motor predominant. Patients typically had a length-dependent pattern (17/23, 74%, which includes those with only distal lower limb involvement) of lower motor neuron signs of areflexia and muscle atrophy, with foot drop. Lower limb-predominant disease was seen (distal > proximal weakness, with normal upper limbs) in 17% (4/23) and upper limb predominant

disease seen in 9% (2/23). Foot deformity was present in 80% (16/20). However, the presence brisk knee jerks, and preserved lower limb reflexes each in 9% (3/23) respectively suggests some mild UMN involvement in these individuals. There were no consistent features outside of the peripheral nervous system.

A prominent feature of the phenotype is its significant asymmetry in terms of limb involvement, seen at onset or at assessment in 61% (14/23). Two cases presented acutely with upper limb weakness on a background of mild or subclinical widespread neuropathy. Neurophysiology was performed in 20 individuals. Detailed numerical study data was available in 15/20; five cases had a report only. All had a motor neuropathy, with variable sensory involvement. Evidence of motor conduction slowing was seen or described in 50% (10/20) and conduction block in 20% of the studies with numerical data (3/15) clinical table is available in appendix 2.

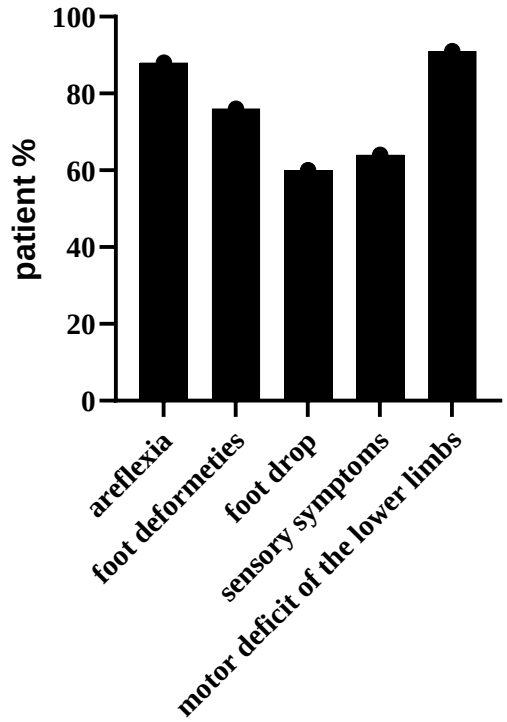


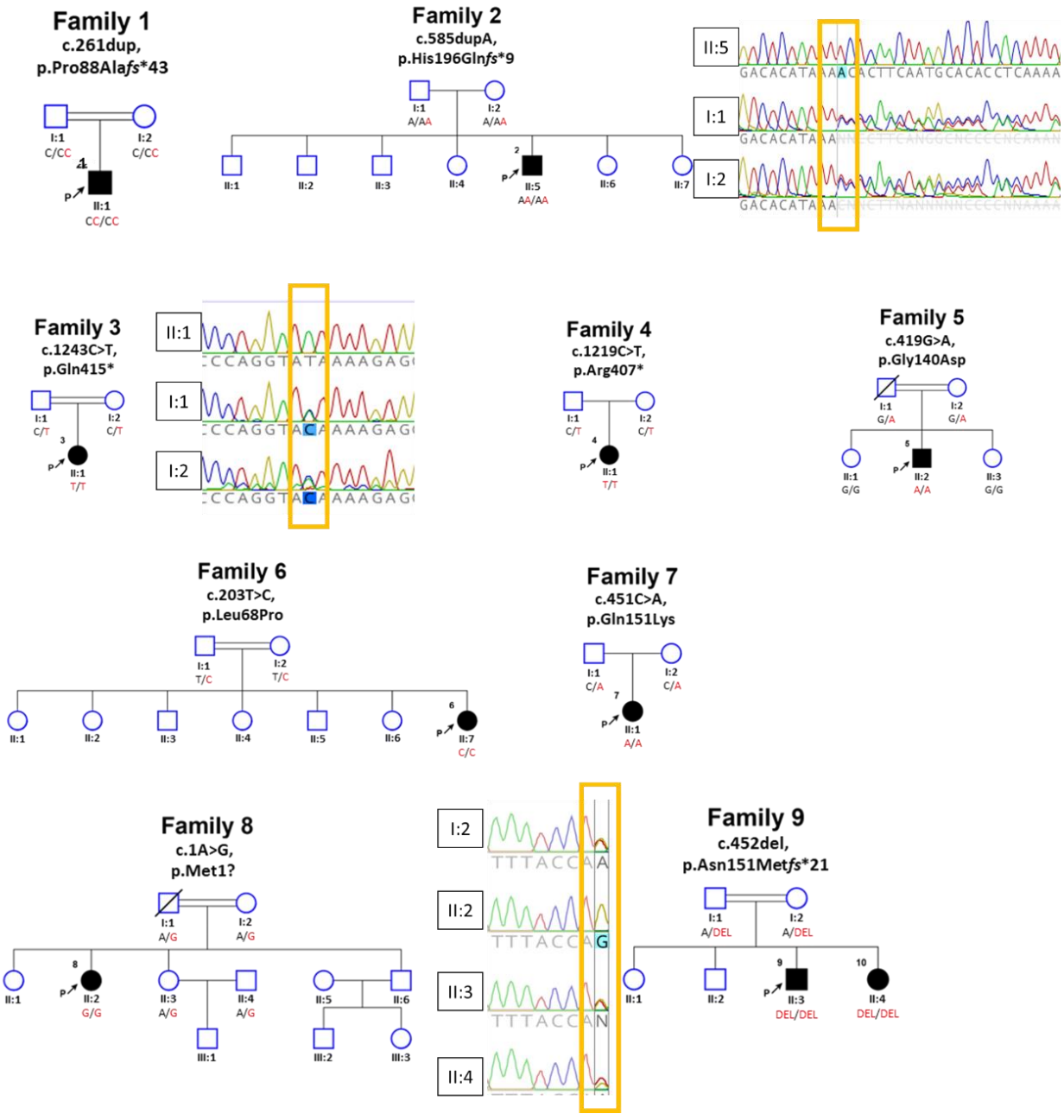
Figure 4.8. Symptoms and signs of patients with biallelic *ARHGAP19* mutations shown as percentage of all cases.

4.3.2.1 Pedigrees and Sanger segregation

The pedigrees of families included in our cohort are presented in figure 4.9 together with the Sanger segregation results across family members. Seven centres shared patient and family DNA for Sanger segregations to be carried out by myself in Houlden Laboratory, segregation in family 10 was not possible due to deceased status of the parents and segregation for the

remaining families were performed in the collaborating centres by means of Sanger sequencing or trio-WES/WGS.

Performing Sanger sequencing is important not only for segregation of the variant within family, but also to confirm the results of WES or WGS. All the *ARHGAP19* variants found in the affected individuals were homozygous whereas the parents carried the corresponding variant on one allele only. No unaffected individuals were found in homozygous state. These results confirm that the homozygous variants segregate with the disease.



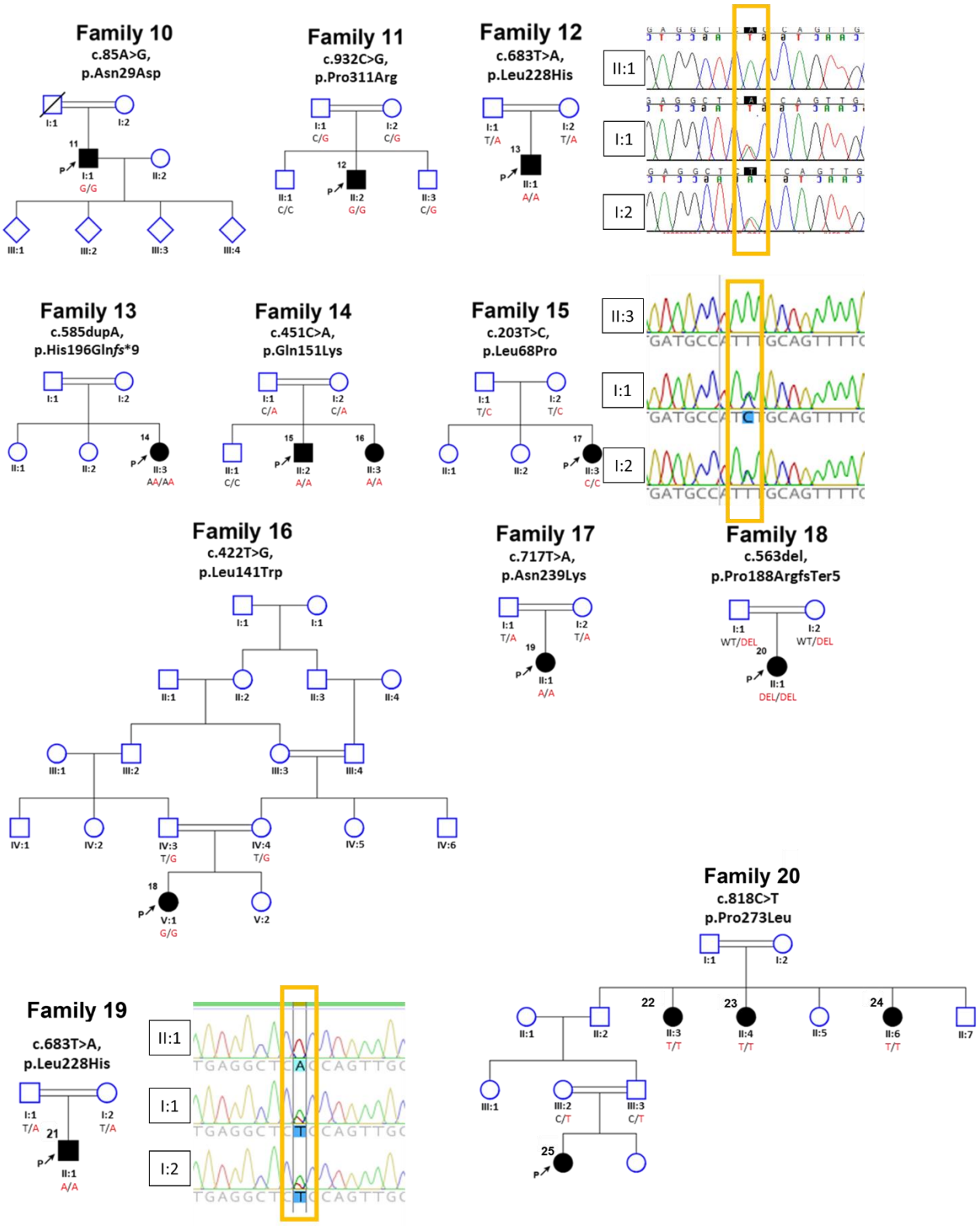


Figure 4.9 ARHGAP19 Family pedigrees. 25 affected individuals with biallelic variants in *ARHGAP19* were collected. Probands are marked with an arrow and letter P. Sanger sequencing traces are included for families where sequencing was performed by myself. Yellow boxes in Sanger traces indicate the variant position.

4.3.3 Ancestorial founder effect

Four mutations occurred in more than one independent, not-related family namely variants p.Leu68Pro, p.Gln151Lys, p.H196Qfs*9 and p.Leu228His.

p.Leu68Pro and p.Gln151Lys occurred in 2 Turkish families each, p.H196Qfs*9 in 2 families originating from Egypt and Anatolia; and p.Leu228His in families from Bangladesh and Afghanistan.

Therefore, homozygosity mapping and haplotype analysis were performed on the raw genetic data from the affected sequenced individuals for each variant and unaffected controls.

Interestingly, haplotype analysis of the p.Gln151Lys and p.Leu68Pro variants identified shared 3.3Mbp and 3Mbp regions respectively, indicating a possible founder effect for these variants, probably originating in Turkey given the origin of the families. Most recent common ancestor (MRCA) analysis further estimated that these founder variants arose ~590 years ago (29.5 generations ago) and 220 years ago (11 generations ago) respectively.

A third recurring variant, p.His196Glnfs*9, showed a small size of shared ancestral haplotype, making it unlikely that this variant is from a recent common ancestor and possibly suggesting two independent *ARHGAP19* mutational events within Arabian Middle Eastern populations. Similarly, a fourth recurring variant p.Leu288His is unlikely to have arisen from a recent common ancestor therefore this variant is recurrent (fig 4.10).

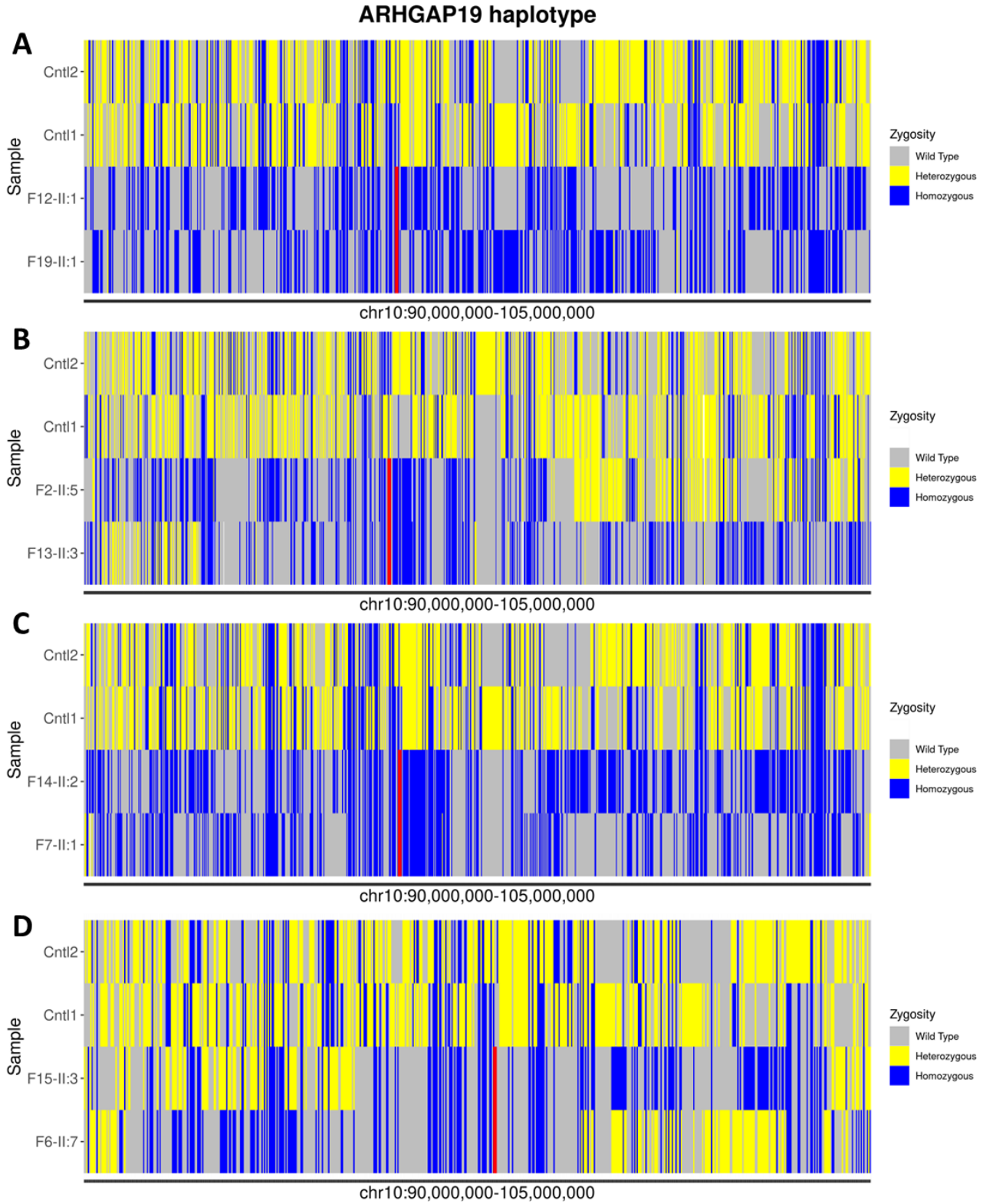


Figure 4.10 Haplotype analysis of ARHGAP19 variants A) chr10-97259559-A-T (p.Leu228His) B) chr10:97263447-G-GT (p.His196Thrfs*10) C) chr10-97263582-G-T (p.Gln151Lys) and D) chr10-97265979-A-G (p.Leu68Pro). Variations pattern of flanking regions of shared pathogenic variants was coloured as follows: homozygous variants as blue, heterozygous variants as yellow, and WT locus as grey. The position of the pathogenic variant is highlighted in red.

4.3.4 In silico ARHGAP19 variant predictions

The in-silico modelling was performed collaboratively by Gina Revenscroft laboratory.

4.3.4.1 Three-dimensional visualisation of ARHGAP19 variants

Figure 4.11 A depicts visualisation of the wild-type ARHGAP19 protein as predicted by AlphaFold2. AlphaFold2 generated mutant protein structures for frameshift variants show substantial deviation from the wild-type protein (fig.4.11 B), likely disrupting the Rho-GAP domain structure. Missense substitutions modelled using AlphaFold2 (fig.4.11 C) display a similar structure to the WT protein, with greater variation only occurring in the final alpha-helix and the subsequent C-terminal end of the protein.

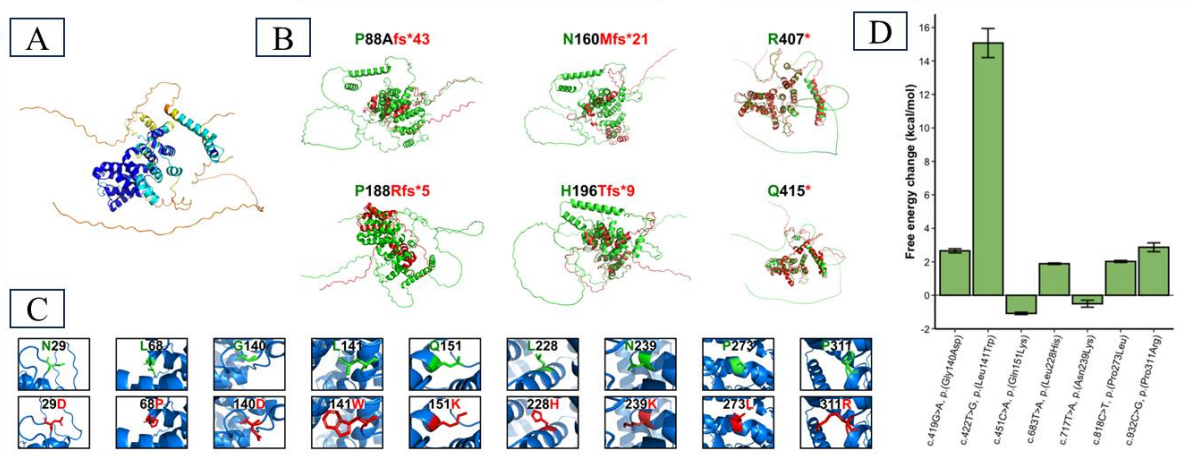


Figure 4.11 AlphaFold2 protein structure predictions. A) Wild type ARHGAP19 B) frameshift ARHGAP19 variants C) missense ARHGAP19 variants D) Free energy calculations for missense ARHGAP19 variants.

Missense substitutions showed little or no change when modelled with the AlphaFold2 pipeline. As such, the 'mutagenesis' function on PyMol was used to predict changes in protein structure and/or folding upon substituted amino acid residue incorporation into the sequence (fig.4.11 C). All substitutions showed changes in steric hindrance with nearby amino acid residues. The calculated free energy changes for p.(Gly140Asp), p.(Leu141Trp), p.(Gln151Lys), p.(Leu228His), and p.(Pro311Arg) substitutions show a decrease in free energy >1.6 kcal/mol, indicating a protein destabilising effect (fig.4.11D). Notably, three substitutions are predicted to result in protein instability with high confidence (p.(Gly140Asp), p.(Leu141Trp), and p.(Pro311Arg). In addition, the p.(Gln151Lys) variant shows an increase in free energy and thus an increase in protein stability. Using further analysis with AlphaMissense, of the nine missense variants identified, eight (89%) are predicted to be likely pathogenic (table 4.7)

Variant (c.)	AA change (p.)	AlphaMissense pathogenicity**	AlphaMissense Class
c.85A>G	p. Asn29Asp	0.0882	likely_benign
c.203T>C	p.Leu68Pro	0.9953	likely_pathogenic
c.419G>A	p.Gly140Asp	0.9985	likely_pathogenic
c.422T>G	p.Leu141Trp	0.9915	likely_pathogenic
c.451C>A	p.Gln151Lys	0.7953	likely_pathogenic
c.683T>A	p.Leu228His	0.8643	likely_pathogenic
c.717T>A	p.Asn239Lys	0.9058	likely_pathogenic
c.818C>T	p.Pro273Leu	0.9793	likely_pathogenic
c.932C>G	p.Pro311Arg	0.9945	likely_pathogenic

Table 4.7 Extracted AlphaMissense scores and associated predicted variant impact for identified *ARHGAP19* missense variants (in yellow shade are variants located within the GAP domain).

4.3.5 GAP activity assay

The GAP activity assay was performed collaboratively by Nathalie Lamarche-Vane laboratory, Canada.

4.3.5.1 In-vitro GAP assay predicts loss of GAP activity in patients' mutations

ARHGAP19 has been previously reported for its RhoGAP activity towards a prototypical Rho GTPase, RhoA (David *et al*., 2014). Interestingly, several mutations to *ARHGAP19* found in patients are clustering around the region encoding its GAP domain. To investigate the GAP activities in these *ARHGAP19* mutations, wild type ARHGAP19-GAP or the mutated proteins were expressed as GST fusion proteins in *E. coli* for *in vitro* GAP assays. Specific GAP activity toward RhoA was measured as the rate of inorganic phosphate released by GTPase-mediated GTP hydrolysis. RhoA alone showed little intrinsic GTPase activity, while the addition of wild type ARHGAP19 significantly accelerated the rate of RhoA-mediated GTP hydrolysis. Two ARHGAP19 missense mutations, p.Gly140Asp, and p.Gln151Lys, abrogated the GAP activity of ARHGAP19, decreasing the GTPase hydrolysis rate to the basal level. Another ARHGAP19 mutation, p.His196Glnfs*9, which led to a truncated GAP domain of ARHGAP19, completely abolished the GAP activity as evidenced by severely impaired phosphate release (fig.4.12). Altogether, these results suggest that mutations in the GAP domain of ARHGAP19 serve as dominant loss-of-function alleles, abolishing its GAP activity.

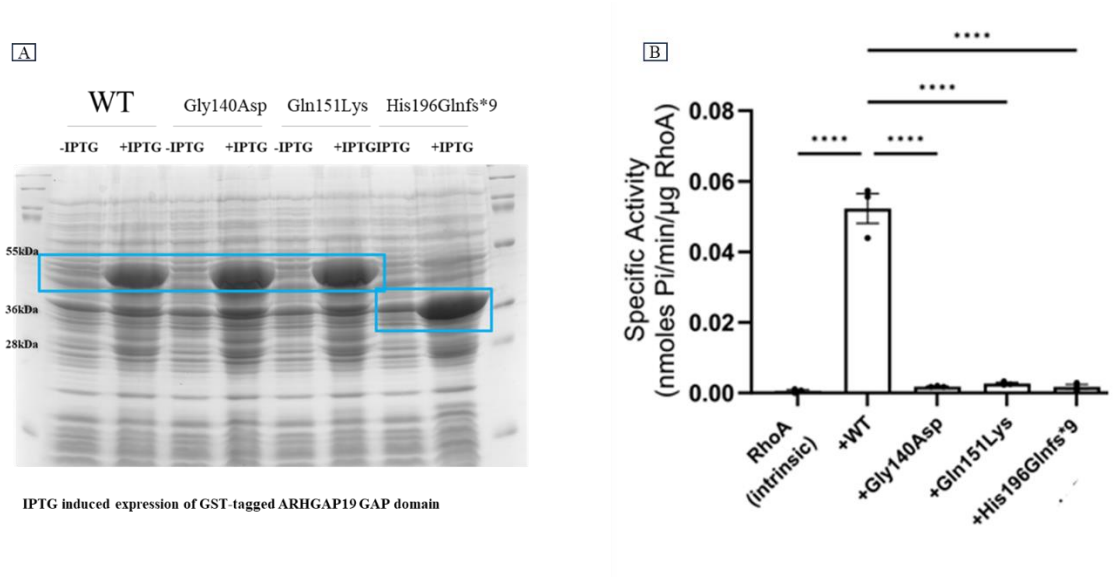


Figure 4.12 GAP activity assay. Two missense (Gly140Asp and Gln151Lys) and one frameshift (His196Glnfs*9) variants, intrinsic RhoA and WT protein tested A) Western blotting expression of GST-tagged ARHGAP19 GAP domains for WT and the mutants B) Two missense (Gly140Asp and Gln151Lys) and one frameshift (His196Glnfs*9) variants cause abolished GAP activity. One-way ANOVA. Significance is shown as **** and indicates $p<0.0001$.

4.3.6 In vitro fibroblasts assays

To further delineate the underlying functional correlation between patients' clinical features and *ARHGAP19* mutations, we derived isolated and cultured fibroblasts via a skin biopsy performed in from the patients of family 5 (p.Gly140Asp), family 6 (p.Leu68Pro) and family 10 (Asn29Asp) paired with age/ gender matched healthy controls (wild type). Since the in vitro GAP assay indicated that *ARHGAP19* c.419G>A (p.Gly140Asp) has a defect on GAP activity towards RhoA, we hypothesized that loss-of-function of *ARHGAP19* will lead to over activation of RhoA which in turn will cause over activation of ROCK and the kinase's downstream effectors. Together with our collaborators or in house, we were able to investigate the consequences of *ARHGAP19* mutations on modulation of cell proliferation and migration, vimentin and beta-actin expression and an attempt was made to investigate the expression of ARHGAP19 protein.

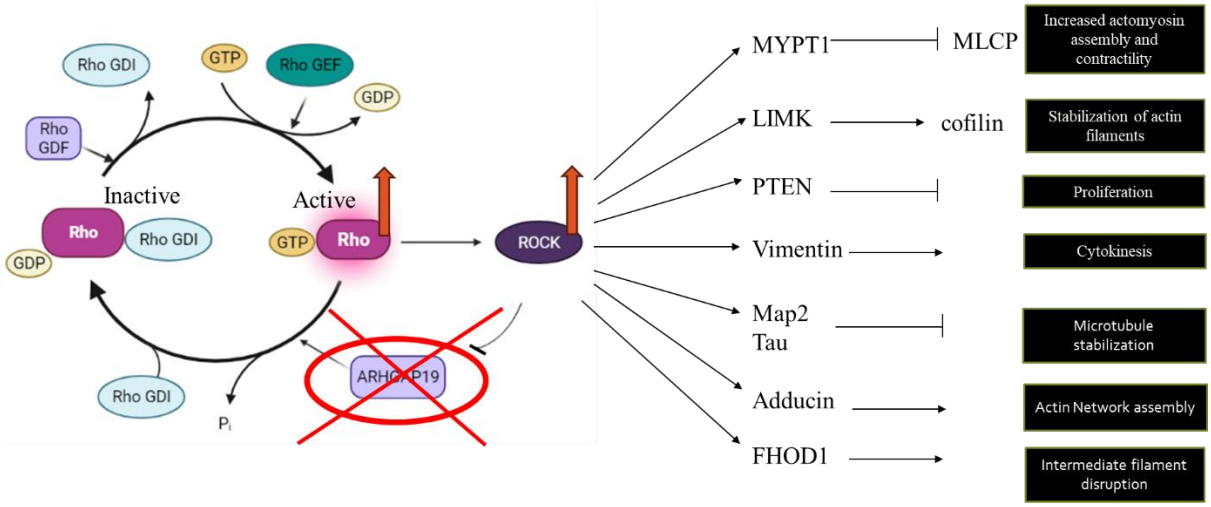


Figure 4.13 Prediction of consequences of loss of function of ARHGAP19 on Rho/ROCK pathway. Rho cycles between inactive GDP bound state and active GTP bound state. The pathway is tightly regulated by regulatory proteins – GAPs GEFs and GDIs. ARHGAP19 is a GAP protein. When in its active state Rho activates ROCK which can further activate many downstream effectors. The effectors in this figure are speculative for the purpose of the functional investigations of ARHGAP19. We hypothesise that loss-of-function of ARHGAP19 will lead to overactivity of Rho and further overactivity of ROCK and the downstream effectors.

4.3.6.1 Cell proliferation

Cell proliferation and viability was assessed in the available fibroblasts using MTT assay (He *et al*., 2023). MTT assay is a coluorimetric assay used to measure cellular activity of the cells and is based on the ability of metabolically active cells to use oxidoreductase enzymes which can reduce yellow tetrazolium salt (MTT) into purple formazan.

ARHGAP19 mutants and wild type fibroblasts had been cultured in 96-well plate for over a 3-day period and we observed no significant differences in the MTT assay between patient and control fibroblasts suggesting the variants have no consequence on cultured fibroblasts proliferation.

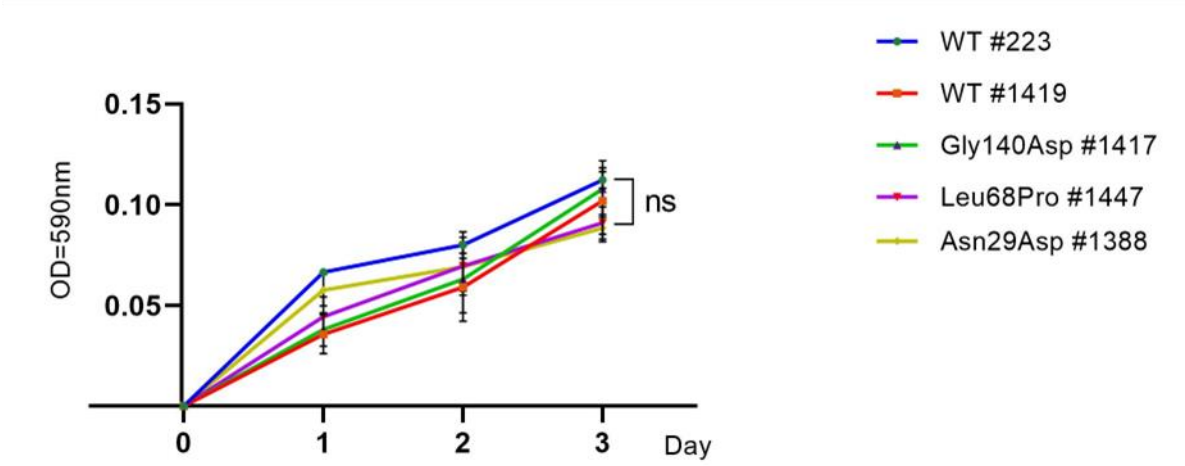


Figure 4.15. MTT proliferation assay. No significant changes are seen in proliferation of patient derived fibroblasts compared to controls during 3 day (x axis) treatment. Values are shown as absorbance at 590nm (y axis). One-way ANOVA. ns=not significant.

4.3.6.2 Cell motility assays

We next investigated the motility of these mutants by performing the Boyden Chamber migration assay and wound healing assay (fig.4.16 A and B). Wound healing scratch assay is a relatively simple assay to measure basic cell migration in response to a crude wound (Cory, 2011).

The Boyden chamber assay is a migration assay based on a medium filled chamber with two compartments separated by porous membrane. The compartment underneath the membrane is filled with chemotactic agents towards which the cells will migrate through the porous membrane. The number of cells migrated is then calculated (Chen, 2005).

Interestingly, fibroblasts from P5 (F5-II:2) Gly140Asp and P11 (F10-II:1) Asn29Asp had a significant reduction in cell migration compared to the wild type, in contrast, there was no altered cell motility in P6 (F6-II:7) Leu68Pro fibroblasts (fig.4.16 C and D), suggesting a role of *ARHGAP19* Gly140 and Asn29 residues in cell migration.

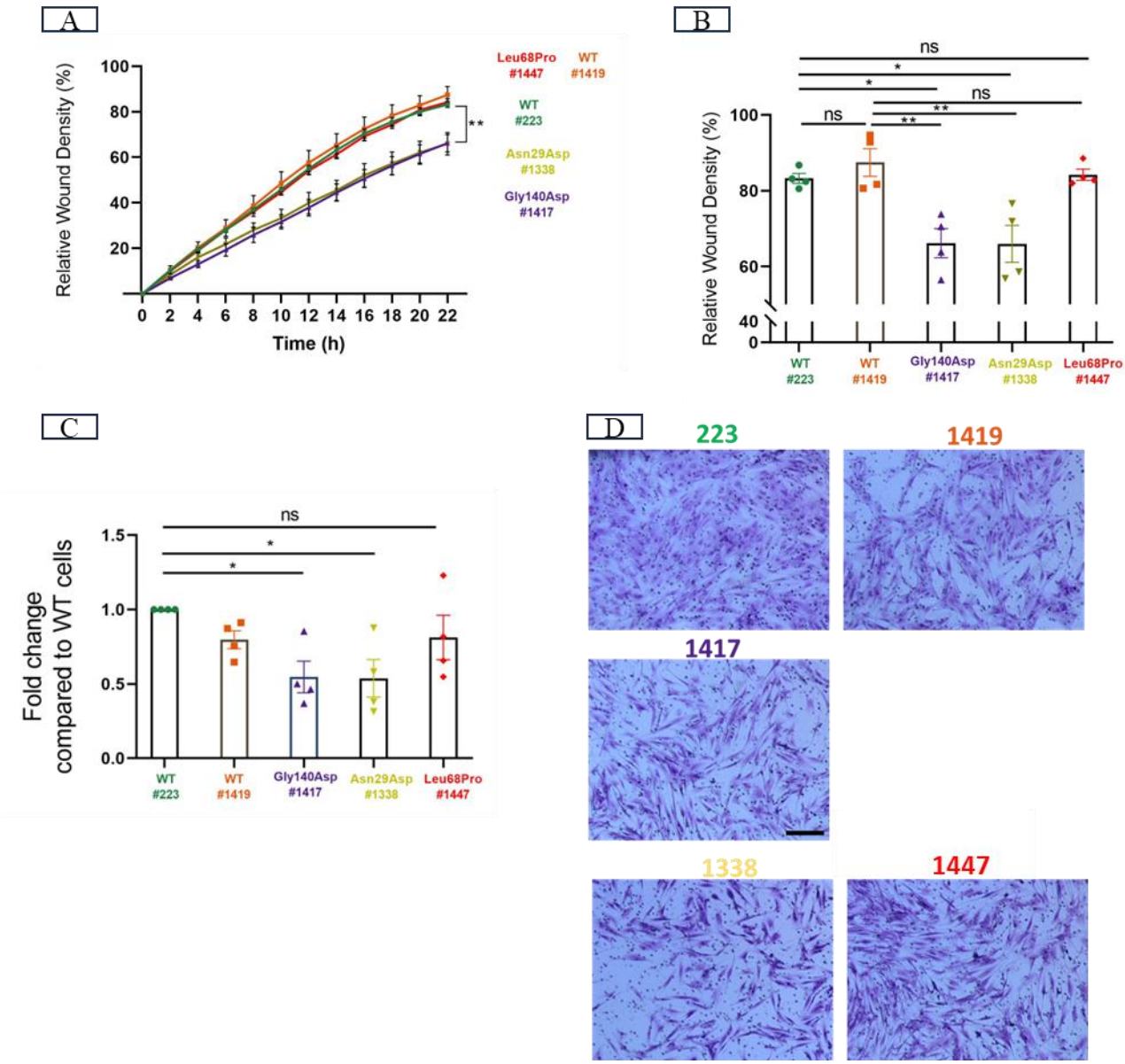


Figure 4.16 Migration assays. A and B) Wound healing scratch assay and C and D) Boyden chamber migration assay showed significant reduction in motility of patient derived fibroblasts of patients with Gly140Asp and Asn29Asp variants as opposed to controls. No significant differences were observed for patient with variant Leu68Pro indicating that this variant may a different mechanism. Scale bar is 50 µm. Data are presented as means ± SEM; * = $p<0.05$, ** = $p<0.01$, ns = not significant, one-way ANOVA

4.3.6.3 No cytokinesis disruption and changes in vimentin expression

Vimentin is an intermediate filament that is highly expressed in Schwann cells and neurons. Intermediate filaments have important functions in nerve development and regeneration and are components of cytoskeleton. Vimentin has a role in cytokinesis, the cell division into two daughter cells (Triolo *et al.*, 2012). We stained fixed patient and control fibroblasts with anti-vimentin antibody and interrogated any potential differences in the expression using Zeiss

confocal microscope. We observed no differences in the cell lines suggesting there is no immediate changes to vimentin expression in fibroblast caused by *ARHGAP19* mutations.

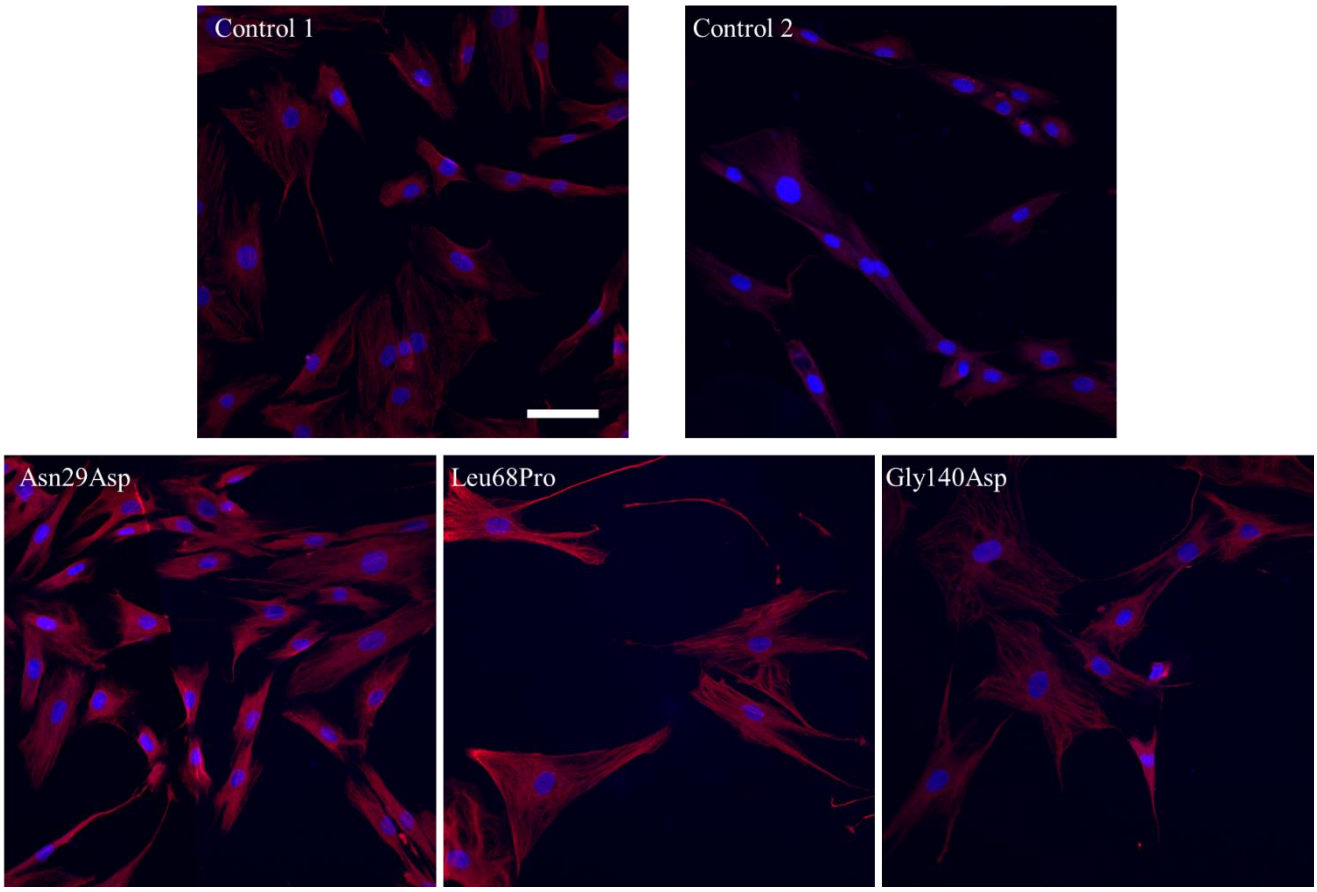


Figure 4.17. Vimentin staining of patient derived (bottom) and control (top) fibroblasts. Three fibroblasts cell lines from patients were interrogated for vimentin expression and structural changes with variants including two lying outside of the GAP domain (Asn29Asp and Leu68Pro) and one lying inside the GAP domain (Gly140Asp). No visual differences were observed in vimentin expression of the patients compared to age matched controls. Vimentin is in red, Hoechst nuclear staining (blue). Scale bar is 100 µm.

4.3.6.4 No actin cytoskeleton disruption

ROCK and downstream effectors have important roles in actin organisation and stabilisation of actin filaments. We interrogated any possible disruption of actin cytoskeleton by staining the available fibroblasts with beta actin which is a non-muscle cytoskeletal actin controlling cell growth and migration.

We saw no visible differences of beta actin expression between the patients and controls.

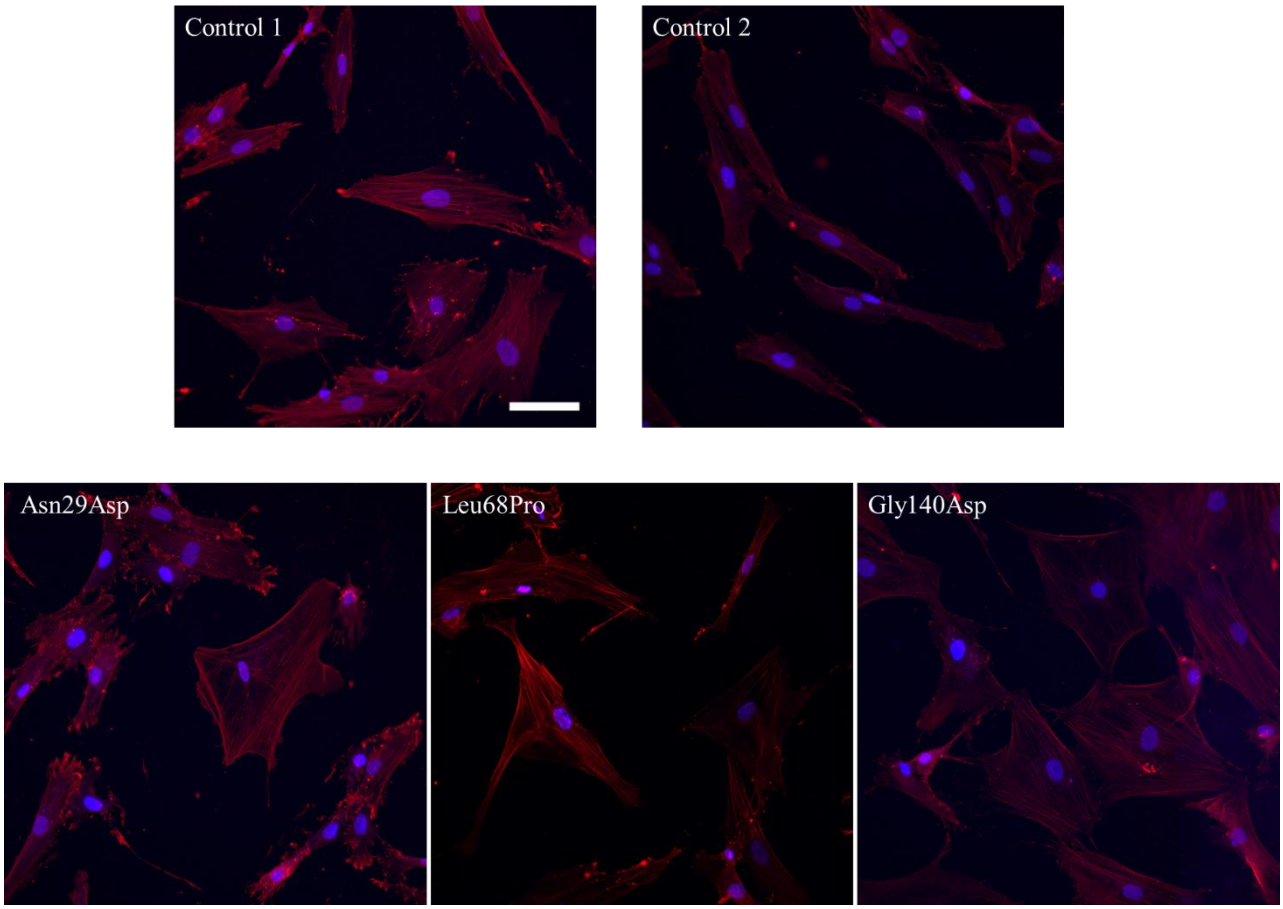


Figure 4.18. Beta actin staining of patient derived and control fibroblasts. Three fibroblasts cell lines from patients were interrogated for beta actin expression and structural changes with variants including two lying outside of the GAP domain (Asn29Asp and Leu68Pro) and one lying inside the GAP domain (Gly140Asp). No visual differences were observed in beta actin expression of the patients compared to age matched controls. Beta actin is in red, Hoechst nuclear staining (blue). Scale bar is 100 µm.

4.3.6.5 ARHGAP19 protein expression was inconclusive in fibroblasts

ARHGAP19 is not a well-studied protein and to date its expression profile has been shown in T-lymphocytes in vitro. It is also known that this is a nuclear protein but it shuttles between various cell locations (David *et al*., 2014).

We tested two commercial antibodies (Genetex and Invitrogen) for immunocytochemistry staining in our available fibroblasts. Both the antibodies failed to produce acceptable quality results and indeed when trialled for reactivity on western blotting, these antibodies did not show any results suggesting that either there is too low expression of ARHGAP19 in fibroblasts, the antibodies are not optimised for use in fibroblasts, or the reaction conditions used need to be optimised. However, a different anti-ARHGAP19 antibody was commercially available (Santa Cruz), and it was used in further experiments.

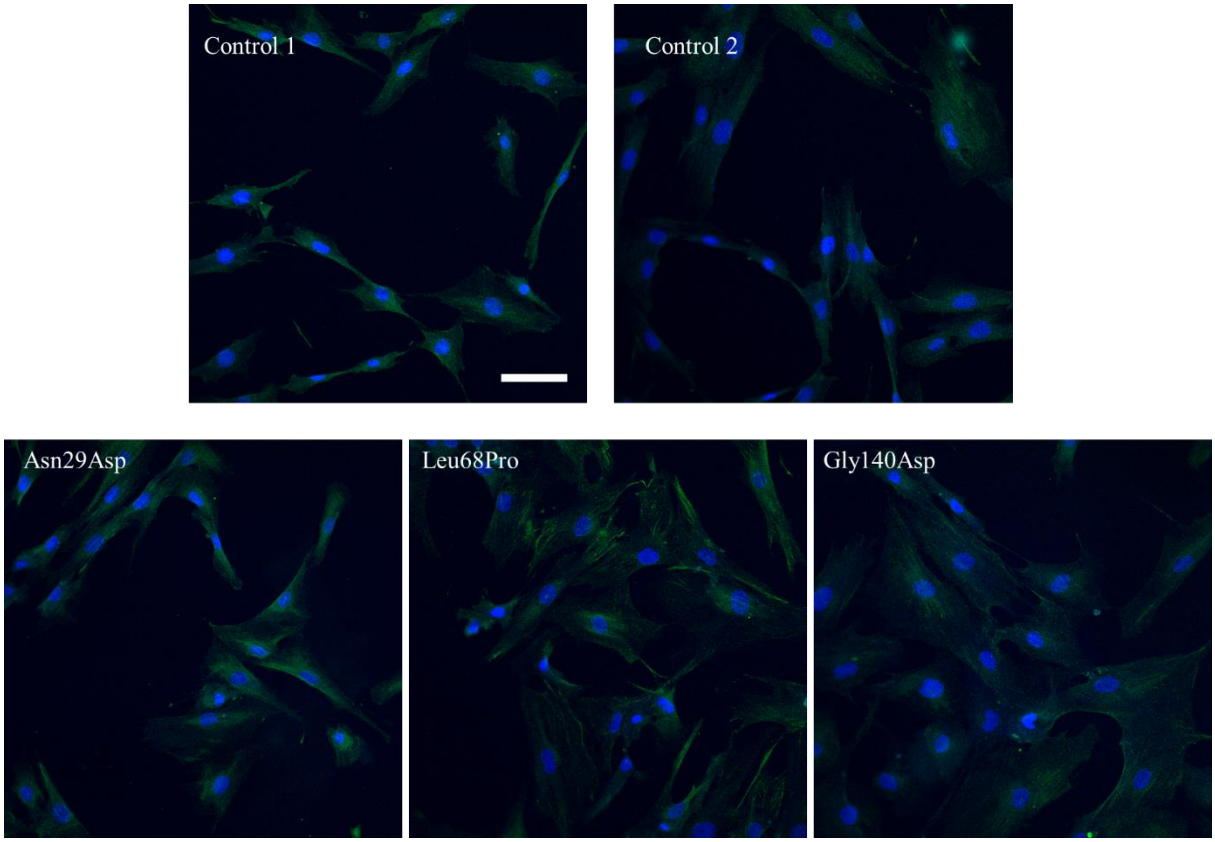


Figure 4.19. ARHGAP19 staining of patient derived and control fibroblasts. Three fibroblasts cell lines from patients were interrogated for ARHGAP19 expression and structural changes with variants including two lying outside of the GAP domain (Asn29Asp and Leu68Pro) and one lying inside the GAP domain (Gly140Asp). No visual differences were observed in ARHGAP19 expression of the patients compared to age matched controls. ARHGAP19 is in green, Hoechst nuclear staining (blue). Scale bar is 100 µm.

4.3.8 Quantitative PCR

I performed qPCR on fibroblasts of the patients and controls; and on fibroblast derived iPSC motor neurons of the patients and corresponding controls to assess the levels of *ARHGAP19* RNA.

There were no significant differences at RNA level between the patients and controls in both fibroblasts and the motor neurons. This is perhaps not surprising as the available cell lines are derived from patients carrying missense mutations. Missense mutations cause difference of amino acids in the translated protein compared to the wild-type protein; however, they are not predicted to cause changes in mRNA levels.

4.3.9 Western blotting

I performed Western blotting on fibroblasts of the patients and controls; and on fibroblast derived iPSC motor neurons of the patients and corresponding controls to assess the levels of ARHGAP19 protein.

There were no significant differences in ARHGAP19 protein expression in fibroblasts of patients versus controls. Interestingly, significant loss of protein expression was observed in patient derived motor neurons suggesting that the loss-of-function of ARHGAP19 is more robust in these cells.

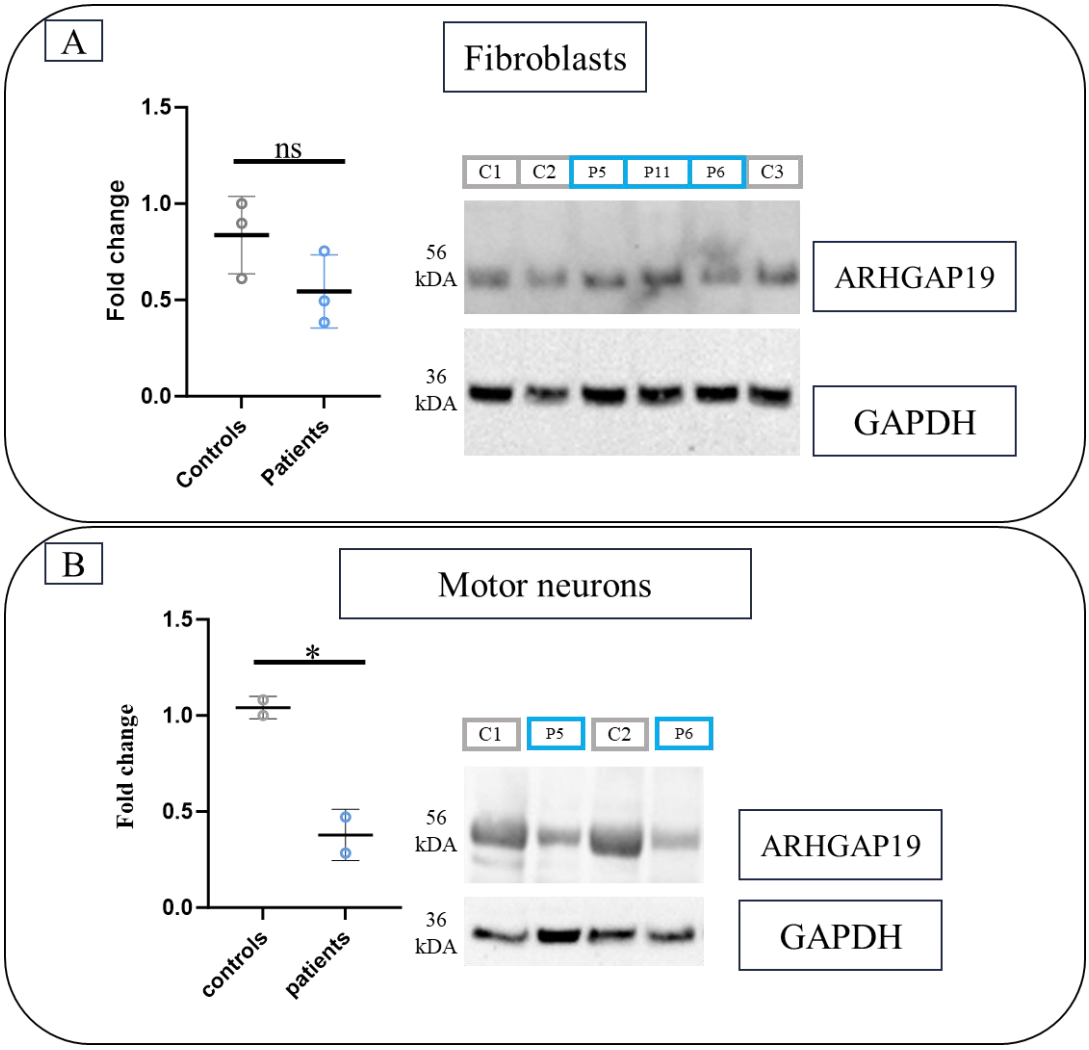


Figure 4.20 ARHGAP19 Western blotting. A) ARHGAP19 protein expression levels in patient derived fibroblasts are unchanged compared to healthy controls. B) ARHGAP19 protein expression levels in iPSC motor neurons show significant decrease in expression between the patients and healthy controls. Data are presented as means ± SEM; * = $p<0.05$; ns = non-significant. Controls are highlighted in grey and the patients in blue. GAPDH was used as a housekeeping protein.

4.3.10 In vivo *Drosophila melanogaster* model

4.3.10.1 The ARHGAP19 ortholog RhoGAP54D promotes movement in Drosophila

To explore the consequences of *ARHGAP19* loss of function in vivo, we first utilised the fruit fly, *Drosophila melanogaster*. I am covering the methods and tools I used for this work as well

as the results in the next chapter 5 titled "*Drosophila melanogaster as model organism for discerning function of RFC1 and ARHGAP19*".

4.3.11 In vivo *Danio rerio* model

4.3.11.1 In situ hybridization detection of arhgap19 mRNA local expression

To analyse the endogenous expression and subcellular localization of arhgap19 during embryonic development, we conducted whole-mount in situ hybridization (WISH) assays at three different embryonic stages, utilizing digoxigenin-labelled antisense RNA probes specific for arhgap19. WISH analyses revealed a ubiquitous expression pattern of arhgap19 across multiple brain regions, notably in the forebrain and hindbrain compartments. This expression was predominantly enriched within neural tissues at the 48hpf. Specifically, heightened arhgap19 expression was observed in anatomically defined regions such as the cerebrum, thalamus, tuberculum, and tegmentum (fig 4.21). Intriguingly, a temporal downregulation of arhgap19 expression was evident as development progressed; by 5dpf, the expression levels had substantially diminished.

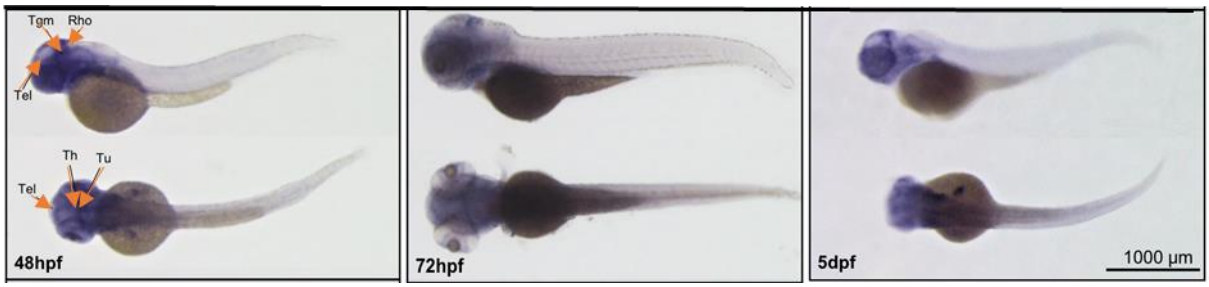


Figure 4.21 arhgap19 protein localisation in *Danio rerio* during embryonic development. At 48 hpf, WISH signal of arhag19 is localized in the forebrain and hindbrain regions; scale bar is 1000 µm. Rho: Rhombencephalon (hindbrain); Tel: Telencephalon; Th: Thalamus; Tu: Tuberculum; Tgm: Tegmentum.

4.3.11.2 Phenotype and locomotor behaviour of arhgap19 mutant zebrafish using CRISPR-Cas9 and morpholino technologies

To investigate the functional role of *Arhgap19* in neuronal and motor development, we generated an F0 biallelic knockout mutant model utilizing CRISPR/Cas9 system. Genomic deletions were introduced at three selected loci within the zebrafish *arhgap19* gene—exons 2, 4, and 5—informed by considerations of targeting efficiency and potential off-target effects. We also induced *arhgap19* knockdown model by injecting two morpholinos, specifically targeting the E1 splice site and the AUG translation start site.

Of the initial batch of 62 eggs, 32 (51.6%) were successfully fertilized. Embryos were monitored on a daily basis, with dead ones discarded. Those that survived were sacrificed at 5 dpf for genotyping and subsequent analysis.

To evaluate the changes in behaviour in the *arhgap19* mutant model, we collected 12 zebrafish larvae at the 5 dpf and analysed their motor activity. Metrics such as the total duration of movement, aggregate distance traversed, and mean velocity were measured. Behavioural assays demonstrated that *arhgap19* knockout induced conspicuous motor deficits. Larvae in the CRISPR-induced mutants (CRISPants) and morpholino-injected mutants (MO) groups exhibited decreased motor activity, alongside idiosyncratic and involuntary movements. In contrast, larvae from the Uninjected Control (UIC) group exhibited normal locomotor behaviour, exploring the well's periphery.

Statistical analysis revealed significant discrepancy in motor parameters among the groups. Specifically, larvae from the CRISPants and MOs groups were significantly different in the total travel distance (fig. 4.22 A and B); ($p < 0.001$). Likewise, the mean velocity also showed significant differences between mutant and control groups (Fig. 4.22 C; $p < 0.001$) Larvae from the CRISPants and MOs groups swam approximately three times more slowly compared to those from the UIC group.

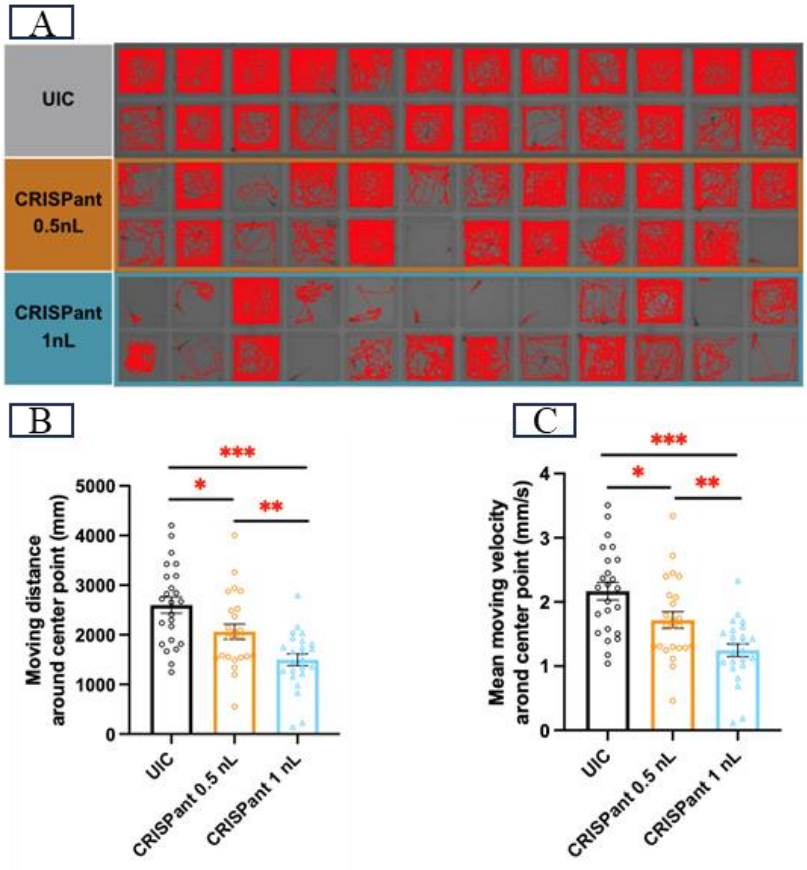


Figure 4.22 *Danio rerio* behavioural assays. (A) reveals the swimming trajectories of each larva. Quantification of total travel distance (B) and travel velocity (C) of UIC and arhgap19 mutant zebrafish larvae for 30 mins. Data are presented as means ± SEM; * = p<0.05, ** = p<0.01, *** = p<0.001, one-way ANOVA. UIC= Uninjected Control

4.3.11.3 Analysing muscle birefringence of zebrafish mutants

To investigate the impact of *arhgap19* knockout on muscular architecture, we quantitatively assessed the birefringence intensity of zebrafish skeletal muscle at the 5-dpf, employing a polarizing light stereomicroscope for imaging. Statistical analysis of the birefringence levels revealed no significant difference between the control group and the three knockout groups (p=0.06). These findings strongly suggest that the skeletal muscle integrity remains largely intact in the absence of *arhgap19*. Consequently, the motor deficits observed in the behavioural analyses are more likely attributed to impairments in motor neuron function rather than muscular deficiencies.

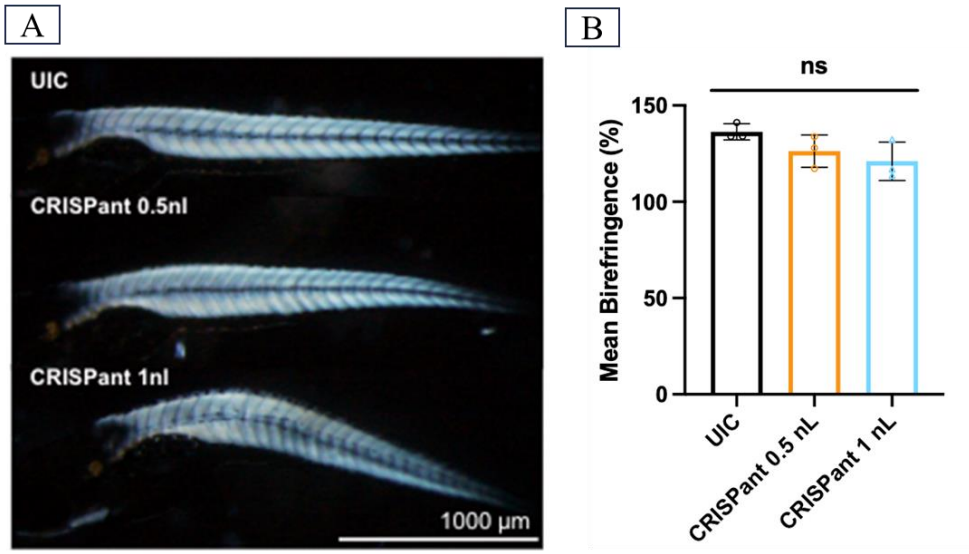


Figure 4.23 *arhgap19* Danio rerio muscle integrity. A) representative image of 1 larva from each treatment group. Scale bar is 1000µm. B) No difference in average birefringence for all zebrafish larvae. Data are presented as means ± SEM; ns = not significant. UIC= Uninjected Control

4.3.11.4 arhgap19 knockout/knockdown cause motor neuron malformations

To investigate the effects of *arhgap19* knockout on spinal motor neurons (spMNs), we employed immunostaining techniques complemented by confocal microscopy for visualization. Notably, a more robust axonal bundle was observed in both CRISPants (fig 4.24A). Quantitative evaluations were conducted on both axonal length and branching complexity. In arhgap19 CRISPant larvae, the branching density of the Caudal Primary (CaP) Motorneurons was higher than that of the control group ($p<0.0001$). The larvae that injected with the two morpholinos also displayed a similar elevation in branching numbers ($p<0.001$) (fig 4.24B). Moreover, axonal length was markedly affected in *arhgap19* CRISPants. Statistical analysis revealed a notable reduction in the average length of CaP and Middle Primary (MiP)

motorneurons in 1nL injected CRISPants , measuring 577.2 µm and 391.2 µm, respectively, in contrast to the control values of 1012.5 µm and 639.8 µm (fig 4.24C).

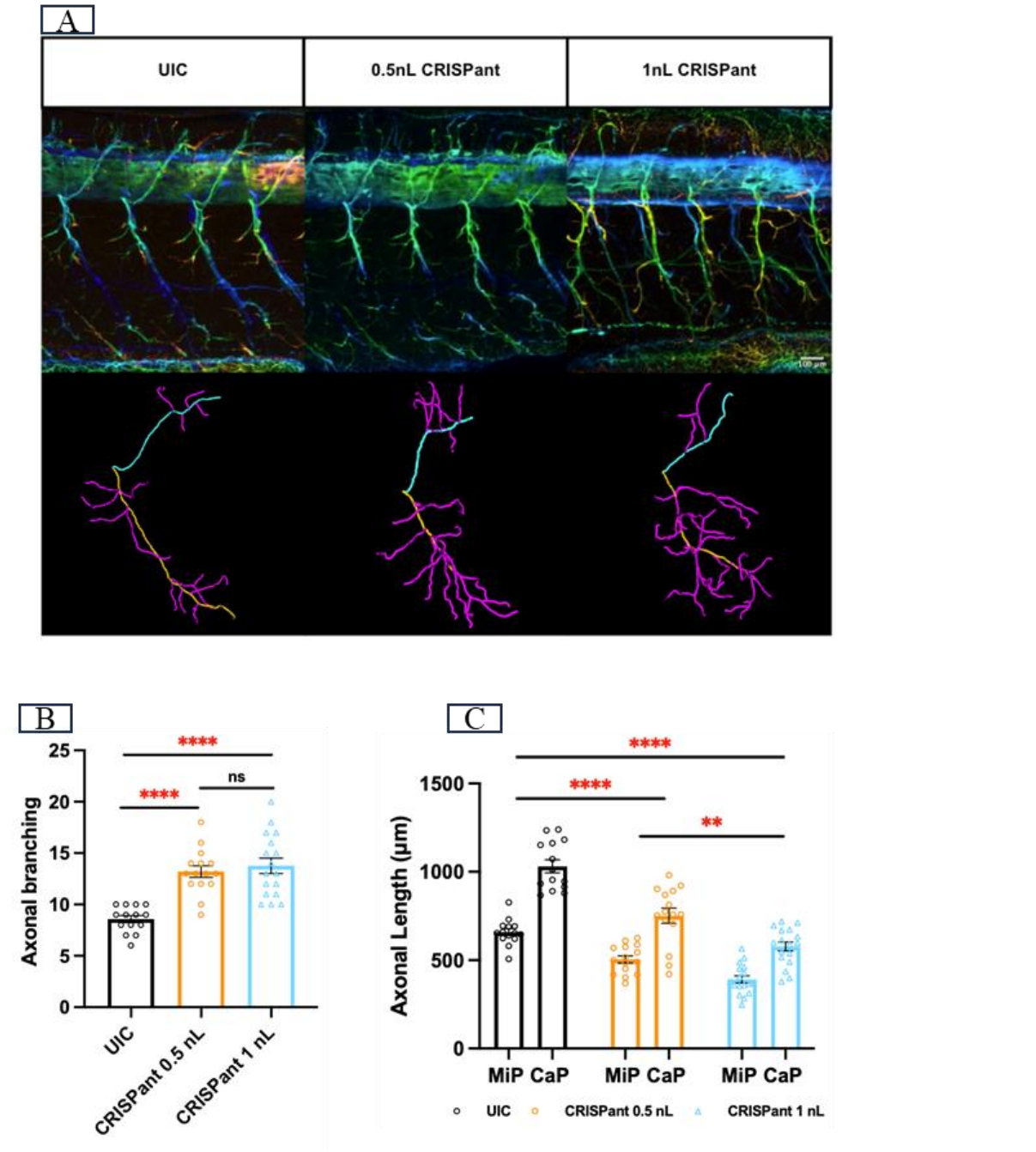


Figure 4.24 Spinal motor neurons morphogenesis defects in *arhgap19* mutant zebrafish larvae. A) Confocal imaging analysis and three-dimensional reconstruction of spinal motor neurons in UIC and *arhgap19* mutant groups at 5-dpf; scale bar is 100 µm. B) Axonal branching number of Cap axons in UIC and *arhgap19* mutant zebrafish larvae. C) Average axonal length of Cap (yellow) and Mip (blue) axons in UIC and *arhgap19* mutant zebrafish larvae. One-way ANOVA. Data are presented as means ± SEM; ** = $p<0.01$, **** = $p<0.0001$, ns = not significant. UIC= Uninjected Control

4.4 Discussion

4.4.1 *ARHGAP19* as novel motor-neuropathy-causing gene

I co-led and contributed to a large international study of patients with inherited neuropathy, and we show for the first time that biallelic *ARHGAP19* mutations are a novel cause of CMT.

We identified 25 individuals from 20 families harbouring missense and nonsense variants laying both in the functional GTPase activating protein (GAP) domain as well as outside this structural domain of ARHGAP19. The patients had a motor predominant length dependant axonal neuropathy with a young age of onset in the first two decades of life apart from pt 10 whose symptoms started in their 30es. There seemed to be no correlation between the variant position and the patient phenotype at the first examination mean AAO of variants in GAP versus non-GAP domain 9.0 vs 11.6 years ($p = 0.32$, $n = 15$ and 8 respectively) and mean ulnar motor conduction velocity 41.0 vs 44.8 m/s ($p = 0.62$ $n = 5$ and 5). However, longitudinal studies would help better delineate the disease progression and the correlation between phenotype and genotype.

Interestingly, we found four recurrent mutations including recessive possible founder variants affecting amino acid residues p.Leu68Pro which lies outside the GAP domain, and p.Gln151Lys, p.His196Glnfs*9 and p.Leu228His that are located within the Rho-GAP domain. Haplotype analysis using genetic data from our cohort as well as from control databases suggested that p.His196Glnfs*9 is unlikely to be from a recent common ancestor and possibly suggests that two independent *ARHGAP19* mutational events within Arabian Middle Eastern populations. Moreover, p.Gln151Lys and p.Leu68Pro variants are founder effect variants, probably originating in Turkey given the ethnicity of patients. Variant p.Gln151Lys which affects a highly conserved residue within the GAP domain, could have a significant structural or functional role.

Together with other substituted residues (p.Gly140Asp, p.Leu141Trp, p.Asn239Lys) and frameshift variants (p.Asn160Metfs*21 and p.Pro188Argfs*5) the recurrent mutations in GAP domain are predicted to disrupt the domain's structure and subsequently its function as a GTPase-activating protein in a variety of cellular processes (David *et al.*, 2014; Amin *et al.*, 2016) including cell cycle control adhesion and migration as well as actin cytoskeleton organisation.

Indeed, the pathogenicity of the chosen *ARHGAP19* mutations located within the GAP domain (p.Gly140Asp, p.Gln151Lys and p.His196Glnfs*9), is further supported by the in vitro GAP

activity assays which show that these mutations cause complete loss of GTPase activity towards RhoA. This data is in line with previously reported variants in the GAP domain of *MYO9B* gene (Cipriani *et al*., 2023), where the variants abrogate Rho-GAP activity. Interestingly, biallelic mutations in this gene also cause a neuropathy, a CMT2 with optic atrophy. Taken together with *PLEKHG5* gene (Chen *et al*., 2021) which has GEF activity towards RhoA and biallelic mutations in which cause CMT phenotype, this highlights the importance of Rho/ROCK pathway in neuropathy and opens up the question for potential therapeutic treatments.

4.4.2 Loss-of-function animal models recapitulate the patients' phenotype

To further support the loss-of-function disease mechanism we established animal models of the disease using *Drosophila melanogaster* and *Danio rerio*. Caution must be taken when exploring CMT disease with *Drosophila* which lacks myelin and will not prove a good model for demyelination, however, our cohort is mostly composed of axonal CMT cases, therefore we saw potential benefits of using the model. The *Drosophila melanogaster* ortholog of *ARHGAP19*, *RhoGAP54D*, is sparsely expressed in variety of different cells in the fly and three global in vivo loss-of-function *Drosophila* models were established. These flies revealed reduced overall and peak movement and startle response to light-off as opposed to corresponding controls supporting the genetic link between mutations in the human ortholog *ARHGAP19* and disrupted movement.

Interestingly, ability to establish global knockouts of *RhoGAP54D* suggests that the protein is not essential for survival in the fly or other proteins might have a compensatory role in those flies. The *Drosophila melanogaster* models are further shown in the next chapter (Chapter 5 *Drosophila melanogaster* as a model organism for deciphering loss of function of *ARHGAP19* and *RFC1*)

Danio rerio loss-of-function models strengthen the data from our fly models. The fish knockout, apart from recapitulating patient movement phenotype showed more pronounced axonal bundles and a significantly increased number of axonal branches in all experimental mutant groups ($p>0.0001$). Importantly, during neuronal development, RhoA activation has been shown to facilitate axonal outgrowth and branching while inhibiting dendritic growth and branching and the *D. rerio* model robustly corroborated these findings.

The data from *Drosophila melanogaster* and *Danio rerio* highlighted that loss-of-function of the ARHGAP19 protein orthologs in the two animal models caused common motor phenotypes

recapitulating those of human patients which further strengthened the human *ARHGAP19* genotype – phenotype linkage. During the course of this thesis, we were unable to pinpoint in which cell types ARHGAP19 orthologs are functioning, however, we saw expression of the protein orthologs likely in subperineural and perineural glia in adult fly brain and in forebrain and hindbrain at embryonic stages of the fish development. This could pose a question whether the expression pattern may suggest potential glial involvement, and this could be a topic for future research. Another important question for future investigations would be whether the increased axonal branching in the fish directly causes its movement phenotype or there are other compensatory mechanisms at play.

4.4.3 Exploration of pathomechanisms of *ARHGAP19* mutations

Since ARHGAP19 has GAP activity towards RhoA, when functioning it stimulates the intrinsic low GTPase activity of RhoA thereby negatively regulating the RhoA/ROCK pathway. Mutations causing loss of GAP activity of ARHGAP19 may therefore cause over activity of RhoA. In this thesis, I hypothesised that over activity of RhoA will sequentially cause over-activity of ROCK and, with help from colleagues and collaborators, I attempted to explore, state and investigate some of the possible downstream effectors of the pathway and the consequences they may have on the cultured fibroblasts of patients with missense mutations (p.Asn29Asp, p.Leu68Pro and p.Gly140Asp) and where possible, fibroblasts-derived motor neurons (p.Leu68Pro and p.Gly140Asp).

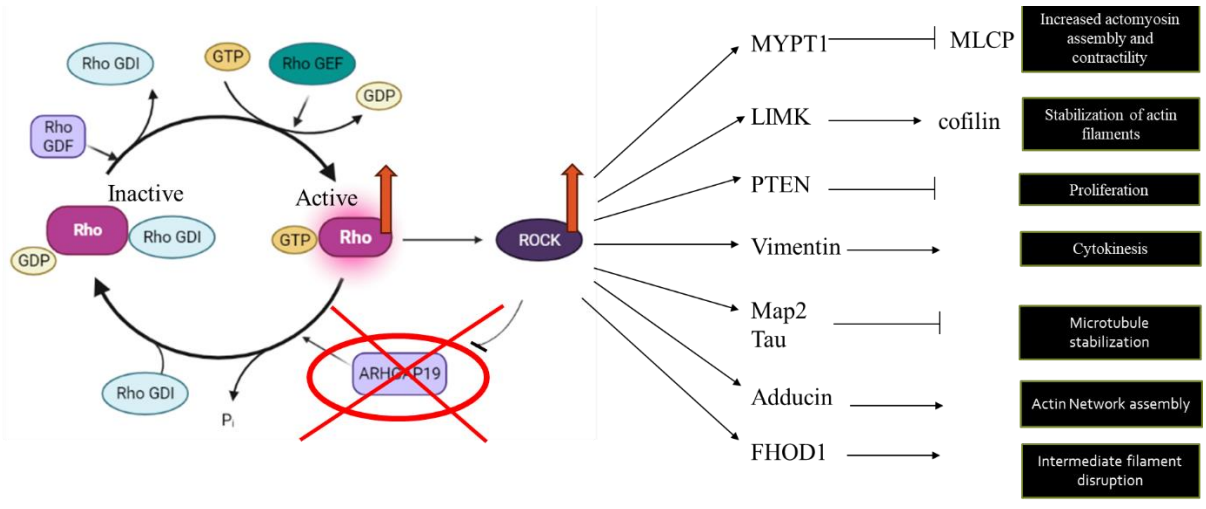


Figure 4.25 Prediction of consequences of loss of function of ARHGAP19 on Rho/ROCK pathway. Rho cycles between inactive GDP bound state and active GTP bound state. The pathway is tightly regulated by regulatory proteins – GAPs GEFs and GDIs. ARHGAP19 is a GAP protein. When in its active state Rho activates ROCK which can further activate many downstream effectors. The effectors in this figure are speculative for the purpose of the functional investigations of ARHGAP19. We hypothesise that loss-of-function of ARHGAP19 will lead to overactivity of Rho and further overactivity of ROCK and the downstream effectors. vim

Despite the theoretical simplicity of an on/off switch model of Rho, the pathway has a complicated mechanism. The high number of GAP and GEF proteins – 66 and 80 respectively (DeGeer and Lamarche-Vane, 2013), who outnumber Rho proteins, coupled with unclear specificity of the protein's function, means that understanding their signalling activities as well as role in disease, remain challenging. We were unable to account for any compensatory mechanisms that could be present and indeed we did not explore any other interactors of ARHGAP19 and RhoA/ROCK pathway outside of the schematic in fig.4.25. Indeed, the downstream effectors shown in fig.4.25 are purely speculative and are taken from various literature not necessarily connected to neurological disease. With this in mind, I, with help of collaborators, explored some potential targets such as actin cytoskeleton, cell proliferation and motility as well as vimentin expression in patient fibroblasts and where possible, fibroblasts derived motor neurons. Due to a large number of potential effectors (many not included in the figure) I decided not to perform qPCR to investigate the expression of any of these. Towards the end of these thesis, we submitted RNA from three patients and three control fibroblasts to perform RNA sequencing in hope for more inclusive and better streamlined process for finding up- or down-regulated effectors. This could shed light on which genes get upregulated/downregulated thus impacting pathways implicated in important processes or which drug targets should be explored in future research. Below, I discuss the effectors that we were able to investigate with the existing current resources.

4.4.3.1 Cell proliferation

The ability of patient fibroblast to proliferate was assessed with MTT assay. Cell proliferation is an important process in which cells grow and divide and this process varies dependent upon the cell type. The possibility of over-active ROCK to inhibit proliferation was therefore assessed and no differences were seen between the available patient fibroblasts and controls suggesting no direct consequences on cell proliferation in fibroblasts.

4.4.3.2 Cell motility

Cell motility was assessed with two motility assays – wound healing scratch assay which assesses basic cell migration and Boyden chamber migration assay which relies on chemotactic agents to attract the cells.

The in vitro data from patient derived fibroblasts (Asn29Asp and Gly140Asp) showed significant decrease in relative wound healing density in wound healing scratch assay ($p<0.05$) as opposed to WT controls; and the same mutations had significant consequences on fibroblast migration ($p<0.05$) in Boyden chamber assay. Interestingly, scratch assay relies on cell

polarisation of microtubules for the ability to close the wound. This process is highly dependent on actin and microtubule dynamics which could point towards the RhoA/ROCK pathway's actin or microtubule roles. Boyden chamber assay relies on chemotaxis, a process which is highly relevant in wiring the nervous system where axonal growth cones must be precisely guided to their targets. Future studies of *ARHGAP19* could investigate whether the disease-causing mutations have an effect on axonal guidance in-vivo.

Another mutation tested with migration and wound healing assays, Leu68Pro, showed no change as opposed to the non-disease controls. Interestingly, the two patients with this mutation in our cohort were first suspected to have an autoimmune cause of neuropathy and were put on intravenous immunoglobulin for treatment of presumed Chronic inflammatory demyelinating polyneuropathy (CIDP with no improvement). These patients also have an upper limb involvement. Leu68 lies outside the GAP domain, but it is predicted pathogenic using various in silico methods. Whether there might be different mechanisms of pathogenicity for this particular *ARHGAP19* mutation remains to be elucidated.

4.4.3.3 Vimentin expression

Vimentin is an intermediate filament that is highly expressed in Schwann cells and neurons. Intermediate filaments have important functions in nerve development and regeneration and are components of cytoskeleton. Vimentin has a role in cytokinesis, the cell division into two daughter cells (Triolo *et al*., 2012). We observed no visual differences in expression of vimentin between the patient and control fibroblasts suggesting it is unlikely to be upregulated in our ARHGAP19 patients.

However, it is interesting to note that vimentin was found to negatively regulate myelination, with a loss of vimentin in neurons causing hypermyelination of axons (Triolo *et al*., 2012). It is therefore plausible to speculate that overexpression of vimentin could cause hypomyelination. Many of the ARHGAP19 patients present with demyelination, and whilst we saw no differences of vimentin expression in patient fibroblasts, whether this hypomyelination could be caused by overexpression of vimentin in either neurons or Schwann cells of the patients is currently unknown. We have available fibroblast-derived iPSC motor neurons for patients 5 and 10 and whilst interrogating their myelination would require a co-culture with Schwann cells which is not possible during the course of this thesis.

4.4.3.4 Actin cytoskeleton

Rho/ROCK pathway regulates various aspects of cytoskeleton formation and control. We interrogated whether the missense *ARHGAP19* mutations may cause disruption of actin cytoskeleton by investigating beta actin expression.

We saw no visible differences of beta actin expression between the patients and controls. However, as beta actin controls cell migration and significant differences were observed in ARHGAP19 patient fibroblasts' ability to migrate versus the wild type fibroblasts in wound healing scratch assay, it could be possible that the magnification used to visualise the beta actin was not sufficient or the fibroblast migration in the assay was not controlled by beta actin. It could therefore be another point for future investigations and the patient iPSC motor neurons could be interrogated in the first instance.

4.4.3.5 mRNA and protein expression levels

Quantitative PCR did not show any significant differences between expression levels of ARHGAP19 in fibroblasts of patients with biallelic mutations and unaffected controls and the same was true in two fibroblasts-derived patient iPSC motor neurons and corresponding controls. This is perhaps unsurprising as the available cell lines were derived from patients with missense mutations which are not predicted to cause changes of expression at mRNA level.

Assessing the protein expression level of fibroblasts did not yield significant expression differences between the two groups. However, and interestingly, iPSC motor neurons derived from ARHGAP19 patients with missense mutations showed significantly ($p<0.05$) less expression of the protein as opposed to healthy controls. This could suggest that the loss-of-function of *ARHGAP19* caused by pathogenic biallelic missense mutations is more robust in motor neurons and indeed the highest expression of ARHGAP19 according to public data set (GTEx project) is expected in tibial nerve. Together these data suggest that patient movement phenotype may be a direct result of under expression of ARHGAP19 in motor neurons.

We were unable to tests any of the frameshift or truncating variants due to unavailability of patient samples, however, these are likely to cause GAP activity abrogation by formation of shorter protein product. Indeed, in patients with frameshift stop variants in the GAP domain (p.Asn160Metfs*21, p.Pro188Argfs*5 and p.His196Glnfs*9) these could result in truncated proteins with incomplete GAP domain. To speculate further, these truncating mutations, together with other two truncating variants lying outside of the GAP domain, in the C-terminus of ARHGAP19 (p.Arg407* and p.Gln415*), may cause mis-localisation of those truncated

protein products. This is because the C-terminal end of ARHGAP19 is known to have three phosphorylation sites – Threonine 404 and 476 that can be phosphorylated by CDK1; and Serine 422 which can be phosphorylated by ROCK. Sequential phosphorylation at these residues by the kinases has been shown to regulate the localisation of ARHGAP19 in lymphocytes during mitosis in vivo (David *et al*., 2014).

ARHGAP19 isoforms are another important consideration. In this thesis, I described a patient cohort with mutations that map to the canonical ENST00000358531.9. However, notably, one patient PT8 (F8-II:2) had a p. Met? variant in the non-canonical *ARHGAP19* transcript ENST000000371027.1. This transcript is predicted to form a 9 amino acid shorter isoform of the canonical 494aa ARHGAP19 protein and affect the N-terminus of the isoform after the Methionine start codon. Mutation in p. Met? is a start loss mutation where no protein will be formed. PT 8 develops neuropathy and has a phenotype compatible with the entire cohort, which suggest that this isoform has a function in human nervous system. Further work is needed to describe the exact function of ARHGAP19 isoforms; however, their existence suggests that ARHGAP19 may regulate various cellular processes depending on their localisation or stages of cell differentiation and the loss of function in non-canonical ARHGAP19 isoforms may be sufficient to trigger the development of neuropathies.

4.4.4 ARHGAP19 causes motor-neuropathy through loss-of-function pathomechanism

We identified 25 individuals from 20 families harbouring homozygous missense and nonsense variants in *ARHGAP19*. In this thesis I described evidence for a loss-of-function pathomechanism of ARHGAP19 causing motor predominant neuropathy. Firstly, the genetic variant association with the patient phenotype points towards the loss-of-function mechanism and indeed Sanger sequencing allowed for variant segregation within the families in our cohort and showed that only variants in homozygous state cause neuropathy, whereas parents and other family members with heterozygous variants remain unaffected.

With this in mind, we developed two loss-of-function animals – I established the fruit fly model and performed the experiments, and in collaboration, zebrafish model. Both the animal models recapitulated patient movement phenotype and further strengthened the human *ARHGAP19* genotype – phenotype linkage. Importantly, the *Drosophila melanogaster* data suggests that ARHGAP19 fly ortholog is either not essential to the animal survival or compensatory mechanisms with other GAP (or GEF) proteins may be involved. It is still unknown when ARHGAP19 critical period takes place or if the protein may be essential throughout

development and adult life - the zebrafish data was strikingly collected in the embryonic stages which suggests developmental stages may be impacted, whereas the fly data was collected in adult animals, however, ARHGAP19 orthologue was knocked down throughout the entire fly lifespan. In addition, axonal length and branching of motor neurons were assessed in the zebrafish knock-out model and interestingly higher number of branching and decreased length of the axons were observed. This might point to loss of polarity in growth, however, the morphology of human iPSC MNs and fly neuromuscular junctions have not yet been explored to conclude whether this is a common mechanism.

In this work, the pathogenicity of *ARHGAP19* variants p.Gly140Asp, p.Gln151Lys and p.His196Gln*fs**9 are supported by *in vitro* GAP activity assays which show that variants within the GAP domain cause complete GAP loss. ARHGAP19 stimulates the intrinsic low GTPase activity of RhoA thereby negatively regulating the RhoA/ROCK pathway. Variants causing GAP loss may therefore cause overactivity of RhoA with further consequences in downstream effectors such as ROCK activation which has important functions in actin organisation, cell migration, and axon outgrowth and guidance.

Importantly, this is in line with *in vitro* data from patient-derived fibroblasts harbouring p.Asn29Asp and p. Gly140Asp variants, that show significant decrease in cell migration. Together, these results reveal that GAP defective *ARHGAP19* variants lead to altered RhoA activity thereby altering cell migration and cytoskeletal dynamics. Interestingly, protein expression analysis by Western blotting revealed significant reduction in ARHGAP19 in iPSC motor neurons of patients compared to controls. Taken together, these findings are consistent with the patients' phenotype and functional assays, explaining that the muscular and motor defects observed in patients as well as animal models could be due to the dysregulation of the ARHGAP19-related signalling cascade.

Taken together, the data provides evidence for loss of function mechanism in ARHGAP19 disease which may be more robust in motor neurons, however, further studies are needed to elucidate the tissue specificity and critical period of ARHGAP19 function.

4.4.5 Future perspective

It is important to note that the findings described in this thesis do not fully elucidate the mechanism of axonal damage caused by ARHGAP19 deficiency nor do they imply the tissue specificity or expression of ARHGAP19 in human or animal models.

Indeed, the patient phenotype and the models described in this thesis suggest an intrinsic axonal deficit, however, the bulk expression from publicly available datasets (GTEX) may imply Schwann cell expression. Therefore, the future work to describe ARHGAP19-mediated neuropathy will include further characterisation of the gene expression, this can be achieved by interrogation of single cell RNA sequencing datasets in combination with ARHGAP19 staining in rodents. A preliminary search of ARHGAP19 in Human Brain Cell Atlas v1.0, an RNA sequencing data of single cell nuclei from three brain donors, showed no ARHGAP19 expression in this dataset, however, it was limited to brain tissues. Interestingly, using CZ CELLxGENE Discover, an open access, large, standardised, and curated collection of single-cell datasets, ARHGAP19 expression was observed in myelinating Schwann cells as well as peripheral nervous system neuron. At the time of this analysis, tibial nerve cell populations were not available in the dataset therefore it remains to be elucidated whether ARHGAP19 neuropathy is due to a primary axonal degeneration or whether loss-of-function of ARHGAP19 in Schwann cells leads to secondary axonal degeneration. However, it is important to note that Drosophila melanogaster lacks myelinating cells and myelin, and loss of function of its ARHGAP19 ortholog leads to significant locomotion phenotype.

In the future, other datasets will be interrogated to expand this analysis. In conjunction, rodent tissues can be dissected and stained with anti-ARHGAP19 antibodies to better describe the expression in central and peripheral nervous system further helping with questions such as whether the protein is expressed in axons and/or Schwann cells or other cells, and expression in the development could be assessed at embryonic stages. The animal models developed in this thesis recapitulated human patient phenotype. In the zebrafish model, we showed striking axonal branching and shortening of axons in the context of ARHGAP19 knockdown. Future studies will aim to examine neuromuscular junctions of *Drosophila melanogaster* larvae as well as patient derived iPSC MNs to compare the morphology of axons in these models which will further help to determine whether loss of polarity in developing axons could play role in this disease. In addition, human derived cells can be used for interrogation of morphology resulting from overactivity of RhoA, such as stress fibres.

Due to *Drosophila melanogaster* model advantages such as short lifespan and large number of animals, this model can be utilised further to answer important questions on ARHGAP19 function. For example, it remains unclear whether ARHGAP19 fly ortholog, RhoGAP54D, acts in the mature nervous system or during development to regulate movement. To investigate this, Temporal and Regional Gene Expression Targeting system can be used to switch off expression

of RhoGAP54D globally at different developmental time-points, and to interrogate the consequences on adult movement. In addition, gross locomotor activity of ARHGAP19 KO larvae can be compared to controls to interrogate a developmental phenotype. This data together with the larvae neuromuscular junction staining will be compared to the zebrafish model where embryonic development was implicated in the KO pathomechanism. Finally, to test for involvement of specific cell-types in the pathology, expression of RhoGAP54D can be knocked down using tissue-specific drivers such as nsyb-Gal4 (neurons), repo (glia), alrm-Gal4 (astrocytes), and mz0709-Gal4 (ensheathing glia). Developmental drivers such as wor-Gal4 (neuroblasts) can also be investigated.

4.4.6 Limitations

This study was possible due to invaluable collaborations from around the globe and willingness of patients to participate. Many of the patients are based in low to middle-income countries where access to healthcare is limited therefore it can be difficult to have repeated clinical assessments important for longitudinal studies or for consenting to additional sampling such as skin biopsies. Much remains to be elucidated in *ARHGAP19* pathogenesis and if biopsies from patients with truncating variants were available, the derived cell lines could shed more light on function and localisation of ARHGAP19.

4.5 Conclusion

In conclusion, this study expanded the genetic heterogeneity of CMT and added a novel gene to the rapidly expanding list of CMT causative genes. It provided evidence that biallelic *ARHGAP19* mutations cause a loss-of-function neurological disease and two animal models, *Drosophila melanogaster* and *Danio rerio*, recapitulate the motor deficit movement observed in patients further supporting the human genotype – phenotype observations. ARHGAP19 is a GTPase activating protein taking part in Rho/ROCK pathway and in vitro studies described here show that the patient mutations cause loss of GAP activity of the protein which may cause overactive RhoA signalling. Importantly, overactive RhoA signalling in neurons, be it due to gene mutations or imbalance between signalling molecules, has been reported in Charcot-Marie-Tooth (Chen *et al*., 2021; Beijer, *et al*., 2022; Cipriani *et al*., 2023). It is increasingly emerging that this is an important pathway not only in neuronal health but also disease and may show a potential for therapeutic treatments in CMT related disorders.

CHAPTER 5. *Drosophila melanogaster* as a model organism for deciphering loss of function of *ARHGAP19* and *RFC1*

5.1 Introduction

5.1.1 *Drosophila melanogaster* as a model organism

The fruit fly, *Drosophila melanogaster*, has been used as a model system for well over a hundred years (Morgan, 1910) and many genetic techniques have been developed in this organism to answer a wide range of biological questions, a large number of which are applicable to study of neurogenetic disorders.

One of the key advantages of *Drosophila* as a model is its quick life cycle (fig.5.1) of about 10 days at 25°C, which allows for rapid generation of large numbers of flies. This also allows for studying the fly at different developmental stages, be it larva or the adult fly. Moreover, flies are highly genetically tractable compared to other animal models (Arias, 2008) and males and females have distinct features that allow for their differentiation. The females can lay eggs shortly after mating that can occur as soon as 12 hours after eclosion (emergence of adult from pupal case), and one female can produce as many as 100 eggs per day. In addition, they are easy to house, do not require much space and they have a low cost of maintenance. They also have a relatively low level of genetic redundancy. Most importantly, homologues of over 70% of proteins that are involved in human disease, can be found the *Drosophila* genome, which means it can be used as a model system for human disease (Reiter and Bier 2002, Rubin *et al.*, 2002). If a human gene has its homolog in the fly, much can be learned about its possible function, interactions, and biochemistry.

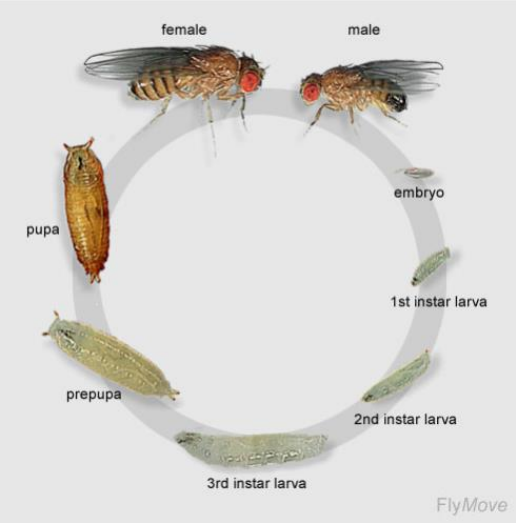


Figure 5.1 Drosophila life cycle in which a fertilised egg becomes an embryo, goes through 3 larval stages, and finally becomes a pupa from which an adult fly emerges. Figure from http://flymove.uni-muenster.de/Genetics/Flies/LifeCycle/LifeCyclePict/life_cycle.jpg

5.1.2 *Drosophila* genetics

Drosophila melanogaster has two copies of sex chromosomes and two copies of three autosomes resulting in 4 diploid chromosomes. The majority of the *Drosophila* genome of fly is well sequenced and annotated (Adams *et al*., 2000).

Large collections of mutations and other resources are available through Stock Centres such as Bloomington (https://bdsc.indiana.edu/), Vienna (https://stockcenter.vdrc.at/control/main) and others, which can be utilised for specific targeted genetic interventions.

Furthermore, due to their simple genome architecture and rapid generation time, *Drosophila* containing complex combinations of mutant alleles and/or transgenic insertions can readily be generated.

5.1.3 Tools in *Drosophila* research

Below, I give a brief background to some tools used in *Drosophila* genetic research and that were used in this thesis.

5.1.3.1 Balancer chromosomes

An important tool in fly work is a balancer chromosome. These chromosomes carry multiple inversions and translocations which suppress recombination on that specific chromosome allowing for stable maintenance of deleterious mutations.

Balancer chromosomes carry dominant mutations causing a visible phenotype and which are generally lethal when homozygous (Stocker and Gallant 2008). When performing mating

schemes, this allows for the precise tracking of alleles carried in trans with a balancer chromosome, thus eliminating the need for genotyping using molecular biology methods. For example, crossing a fly with a balancer called CyO - which contains the dominant loss of function allele *curly* which results in curly wings - of genotype *actin-Gal4*/CyO with a wild-type fly will produce half of the offspring with curly wing and no *actin-Gal4* and half of the offspring with straight wing which will be heterozygous for *actin-Gal4*. Balancers are incredibly useful, and they have been developed for the 3 chromosomes that can undergo meiotic recombination. It is however important to note that flies with balancer chromosomes are usually less fit than wild type flies and they do not produce as many offsprings. Examples of some balance chromosomes and their phenotypes shown in fig.5.2 A and B.

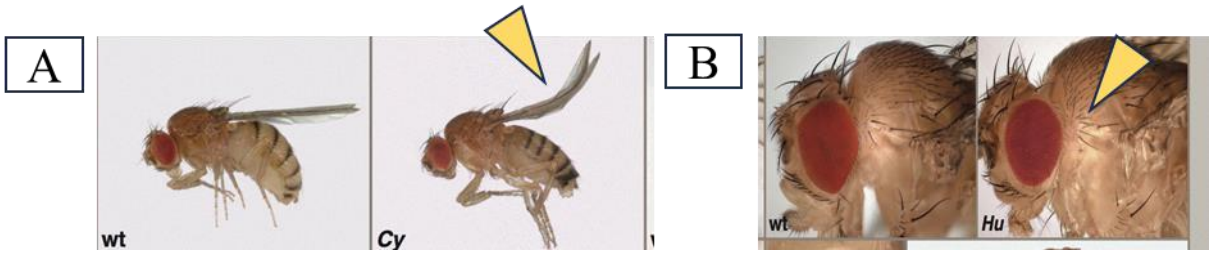


Figure 5.2 Example of balancer chromosomes and dominant markers. A) A second chromosome balancer, Curly of Oyster (CyO). Yellow arrow points to the curly wings. B) A third chromosome balancer, TM6b with dominant marker Humoral (Hu). Yellow arrow points to the more dense hairs.

5.1.3.2 GAL4 – UAS system

GAL4 – UAS system is a bipartite tool allowing for expression of a particular gene or short harpin (shRNA) in tissue a specific manner (figure 5.3) which will lead to RNA interference (RNAi) mediated knockdown. It relies on two separate transgenic fly stocks, a driver line with yeast GAL4 transcription factor expressed in a cell or tissue specific pattern, and a responder line with a gene of interest under upstream activating sequence (UAS) control (the binding site for GAL4) (Brand and Perrimon, 1993; McGuire *et al.*, 2004). When these transgenic flies are kept in separate stocks the UAS construct is silent. Only after crossing the driver and responder lines, will the progeny express the transgene/mRNA downstream of UAS, which is trackable owing to the presence or absence of balancer chromosomes depending on the known genetic background of the stock flies. Depending on the responder element in the responder line, downstream of UAS, various outcomes can be achieved such as knockdown of specific gene using shRNA lines, fluorescent labelling of cellular or sub-cellular populations using fluorescent proteins such as green fluorescent protein (GFP), or over-expression of wild-type/mutant forms of a protein of interest.

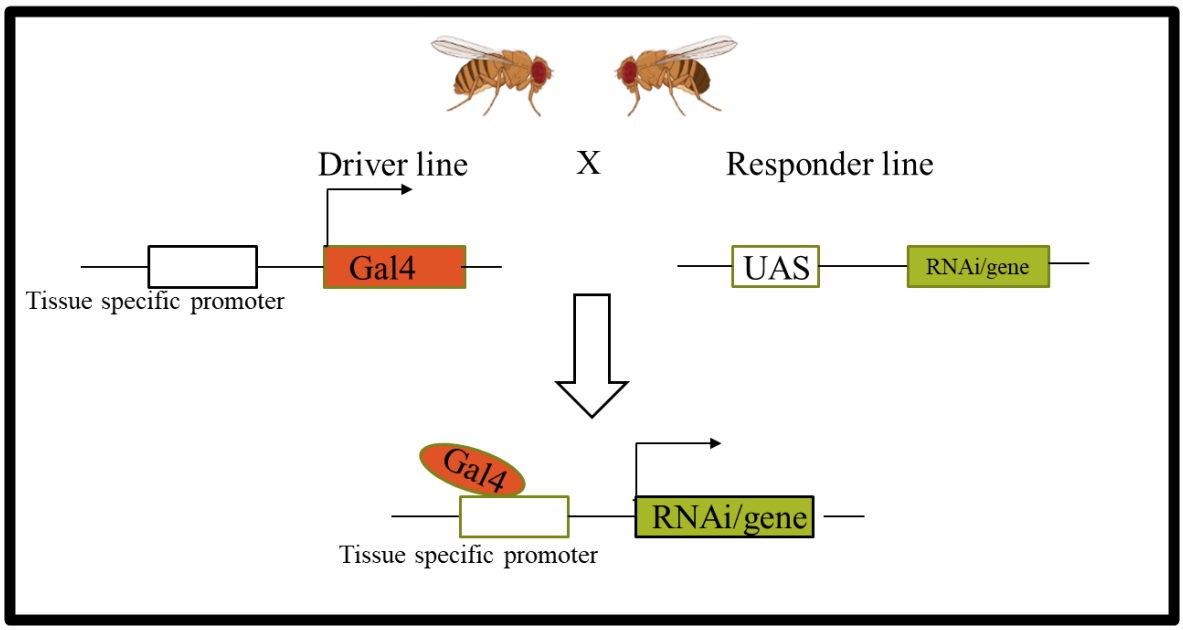


Figure 5.3 GAL4-UAS system schematic. A driver line fly expresses Gal4 in cell or tissue specific pattern and is crossed with a responder line fly who has an RNAi construct under UAS. The progeny of those flies expresses the shRNA or a gene of interest

A multitude of tissue specific promoters can be found, which include global drivers such as *actin*-Gal4, drivers specific for neurons (*nsyb*-Gal4 or *elav*-Gal4), or glia (*repo*-Gal4), and many others.

5.1.3.3 CRIMIC lines

CRISPR Mediated Integration Cassette (CRIMIC) lines are CRISPR/Cas9 targeted insertions into fly genome that can be replaced with desired DNA sequence due to flanking recombination sites. Importantly, these CRIMIC insertions contain T2A-Gal4 (fig.5.4) as part of the insertion cassette which results in gene specific reporter activity (Lee *et al.*, 2018).

Hence in addition to generating loss-of-function alleles by interrupting transcription, Crimic lines can also be leveraged for expressing a fluorescent protein under UAS control to visualise the gene expression patterns.

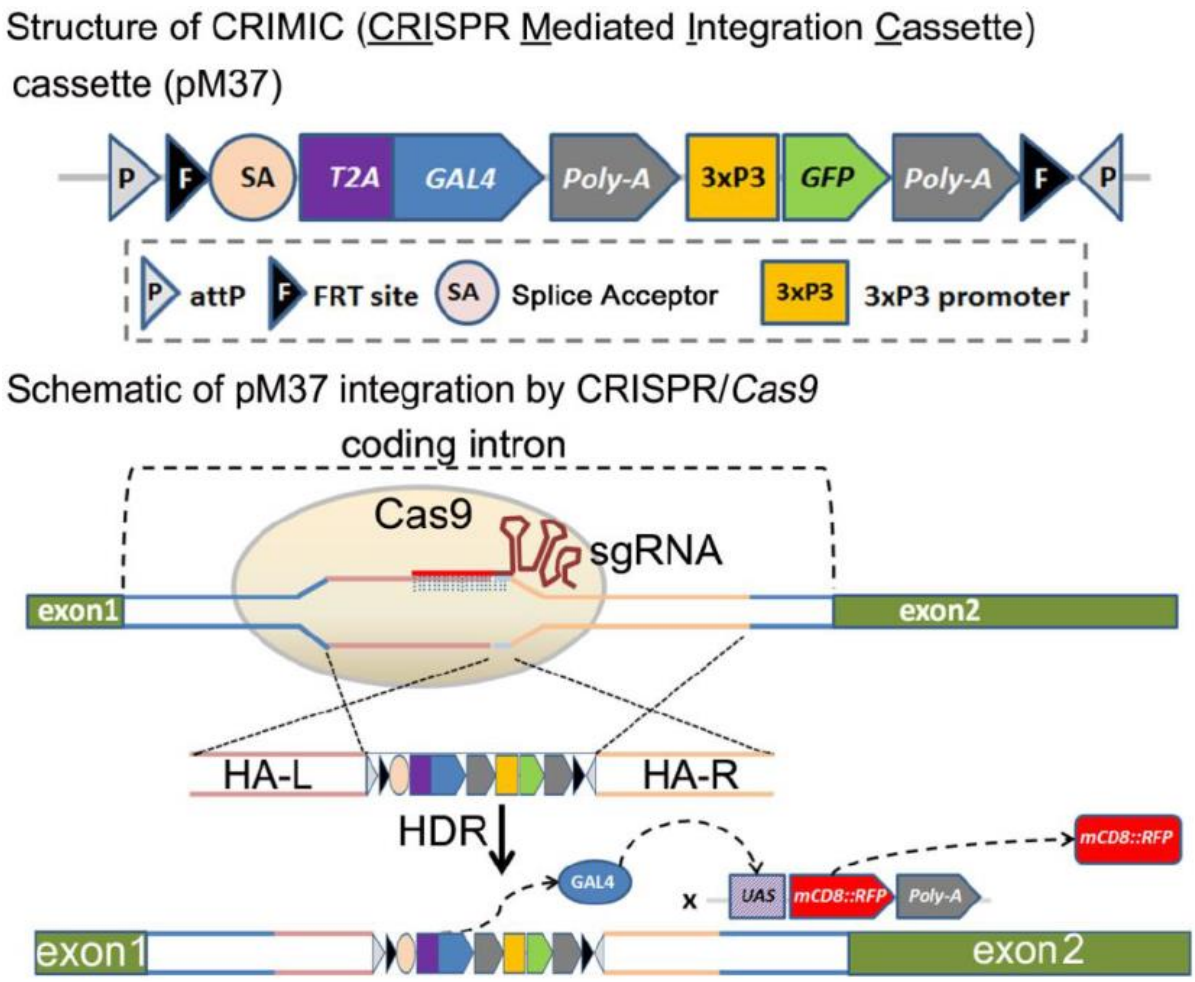


Figure 5.4 Schematic of CRISPR-mediated integration cassette with T2A-Gal4s forming CRIMIC lines. Adapted from Lee *et al*., 2018.

5.1.3.4 RNA interference

RNA interference (RNAi) is a mechanism which leads to degradation of messenger RNA (mRNA) transcripts by short, homologous, short harpin RNA (shRNA) (Elbashir *et al*., 2001). This can be used to reduce expression of genes by introducing exogenous shRNA into cells. Following expression of shRNAs by use of UAS-GAL4 system, the shRNA forms a harpin structure in the cell. The double stranded RNA harpin is cleaved by the Dicer-2 endonuclease to produce 20-25bp dsRNA fragments containing guide and passenger strands. The guide strand is further associated with RNAi silencing complex (RISC) protein complex to bind to native mRNA. Additional endonucleases such as Ago-2 bind to the complex and cleave the native mRNA (Carmichael *et al*., 2004).

Leveraging GAL4-UAS system, RNAi can be induced in a tissue specific manner in the fly. RNAi transgenic flies are available for multitude of known genes and can be ordered from stock centres mentioned in previous section. RNAi, however, is not a perfect tool, since the

available transgenic lines are often not validated and may only be partially efficacious. Moreover, RNAi can potentially have an off-target effect by knocking down expression of other genes.

Even though these disadvantages exist, the use of RNAi is one of the most common methods for inducing gene knockdown because it is relatively cheap and not time consuming. This is a good first method in establishing a gene function and can be validated by other models or molecular biology methods.

5.1.3.5 deGradFP

deGradFP is a system that allows for degradation of GFP-tagged proteins (tagged at endogenous locus) and therefore for inducing loss-of-function. This is another important method of genetic knockdowns further down the central dogma – at the protein level. It is not limited by potential off target effects unlike RNAi, and it is a universal method because it relies on ubiquitin-proteasome pathway that is evolutionary conserved in eukaryotes (Caussinus *et al*., 2011)

In *Drosophila*, deGradFP relies on series of genetic crosses to create a fly with the GFP-tagged protein and UAS-Nslmb-vhhGFP4 which induces degradation signal by an anti-GFP single chain nanobody that fuses to ubiquitin/proteosome signalling machinery (fig.5.4). The knockdown can be achieved in tissue specific manner by use of Gal4 drivers. However, one of the disadvantages of using the model is that it requires homozygous GFP alleles, it may be difficult to create the fly lines and some flies with tagged genes may not be viable.

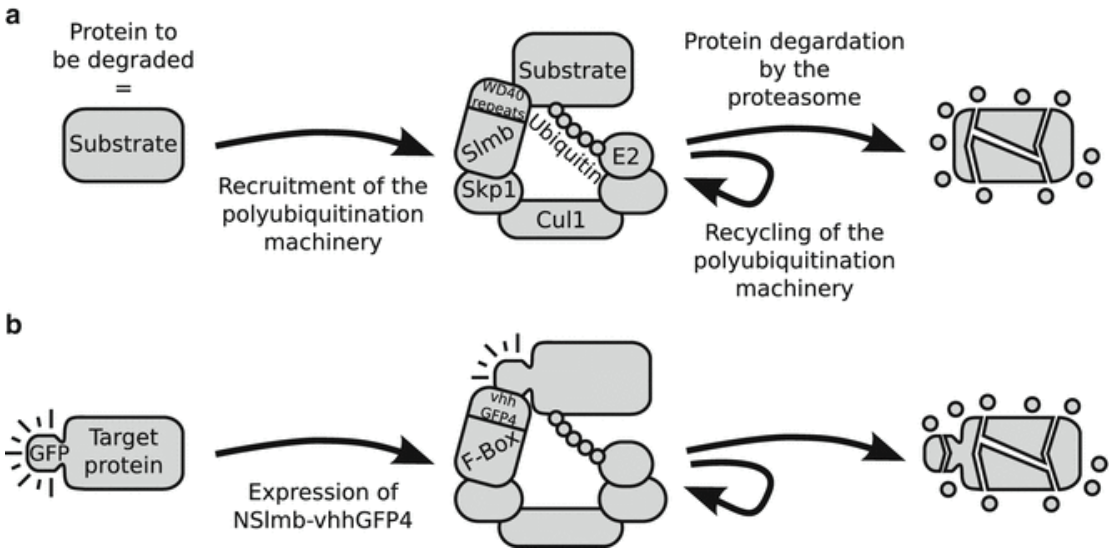


Figure 5.5 Schematic illustration of degradFP. A) Ubiquitin proteasome degradation pathway can be highjacked B) to drive expression of Nslmb-vhhGFP4 fusion protein in tissue specific manner. Now, GFP tagged proteins can be polyubiquitinated and sent to proteasome for degradation. Adapted from Caussinus and Affolter, 2016.

5.1.3.6 CRISPR/Cas9

Clustered Regularly Interspaced Short Palindromic Repeats/ CRISPR-associated endonuclease 9 (CRISPR/Cas9) is a technique in which Cas9 nuclease in a complex with synthetic guide RNA is delivered into a cell, allowing for cutting of the genomic material at site specified by the guide RNA and subsequent modifications such as insertions of a portion of the genome (Basset and Liu, 2014).

In *Drosophila*, CRISPR/Cas9 editing relies on synthesis and cloning of specific vector containing the Cas9 and gRNA complex. This plasmid is then microinjected to *Drosophila* embryos for gene editing, and the editing in the animals is validated by PCR and sequencing.

5.1.4 FlyBase

Fly base (http://www.flybase.org) is the primary database for information about the fruit fly genome and a range of information can be found on the website. This ranges from gene sequence and its function to expression patterns, phenotypes and genetic interactions, curated from variety of sources such as research publications, large genome sequencing projects and other online resources. The website also provides links to fly strains available for ordering from stock centres. This is therefore an invaluable tool in *Drosophila* research.

5.1.5 Modelling loss of function of *RFC1* and *ARHGAP19* using *Drosophila melanogaster*

Studying loss of function of a gene is an important area of research because it allows to model phenotypic changes in a system, caused by the absence or reduced expression of a gene or its transcriptional or translational products. This can lead to discovery of important interactions between genes, their products or indeed pathways that they play a role in. Importantly, if an animal model displays similar phenotype to this of human patients, it provides better confidence in the genotype-phenotype correlation.

Techniques such as RNA interference or gene knockouts can be used and after careful consideration of various model systems, we saw *Drosophila melanogaster* as a beneficial model for gaining understanding of function of *RFC1* and *ARHGAP19*.

Considering that both the genes, thoroughly described in chapters 2 and 3 for *RFC1* and chapter 4 for *ARHGAP19*, are inherited in autosomal recessive mode with considerable evidence for loss-of-function mechanisms for both, I used *Drosophila* knock-down and knock-out models to assess phenotypic changes in the generated fly models and the implications of these findings to patient genotype-phenotype correlations.

5.1.5.1 RFC1 and its fly homolog Gnf1

5.1.5.1.1 Gnf1 is a *Drosophila melanogaster* ortholog of human *RFC1*

The *Drosophila* genome contains a single *RFC1* ortholog termed *Gnf1*, Gnf1 exhibits 55% similarity and 40% identity to the human RFC1 protein (fig.5.6B). Similarly to its human ortholog, *Drosophila* Gnf1 is essential for cell cycle progression and has been shown to physically interact with PCNA during DNA replication (Tsuchiya *et al*., 2007). Recent data from single cell RNA sequencing shows that Gnf1 is moderately expressed in neurons throughout the adult fly brain (Janssens *et al*., 2022). However, the roles of Gnf1 in post-mitotic cells, including neurons, remain poorly defined.

In this thesis, in collaboration with Prof James Jepson and his lab at UCL, I use RNA interference to knock down the expression of Gnf1 in *Drosophila* neurons, and to test how this impacts survival, motor function and the DNA damage response.

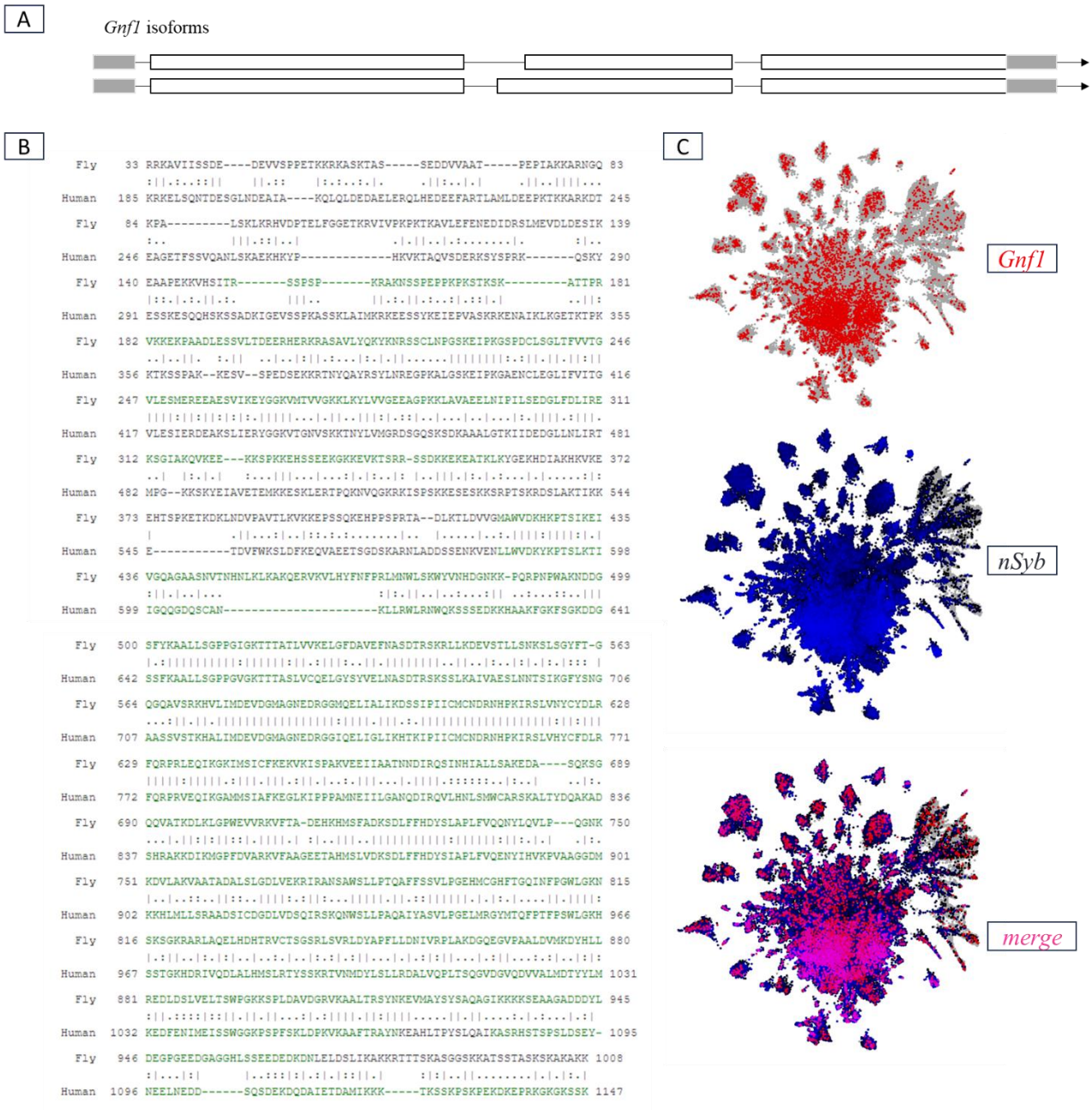


Figure 5.6 Gnf1 is a *Drosophila melanogaster* ortholog of human RFC1. A) A schematic of two known Gnf1 isoforms. B) A protein sequence alignment of Fly and Human Gnf1 and RFC1 proteins. Gnf1 exhibits 55% similarity and 40% identity to the human RFC1 protein. C) Adult fly brain RNA expression pattern (SCope, Davie *et al.*, 2018) of Gnf1, neuronal nSyb and a merge of the two show that Gnf1 is expressed in adult fly neurons

5.1.5.2 ARHGAP19 and its fly homolog RhoGAP54D

5.1.5.2.1 RhoGAP54D is a *Drosophila melanogaster* ortholog of human *ARHGAP19*

The *Drosophila* genome contains a single ARHGAP19 ortholog termed RhoGAP54D which exhibits 51% similarity and 31% identity (fig.5.7). Importantly both proteins contain GAP domains that are 37% identical. Recent data from single cell RNA sequencing suggest very sparse expression of RhoGAP54D in neuron ad glia in the adult fly brain. However, since the protein is not well characterised, we investigated its expression pattern in the fly. Knowing where a gene is expressed and the resulting protein is localised, provides invaluable clues to its possible function and interactions with other genes, RNAs or proteins.

In this thesis, I use a range of tools to knock down the expression of RhoGAP54D in the fly. I further assess the fly phenotype and provide evidence for loss-of-function mechanism underlying the patient phenotype.

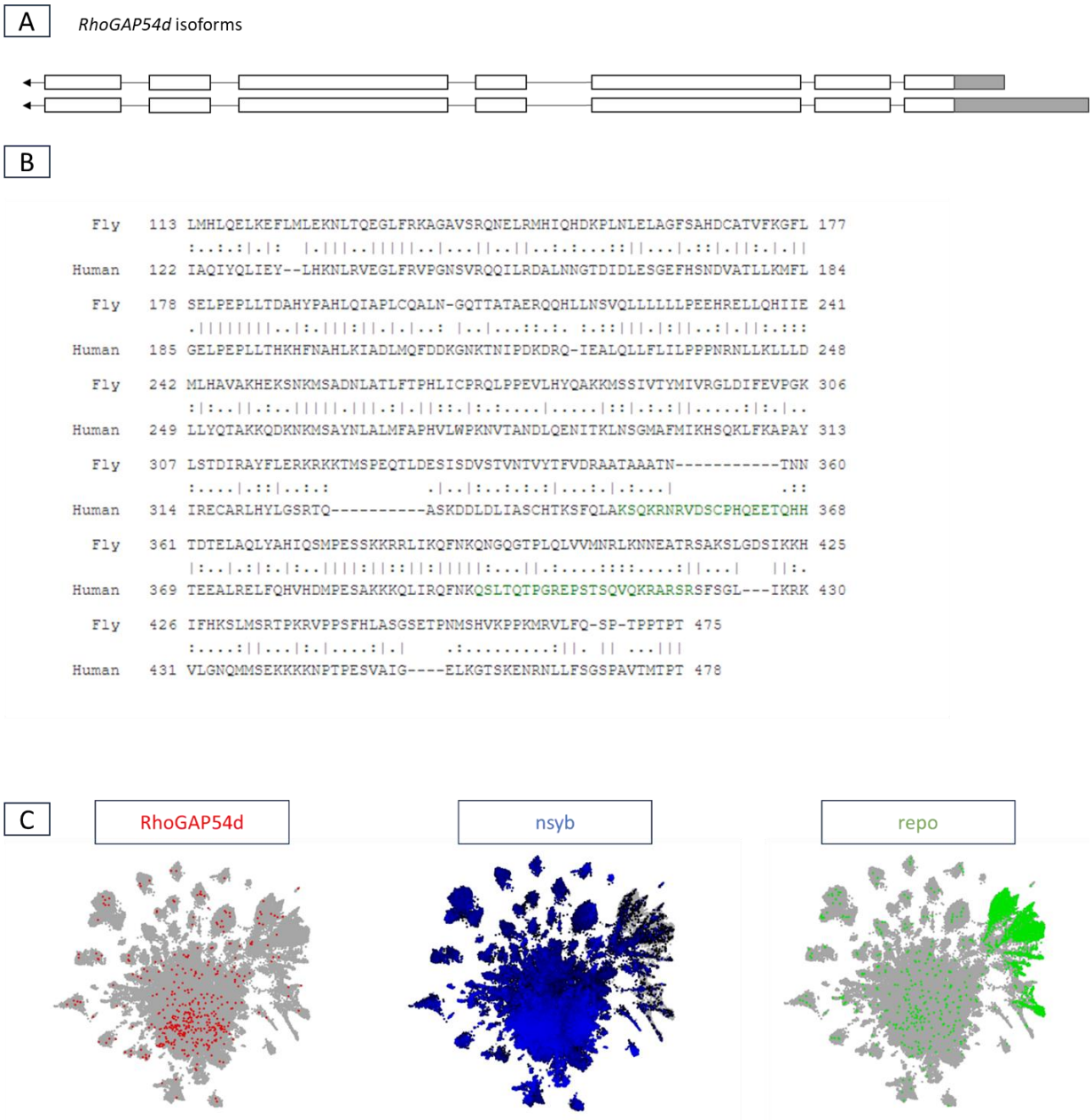


Figure 5.7 RhoGAP54D is a *Drosophila melanogaster* ortholog of human ARHGAP19. A) A schematic of two known RhoGAP54D isoforms. B) A protein sequence alignment of Fly and Human RhoGAP54D and ARHGAP19 proteins. RhoGAP54D exhibits 51% similarity and 31% identity to the human ARHGAP19 protein. C) Adult fly brain RNA expression pattern (from SCope, Davie *et al*., 2018) of *RhoGAP54D*, neuronal *nsyb* and glial *repo* indicate that RhoGAP54D is sparsely expressed in adult fly brain

5.2 Materials and methods

Main contributors to the methods (and results) in this chapter are listed in table 5.1

Contribution table	
Methods used	**Contributors**
Drosophila crossing and husbandry	**Natalia Dominik**
Drosophila behavioural studies	**Natalia Dominik**
Drosophila lifespan assays	**Natalia Dominik**
Drosophila brain dissections	James Jepson, Simon Lowe, Gabriel Aughey
Immunostaining	Simon Lowe, James Jepson, Gabriel Aughey, **Natalia Dominik**
Confocal microscopy	Simon Lowe, **Natalia Dominik**

Table 5.1. Main contributors to the methods (and results) in this chapter

Throughout this thesis chapter I used common tools and techniques in *Drosophila* research of both Gnf1 and RhoGAP54D and I provide the description for those in Materials and Methods section. However, I specify when a method was only employed for study of one of the genes.

5.2.1 Basic considerations

Cultures of flies for manipulation were maintained at 25°C and flipped onto new food as necessary, every 10-14 days. Flies were kept in standard fly food as shown in table 5.2 unless otherwise stated.

Flies were anaesthetised on CO_2 mats for sorting and collections. Female virgins of stocks were collected as and when needed and sexing of flies occurred by visual comparison of sex organs. For all experimental assays, only male flies of correct genotypes were used.

reagent	concentration
agar	10 g/L
surose	15 g/L
glucose	33 g/L
yeast	35 g/L
maize meal	15 g/L
wheatgerm	10 g/L
treacle	30 g/L
soya flour	7.22 g/L
nipagin	1g (in 10 mL ethanol)/L
propionic acid	5 ml/L

Table 5.2 Fly food ingredients for the standard fly food used

5.2.2 Drug application in Gnf1 assays

cis-Diammineplatinum(II) dichloride (Sigma) (referred to as cisplatin hereafter) was dissolved in distilled water at 37°C to achieve stock concentration of 5uM and mixed into 10 ml of fly food at a final concentration of 100 µg/ml. An equal amount of distilled water was mixed into fly food as control.

5.2.3 GAL4 – UAS for gene knockdown using RNAi

GAL4-UAS system was leveraged for tissue specific knockdown of gene products of interest using RNAi. The tissue specific drivers and RNAi lines used in this thesis are listed in table 5.3. The drivers used were outcrossed into an isogenized background (iso31) for five generations, with the X-linked y[1] and v[1] markers removed in the process. The RNAi used in the project were however not outcrossed into an isogenised background, therefore, to ensure that any X-linked y[1] and v[1] markers were removed in the progeny used for experimental assays, virgin females of tissue specific Gal4 drivers were always crossed with the males of RNAi of interest and subsequently only male progeny of the correct genotype was used for experimental assays.

RNAi ID	**genotype**	
Gnf1 RNAi 35423	y[1] sc[*] v[1] sev[21]; P{y[+t7.7] v[+t1.8]=TRiP.GL00346}attP2/TM3, Sb[1]	
RhoGAP54d RNAi 54051	y1 v1; P{TRiP.HMS03522}attP40	
RhoGAP54d RNAi 31144	y1 v1; P{TRiP.JF01619}attP2	
RhoGAP54d RNAi 54459	y1 sc* v1 sev21; P{TRiP.HMS03719}attP2	
Driver ID	**genotype**	**tissue specificity**
actinGAL4	y1 w*; P{Act5C-GAL4}25FO1/CyO, y+	global
tubGAL4	y[1] w[*]; P{w[+mC]=tubP-GAL4}LL7/TM3, Sb[1] Ser[1]	global
nsybGAL4	y1 w*; P{nSyb-GAL4.S}3/TM2	neuronal
repoGAL4	w1118; P{GAL4}repo/TM6b, Sb1, y	glial
RhoGAP54d CrimicGAL4	y[1] w[67c23]; P{y[+mDint2] w[+mC]=EPgy2}RhoGAP54D[EY22686]	RhoGAP54d Crimic Line

Table 5.3 RNAi lines and GAL4 drivers used together with their genotype. GAL4 drivers are listed with the tissue they drive expression in.

Each experimental cross of GAL4>RNAi must have an appropriate control, as presented in figure 5.8, and these include: 1, the UAS-RNAi responder line crossed with a wild-type fly to produce genetically controlled offspring with silent RNAi construct (RNAi>+) and 2, the tissue specific driver line crossed with a wild-type fly to produce genetically controlled offspring with GAL4 (GAL4>+).

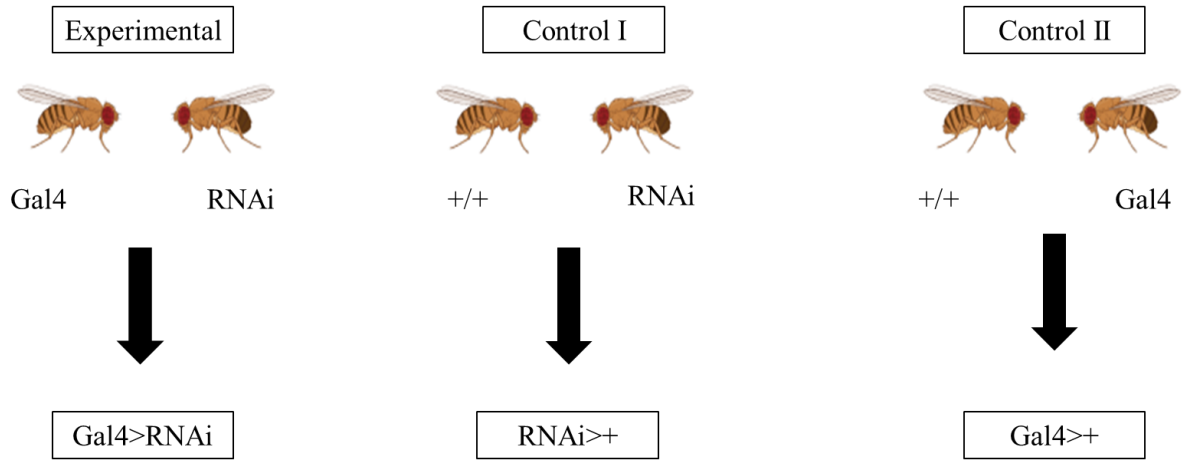


Figure 5.8 Experimental fly and its controls schematic. Each experimental genotype must have specific control lines and they include silent RNAi construct (RNAi>+) and a silent tissue specific driver (Gal4>+)

5.2.3.1 Fly husbandry - RNAi crossing schemes:

The RNAi lines often have a genetic background on the X sex chromosome which may interfere with correct comparison between the experimental and control crosses. It is therefore imperative that males of the RNAi line stocks are crossed with female virgins of the driver line stocks to remove unwanted genetic background mutations. For all the experiments involving adult flies I used male flies of the correct genotype only. Crossed flies were kept at 25°C in 12-hour light 12-hour dark conditions.

5.2.4 GAL4-UAS system for RhoGAP54D knockdown using degradFP

In addition to RNAi knockdowns, degrad:FP system was employed for genetic knockdown in RhoGAP54D project to establish additional model validating the findings of RNA interference model. DegradFP is a genetic system that promotes degradation of GFP-tagged fusion proteins via the ubiquitin pathway. We obtained a RhoGAP54D::GFP knock-in allele from Yohanns Bellaiche Curie Institute, France, who systematically analysed RhoGAPs and GEFs in *Drosophila* by tagging the GAP and GEF proteins with GFP (di Pietro *et al*., 2023). The degrad:FP knock-down was driven by ubiquitously expressed *tubulin*-Gal4 (*tub*-Gal4) driver.

5.2.5 RNA extraction

RNA extraction for validation of RNAi knockdowns by qPCR was performed for *actin-Gal4* RhoGAP54D RNAi crosses and for *nsyb-Gal4* Gnf1 RNAi crosses. An unrelated RNAi targeting mCherry was crossed with the *actin-Gal4* and separately with *nsyb-Gal4* and used as controls.

Flies were decapacitated at 3-5 days after eclosion and male flies were used. RNA extraction followed from 10-12 heads per sample. Each genotype had 3 biological replicates where possible.

Heads were immediately put into 500µl ice cold trizol. Homogenisation with a hand mixer followed for approximately 1 minute per sample or until visibly homogeneous. 300µl additional trizol was added per sample and incubated at room temperature for 5 minutes. Subsequently, 160µl of chloroform was added and shaken vigorously for 15 seconds. 3 minutes incubation at room temperature followed. The samples were centrifuged at 11,000 *xg* for 15 minutes at 4°C to separate into lower red phenol-chloroform phase, an interphase, and a colourless upper aqueous phase. The aqueous phase containing RNA was collected into fresh Eppendorf tubes and 1 µL glycogen (20 µg/µl) and 400 µL isopropanol were added. Tubes were vortexed for 10 seconds and incubated for 1 hour at -20°C. Samples were centrifuged at 11,000 *xg* for 10 minutes at 4°C and supernatant discarded. The pellet was immediately washed with 1ml 75% ethanol followed by brief vortexing and another centrifuge spin at 6,000 *xg* at 4°C for 5 min. The supernatant was discarded, and the pellet dried for 2-5 minutes. RNA was resuspended with 30µl DPEC water RNA concentration and purity of the samples were assessed using NanoDrop equipment (NanoDrop Technologies Inc., Wilmington, DE).

5.2.6 cDNA synthesis

cDNA synthesis was performed using Superscript III (Thermofisher). A master mix was prepared for each sample by using 1µg of RNA, 1µl random hexamers (50ng/uL), 1µl of dNTPs and the total volume was made up to 10µl using DEPC-treated water. The mixture was incubated at 65°C for 5 min, then placed on ice for 1min. cDNA synthesis mix was prepared by mixing 2µl of 10X RT buffer, 4µl of 25mM MgCl2, 2µl of 0.1 M DTT, 1µl of RNase OUT and 1µl of SuperScript. 10µl of the cDNA mixture was added to the RNA on ice and incubated in the following cycler conditions: 10minutes at 25°C, 50 minutes at 50°C and 5 minutes at 85°C. Subsequently, 1µl RNase H was added and further 20 minutes incubation at 37°C degrees followed.

5.2.7 Quantitative PCR

I performed qPCR for validation of RNAi knockdowns in both RhoGAP54D and Gnf1 crosses with a driver crossed to *mcherry* RNAi as control. Primer design was performed in-silico in the following website https://www.flyrnai.org/flyprimerbank. The primers used are listed in table 5.4.

Primer name	Primer sequence
RpL4 forward	TCCACCTTGAAGAAGGGCTA
RpL4 reverse	TTGCGGATCTCCTCAGACTT
Gnf1 forward	CAACGCGGCATTGACTCCT
Gnf1 reverse	CGTCTCTCCATTTTCGGCCTC
RhoGAP54d forward	ATGGAAGCAACGATGGATACG
RhoGAP54d reverse	CTCGTGACAGGGGAGATCGAA

Table 5.4 Primers used for qPCR validation of RNAi knockdowns. RpL4 is a housekeeping primer used in both *RhoGAP54D* and *Gnf1* validations

Reaction volume of 15µl was prepared by mixing 7.5µl of Fast SYBR Green Master Mix (2x) with 5µl of DPEC water and 0.75µl each forward and reverse primer (10uM) and 1µl of DNA. Samples were plated on 96 well plate in technical duplicates and plate run in QuantStudio Real-Time PCR machine in the following standard conditions:

qPCR cycler conditions	**95°C 20 secs;** **[95°C 1 secs** **60°C 20 secs] x40;** **Ramp rate increase 2.63 °C/sec** **95°C 15 secs** **60°C 1 min** **95°C 15 secs**

Table 5.5 Quantitative PCR cycler conditions

For relative gene expression, the comparative cycle threshold (ΔΔCT) values were calculated with the QuantStudio Design&Analysis Software (Thermo Fisher Scientific) with RpL4 as housekeeping gene and expressed as x-fold change to controls.

5.2.8 *Drosophila* Activity Monitor

To track the adult locomotion, the *Drosophila* Activity Monitor (DAM) (Trikinetics) measures fly movement over 24-hour period (figure 5.9). Flies were loaded into glass tubes, one fly per tube, which in turn is loaded into a monitor that connects to a computer. There are 32 slots available per system and it is put into an incubator where a controlled environment is set up. I used a 24-hour period with 12 hours lights turned off and 12 hours lights turned on and the temperature of 25ºC. The experimental flies and the corresponding controls are loaded in the monitor 48 hours prior to experimental period to become accustomed to the environment. The movement of the flies is measured when they cross midline where an infrared beam is broken (Pfeiffenberger 2010).

For *RhoGAP54D* experiments, the flies were assessed in the DAM at 3-7 days old only in line with young age of onset of neuropathy in patients with *ARHGAP19* mutations. For the *Gnf1* experiments, the flies were assessed at 3-7 day old, 21-23 day old and 40-42 day old in line with the older age of onset of CANVAS. Additionally, cisplatin-treated male flies were recorded by DAM at 10-13 days old only. The behavioural tubes contained 4% sucrose and 2% agar for untreated flies and 4% sucrose and 2% agar with 100ul/ml cisplatin for treated flies.

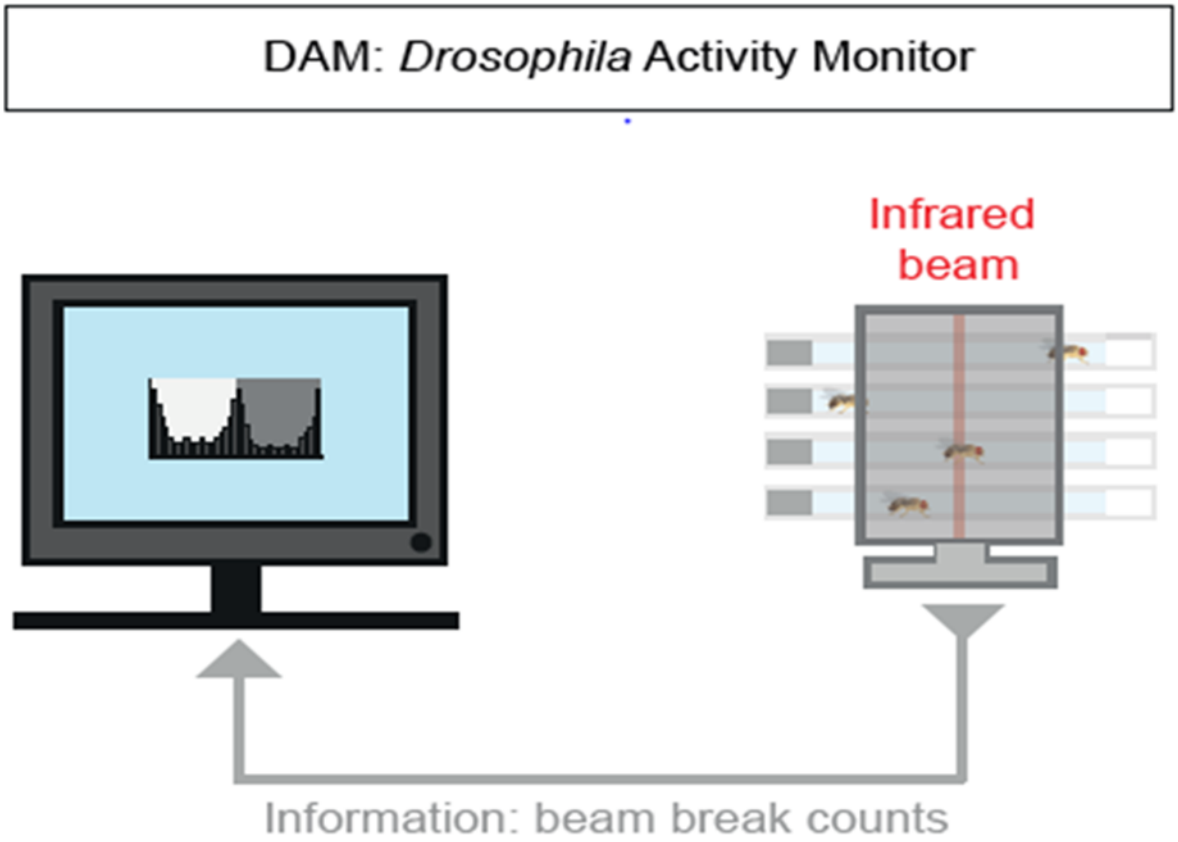


Figure 5.9 Drosophila activity monitor schematic. DAM measures the fly movement over a defined period of time and consist of glass tubes in which experimental and control flies are loaded.

5.2.8.1 Data analysis

Raw data from DAM was inputted into Excel showing the number of beam breaks per minute per fly. This was summed up for 24-hour period per fly, and GraphPadPrism is used to visualise results of each genotype versus its controls. Statistical analyses were carried out in Prism depending on Shapiro-Wilk normality test and they might include one way ANOVA followed by post hoc Dunnett's test for multiple comparisons or Kruskal-Wallis test when normality could not be assumed.

5.2.9 Immunostaining

5.2.9.1 Protein expression investigation of RhoGAP54D

To investigate the expression of *RhoGAP54D,* we utilised RhoGAP54DCrimicGal4 in concert with the UAS-tdTomato reporter.

Brains were fixed for 20 min at room temperature via incubation in 4% paraformaldehyde (MP Biomedicals) and blocked for 1hour in 5% normal goat serum in Phosphate Buffered Saline (PBS) containing 0.3% Triton-X (Sigma-Aldrich) (0.3% PBT). Primary antibodies were as follows: mouse anti-Bruchpilot (BRP) (Developmental Studies Hybridoma Bank), 1:50; rabbit anti-dsRed (Clontech), 1:1000. Secondary antibodies were: goat anti-rabbit AlexaFluor-555 (ThermoFisher), 1:1000; and goat anti-mouse AlexaFluor-647 (ThermoFisher), 1:500. Brains were incubated in primary and secondary antibodies overnight at 4°C. Brains were washed then mounted and imaged in SlowFade Gold anti-fade mounting solution (ThermoFisher Scientific).

5.2.9.2 H2AV staining for Gnf1 flies

To investigate presence of DNA damage markers (Lake *et al*., 2013), Gnf1 knockdown and control brains were stained with H2Av primary antibody (Developmental Studies Hybridoma Bank) which is an antibody against phosphorylated H2Av which is a marker for double stranded DNA breaks (Lake *et al*., 2-13). H2Av antibody was used at a final concentration of 1:200 following the immunostaining protocol described above. DAPI counterstain (ThermoFisher Scientific) was used at 1:1000 for normalising the fluorescence to control for technical differences.

Fly brains were dissected at 12 days old for cisplatin treated flies and 40-42 days old for aged flies on normal fly food.

5.2.9.2.1 Confocal microscopy

Images were taken with a Zeiss LSM 710 confocal microscope with an EC ‘Plan-Neofluar’ 20x/0.50 M27 air objective, taking z-stacks through the entire brain with step sizes of 1-5µm.

5.2.9.2.2 Image-J for H2AV staining quantification

Images were analysed using ImageJ: z-stacks were 3-D projected using a maximum intensity projection and the region of interest (ROI) around the central brain was drawn in the DAPI image with the drawing tool as shown in figure 5.10 and measured with image J. The same ROI was taken in the H2Av image, and both measurements recorded. An area of about 1cm^2 was also drawn outside of visible tissue to compare background noise on both DAPI and H2AV images.

Images were compared to controls imaged on the same day and the data was normalised to expression of H2Av in *nsyb-Gal4*>+ and statistical analysis performed in Graphpad Prism.

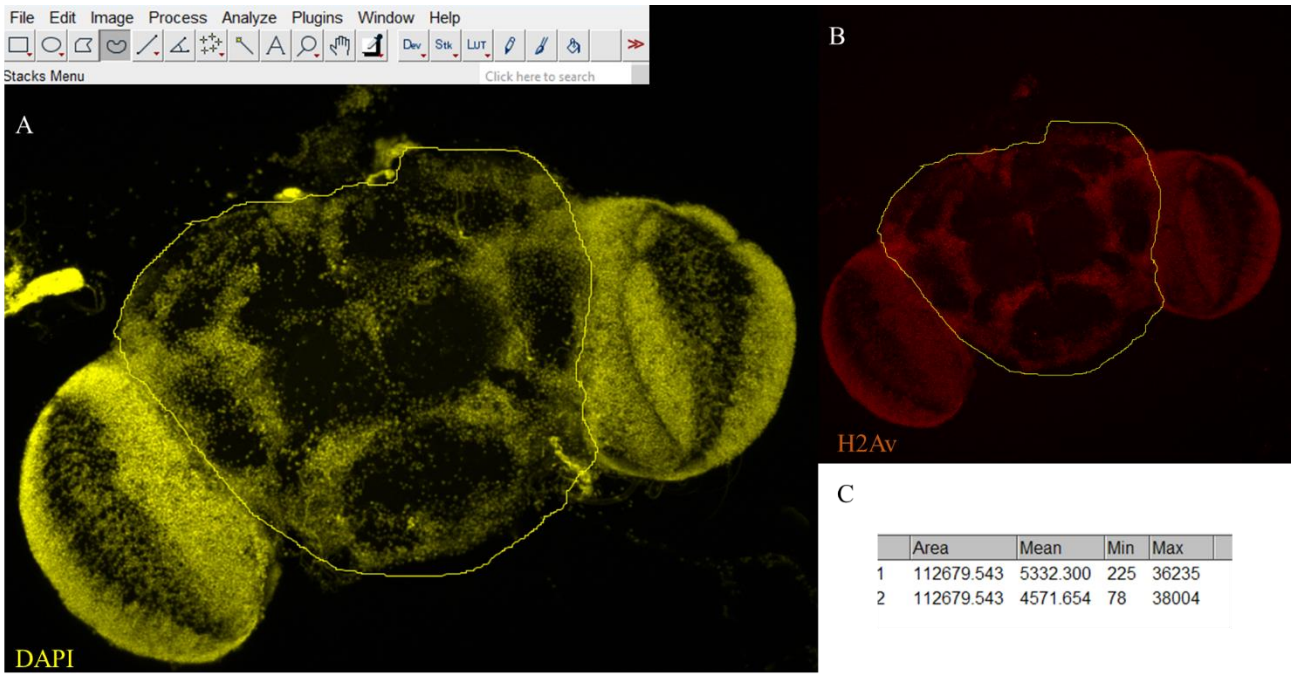


Figure 5.10 Image J analysis of H2Av staining in adult fly brain. A) An area of interest is drawn with the free drawing tool around the central brain in DAPI stained section. B) The same area of interest is copied and pasted into the H2AV stained section, and both used to measure the staining intensity which is further analysed in Excel and Graphpad Prism

5.2.10 Drosophila lifespan assays

Gnf1 experimental and control flies were subjected to survival assay.

Lifespan assays were performed on 100 newly eclosed male flies from each experimental and control group, both fed with cisplatin and on standard fly food only. Flies were placed in vials of 10 flies per group and transferred onto fresh food every 2 days. The number of dead flies were recorded at every transfer and the data were plotted as a Kaplan-Meier survival plot.

5.3 Results

5.3.1 Modelling loss-of-function in Gnf1

5.3.1.1 Knocking down Gnf1 in post-mitotic neurons results in a movement phenotype in 40-day old flies

Actin-Gal4 is often used with RNAi to confirm viability of the progeny – if there is no progeny it suggests that the RNAi lines have strong effect on silencing the gene expression and/or that the gene is necessary to fly function and survival. I performed ubiquitous knock-down of Gnf1 using global *actin-Gal4* driver and Gnf1 RNAi 35423 line obtained from Bloomington *Drosophila* stock centre. No viable progeny was observed, suggesting *Gnf1* has a crucial role for organism survival. Indeed, Gnf1 has been shown to be essential for cell cycle progression (Tsuchiya *et al*., 2007).

Recent single-cell RNA sequencing data have demonstrated that Gnf1 is expressed in post-mitotic neurons throughout the adult *Drosophila* brain. Since a global knockdown of Gnf1 resulted in total lethality, I tested whether *Gnf1* KD in post-mitotic neurons recapitulated aspects of the RFC1 linked CANVAS, focusing on ataxia as a key clinical phenotype.

Gnf1 RNAi 35423 line was crossed with *nsyb-Gal4*, and the progeny was collected at three time points for subjecting to DAM assay. Young KD flies of 3-5 days and middle-aged flies of 21 days showed no significant locomotion differences compared to the controls (fig.5.11).

Since RFC1 disease in humans has late onset, I next assessed whether knocking down the expression of Gnf1 in neurons and aging the flies to older age of 40 days recapitulated the patient movement phenotype. The DAM experiments for 40-day aged flies showed a significant decrease in movement in Gnf1 RNAi 35423 > *nsyb-Gal4* compared to the controls ($p<0.0001$) (fig.5.12).

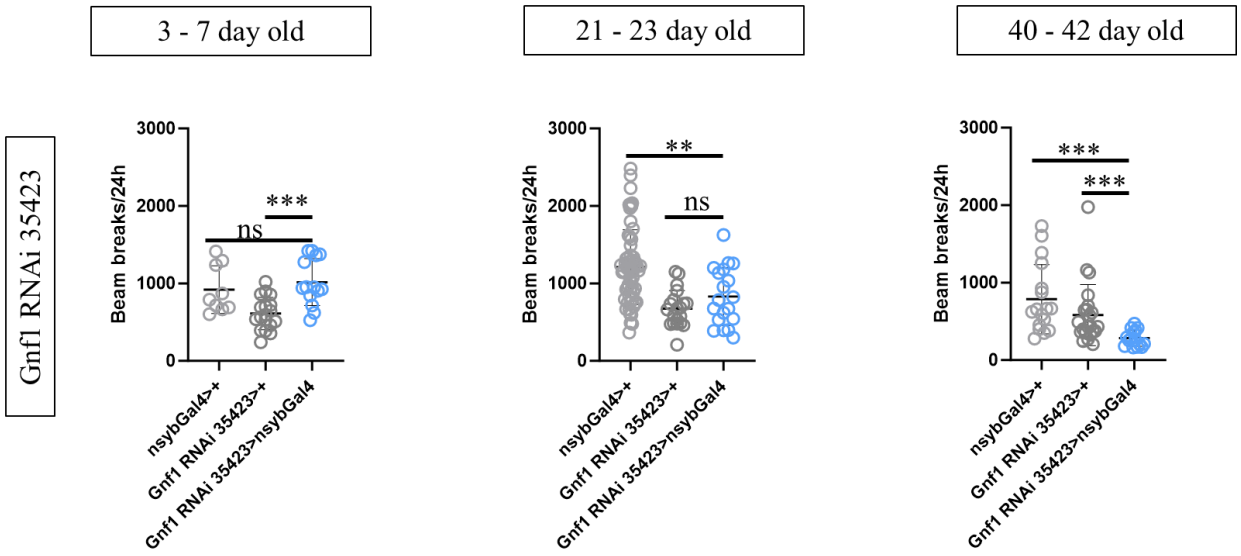


Figure 5.11 Gnf1 RNAi knockdown DAM results for 7-7, 21-23 and 40–42-day old flies with neuronal nsybGal4 driver. Statistically significant movement decrease is observed for RNAi 35423 at 40 days as compared to RNAi and driver controls respectively ($p<0.001$). No significant movement decrease is observed in the 2 earlier time-points of 3-7 and 21 – 23 day old flies ($p>0.05$). Data are shown as means ± SD. One-way ANOVA. ns = non-significant. ** = $p<0.01$, *** = $p<0.001$.

5.3.1.2 Knockdown of Gnf1 causes reduced lifespan

Leveraging advantages of *Drosophila* as a model organism, such as quick lifespan and robust numbers, I was able to collect high numbers of flies (100 per genotype) and subject them to a survival assay to determine whether pan-neuronal *Gnf1* KD has a consequence on lifespan on the fly. Determinations of life span are often challenging in human patients due to variables such as comorbidities, unequal access to healthcare and environmental factors; and a model organism allows for monitoring of an animal within a set environment without external variables.

Gnf1 RNAi 35423>*nsyb-Gal4* showed significantly reduced lifespan compared to controls (fig.5.12) ($p<0.001$), with death of these pan-neuronal *Gnf1* KD flies first being observed at ~ 20 days compared to ~ 40 days in controls, and median lifespan being reduced from 48 and 56 in controls to 39.5 days in pan-neuronal *Gnf1* KD flies.

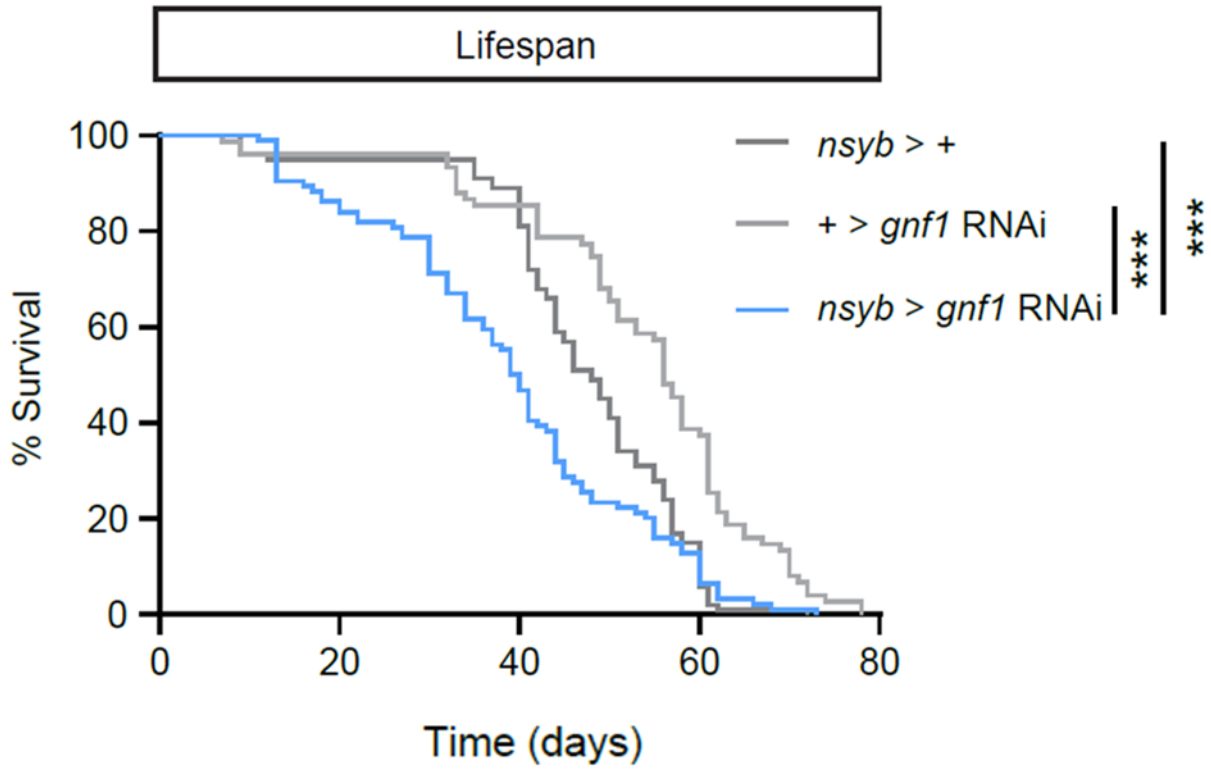


Figure 5.12 Gnf1 KD results in reduced lifespan compared to the driver and RNAi controls ($p<0.001$). *** = $p<0.001$, Log-rank test. A 100 flies were tested per each group.

5.3.1.3 qPCR for efficacy of RNAi

I performed qPCR on the dissected brain tissue of Gnf1 knockdown model and mCherry RNAi>*nsyb-Gal4* as a control to assess degree of genetic knockdown. A high degree of knockdown is observed, and it reaches as much as 60% suggesting that the RNAi line is efficacious (5.13).

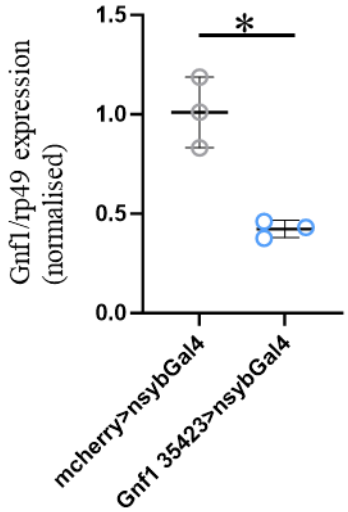


Figure 5.13 Quantitative PCR result of Gnf1 knockdown. Significant knockdown ($p<0.05$) of Gnf1 is observed with Gnf1 RNAi 35423 expressed in neuronal tissue. Data are shown as means ± SD. T-test, * = $p<0.05$.

5.3.1.4 DNA damage is induced in Gnf1 knockdown

The ability to recognize and repair DNA damage is essential for cellular and organismal survival. Gnf1 is an important gene in DNA damage and repair in dividing human cells, yet whether *RFC1* or its *Drosophila* ortholog Gnf1 plays a similar role in post-mitotic neurons is unclear. Since such a function could potentially explain the progressive phenotypes observed in pan-neuronal Gnf1 knockdown (KD) flies and by extension, CANVAS patients - we immuno-stained brains of Gnf1 RNAi 35423>*nsyb-Gal4* KD flies and controls with antibodies against phosphorylated H2Av, a marker for double stranded DNA breaks (Lake *et al*., 2013).

The flies were investigated at 40 days old when locomotor defects and substantial mortality are apparent in Gnf1 RNAi 35423>*nsyb-Gal4* KD flies and a significant increase ($p<0.05$) in brain-wide H2Av staining was observed in the experimental fly as opposed to controls (fig.5.14). This suggest that knocking down Gnf1 in neurons leads to an increase in DNA damage at this age.

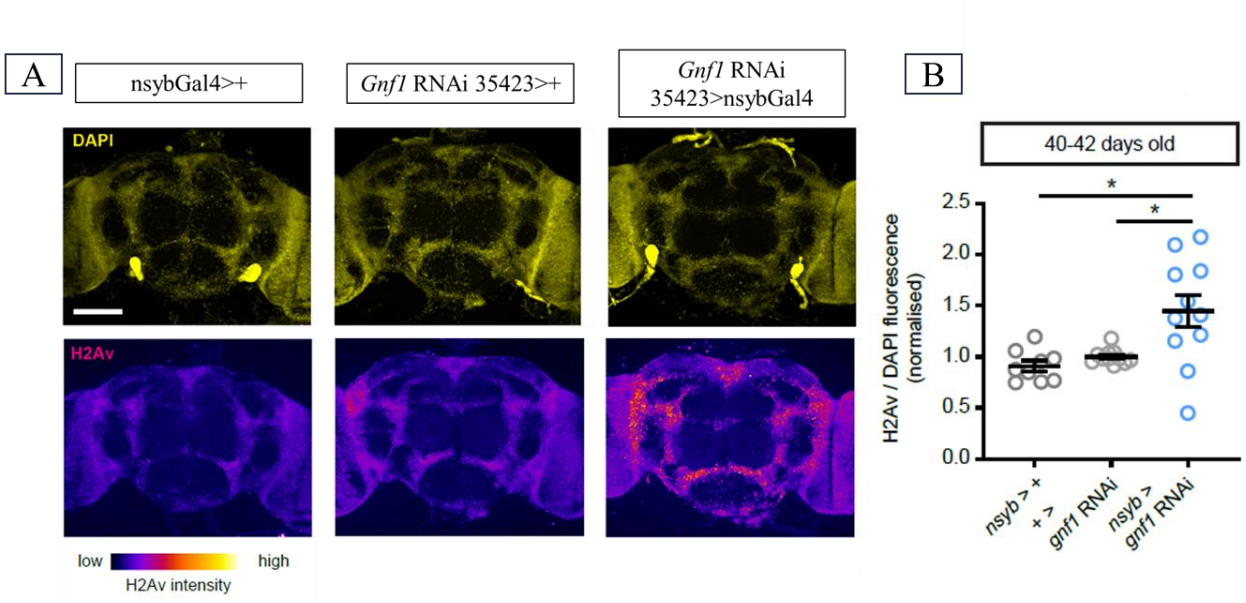


Figure 5.14 DNA damage accumulation A) H2Av staining in 40 day old adult fly brains. DAPI top; H2Av bottom. B) A significant ($p<0.05$) increase in neuronal H2Av staining in pan-neuronal Gnf1 KD flies compared to controls. *Data are shown as means ± SEM. One-way ANOVA, * = $p<0.05$.*

5.3.1.5 DNA damaging agent treatment

To further elucidate the loss of function of Gnf1 in neuronal tissues and whether the observed phenotypes are connected to DNA damage, the flies were fed with 100µg/ml of cisplatin – a known DNA damage agent. A multitude of DNA damage agents are available; however, cisplatin was specifically chosen for the purpose of this study as it is a known DNA damaging agent with neurotoxic properties and causative of sensory neuropathy (Chen *et al*.,2024) similarly to RFC1 repeat expansion found in CANVAS patients.

5.3.1.5.1 Cisplatin treatment significantly reduces fly survival rate

To investigate any possible consequences of cisplatin on lifespan of the Gnf1 flies, a hundred flies per genotype were collected and fed with cisplatin enriched food. They were transferred onto new food every two days and the survival measured. Strikingly, a stark decline to survival rate was observed with the experimental flies dying off around 20 days mark and the respective controls surviving only to about 30 days. It is important to note that the controls follow near identical survival curve, and the experimental flies have significantly reduced lifespan (fig.5.15).

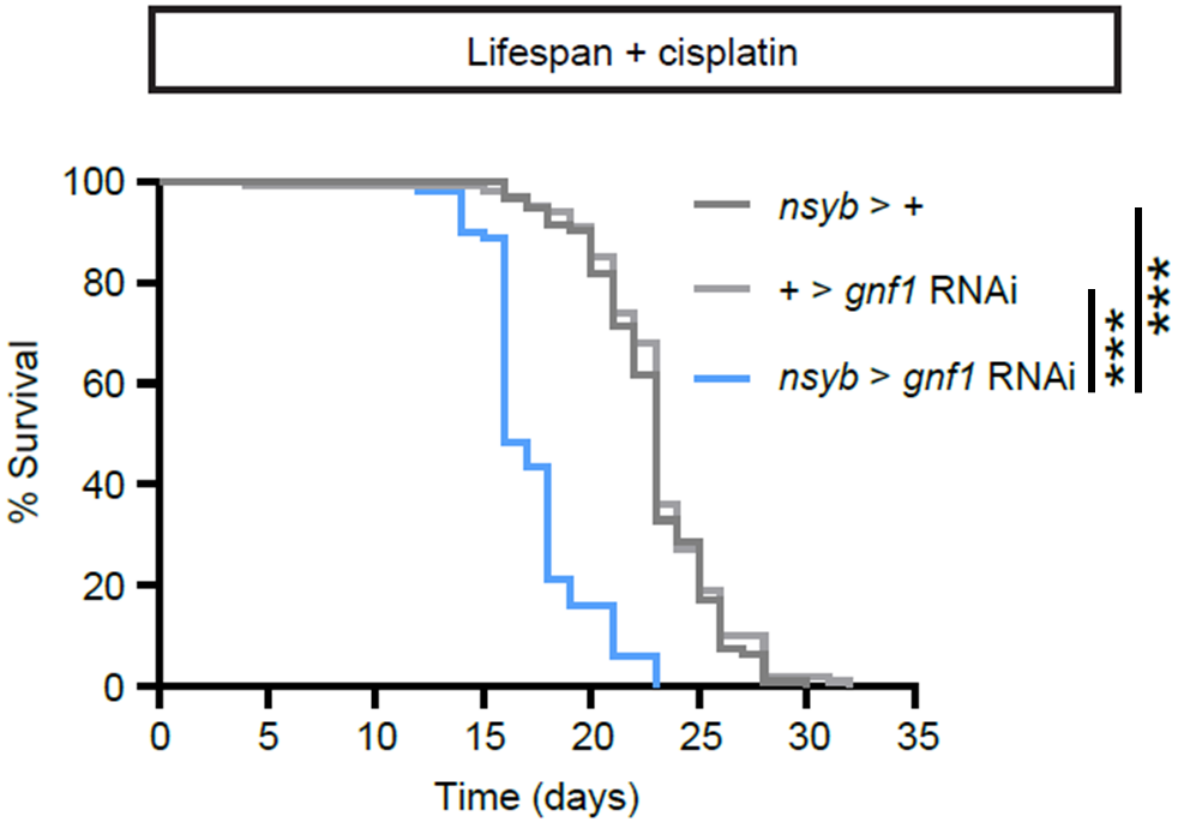


Figure 5.15 Gnf1 KD on cisplatin enriched food results in reduced lifespan as opposed to the driver and RNAi controls ($p<0.001$). *** = $p<0.001$, Log-rank test. A 100 flies were tested per each group.

5.3.1.5.2 H2AV staining in cisplatin treated flies reveals significant increase of DNA damage

Having established that cisplatin experimental flies die by 20 days, we decided to use 12 days old flies, as a last timepoint before stark decline in viable flies, to investigate possible exacerbation of DNA damage with cisplatin. We observed a significant increase in neuronal H2Av staining in pan-neuronal *Gnf1* KD flies compared to controls (fig.5.16) which may suggest accumulation of DNA damage due to faulty DNA damage repair system in these flies.

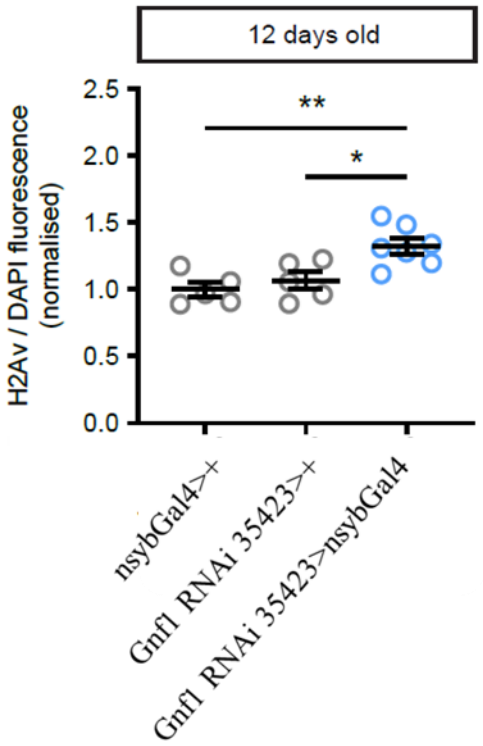


Figure 5.16 H2Av staining in 12-day old flies fed with cisplatin. Significant increase of H2Av staining is observed in the Gnf1 KD compared to controls. Data are shown as means ± SEM. One-way ANOVA, * = $p<0.05$, ** = $p<0.01$

5.3.1.5.3 Cisplatin treatment causes hyperactivity phenotype

Since we investigated DNA damage in 12-day old flies, we subjected cisplatin treated flies of the same age to DAM to investigate their total movement. Interestingly, we observed a novel locomotor phenotype that emerged prior to early mortality in 12-day old neuronal Gnf1 KD flies fed cisplatin. This phenotype was characterised by overall locomotor hyperactivity over 24 h relative to controls (fig 5.17A).

DAM allows for extracting data on the temporal patterns of activity in *Drosophila*. Since the flies are in the monitor for 48 hours prior the experimental period, they are accustomed to the 12 h light 12 h dark conditions. They follow a pattern of activity where they exhibit low activity during dark and light periods (corresponding to periods of sleep), and elevated activity prior to lights on or off termed morning and evening anticipation. They then display startle responses during light changes and rapidly fall back asleep after these periods of activity (Brown *et al*., 2024).

Interestingly, the hyperactivity movement phenotype may be attributed to increased locomotor activity prior to lights-off, at Zeitgeber Time (ZT) of 9 to 11 hours, where a fly is expected to anticipate the light change due to being acclimatised to the environment. Conversely, at ZT 12-

14, when the anticipation, startle response and awake window has passed, the Gnf1 RNAi 35423>*nsyb-Gal4* now show reduced locomotion (fig.5.17).

Therefore, Gnf1 RNAi 35423>*nsyb-Gal4* KD flies treated with cisplatin display premature and heightened lights-off anticipation and a likely diminished startle response. It is yet unclear why cisplatin treatment in Gnf1 KD causes this advanced anticipation phenomenon. However, it is likely that clock neurons may be affected in these flies therefore circadian rhythms may be disrupted, which is consistent with loss of evening anticipation, and this could be a point for future investigations.

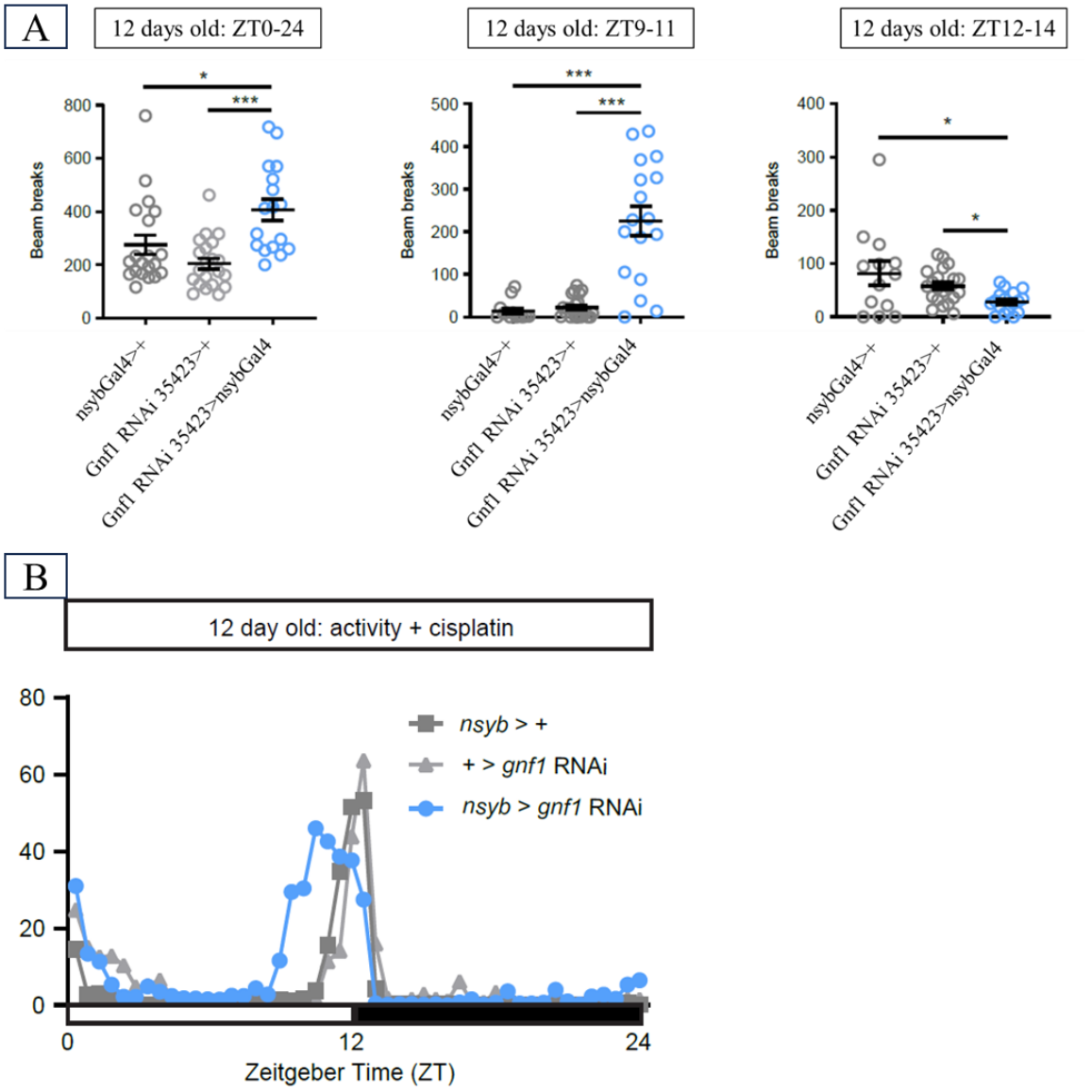


Figure 5.17 Novel locomotion hyperactivity phenotype A) Drosophila activity monitor locomotion of Gnf1 RNAi knockdown and controls treated with cisplatin. Hyperactivity is observed in 24-hour locomotion (ZT0-24) and in two hours prior to lights off (ZT9-11) however, significant movement decrease is observed after lights off (ZT12-14) B) The hyperactivity arises from increased locomotor activity prior to lights-off. Data are shown as means ± SEM. One-way ANOVA, * = $p<0.05$, *** = $p<0.001$

5.3.1.6 Conclusion to Gnf1 results

Through the use of Gnf1 RNA interference lines controlled by Gal4-UAS system and expressed pan-neuronally with *nsyb-Gal4* driver, I established Gnf1 knockdown model and used qPCR to validate its efficacy to over 50% knockdown in dissected adult brain tissue.

I observed age-dependant locomotion phenotype which recapitulated that of human patients. Moreover, these flies displayed significantly reduced lifespan as opposed to controls and an accumulation of DNA damage at 40 days assayed by immunofluorescence of H2Av DNA damage marker. These data support the premise that pan-neuronal *Gnf1* KD promotes the accumulation of DNA damage in fly neurons, leading to progressive neurological defects that impair motor function and lifespan.

5.3.2 Modelling loss-of-function in RhoGAP54D

5.3.2.1 A global knockdown of RhoGAP5dD results in a movement phenotype in 3-5 day old flies

I performed ubiquitous knock-down of RhoGAP54D using the global *actin-Gal4* driver and 3 available RhoGAP54D RNAi lines: RhoGAP54D RNAi 54459, RhoGAP54D RNAi 31144 and RhoGAP54D RNAi 54041.

Having set up the crosses, I observed viable progeny of *actin-Gal4*-RNAi for all the RNAi lines used. Given that *ARHGAP19* mutations perturb movement in humans, I collected the male progeny and subjected them to DAM to assess whether a global knockdown of RhoGAP54D has a consequence on locomotion of adult fly.

Of 3 RNAi lines, 2 showed no overall effect on movement of the flies: RhoGAP54D RNAi 54459>*actin-Gal4* and RhoGAP54D RNAi 31144>*actin-Gal4* (fig.5.18 A). However, RhoGAP54D RNAi 54051>*actin-Gal4* RNAi line showed a significant movement decrease and indeed when sleep metrics were analysed through the flies' immobility and further response to lights off, the experimental flies showed significantly reduced startle response (fig.5.18 B&C). These results may suggest that the RNAi line is likely working and resulting in reduced expression of *RhoGAP54D* globally. Conversely, as *actin*-Gal4 is a global driver and no effect was observed for two RNAi lines, it might suggest that the RNAi lines were inefficient/non-functional, or the phenotype observed in RhoGAP54D RNAi 54051>*actin-Gal4* line is caused by an off-target effect. Indeed, the RNAi lines have not been previously characterised, therefore I performed qPCR to investigate the efficiency of the lines.

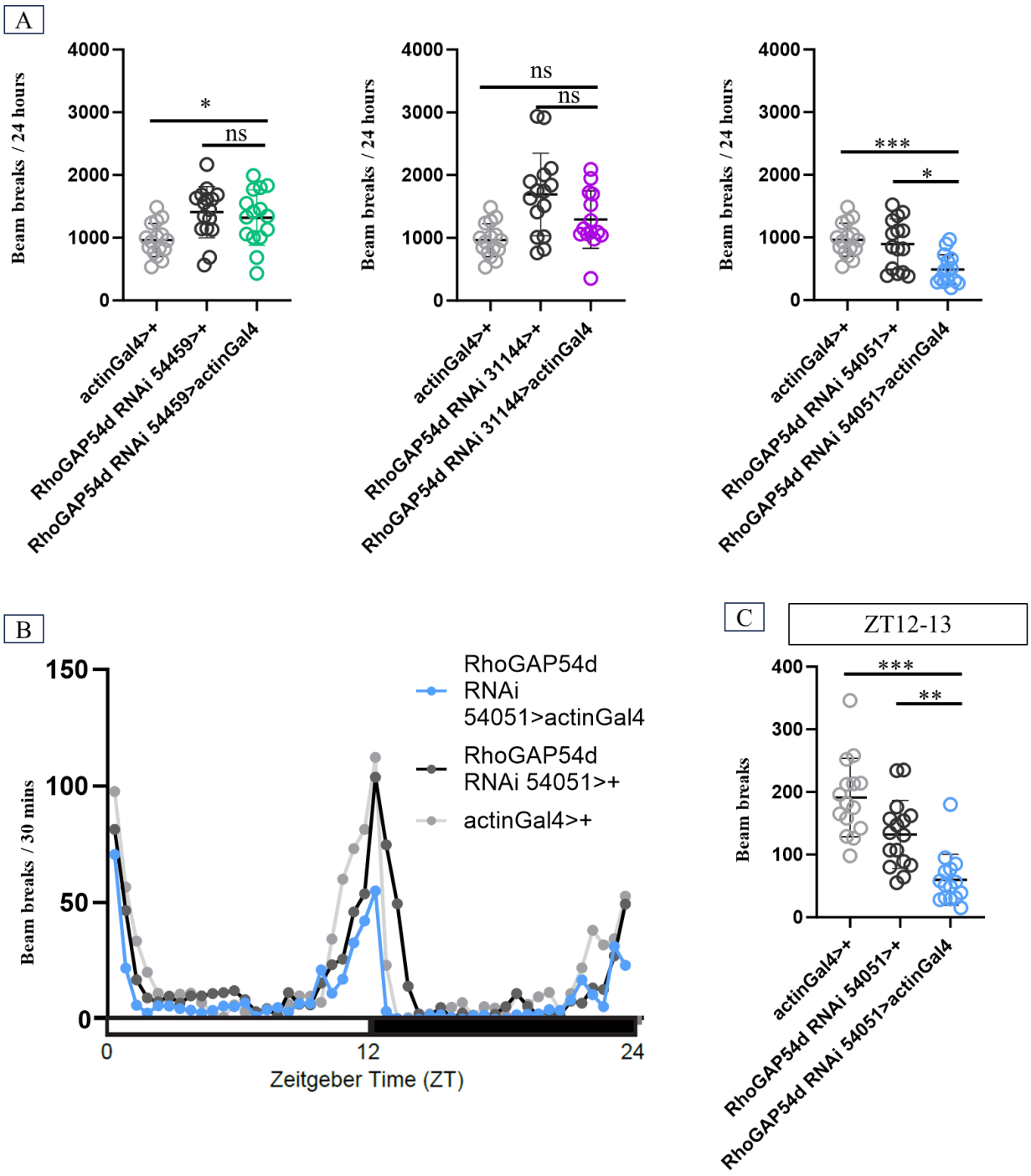


Figure 5.18 RhoGAP54D Drosophila activity monitor. A) total locomotion of adult fly in 24 hours, only RhoGAP54D RNAi 54051>actin-Gal4 displays a movement phenotype B) sleep and wake periods of RhoGAP54D RNAi 54051>actin-Gal4 and controls in 24 hours. C) Significant locomotion decrease at period immediately after lights off (ZT12-13). Data are shown as means ± SD. One-way ANOVA, * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$, ns = non-significant.

5.3.2.2 Quantitative PCR for efficacy of RNAi

qPCR was performed on RNA extracted from correct progeny of available RNAi crosses with *actin-Gal4* and *mcherry* RNAi crossed with *actin-Gal4* was used as the control (fig.5.19).

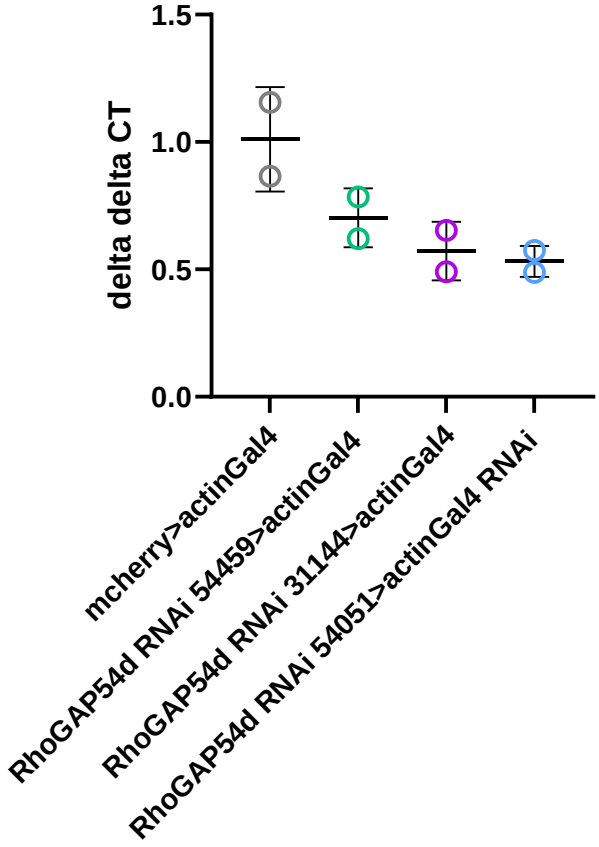


Figure 5.19 Quantitative PCR results for RhoGAP54D RNAi lines used in this project. Highest degree of knockdown is seen in RhoGAP54D RNAi 54051>*actin-Gal4. Data are shown as means ± SD.*

The efficiency of the RNAi lines appears to vary, with RhoGAP54D RNAi 54459>*actin-Gal4* producing about 30% of knockdown. The RNAi line that resulted in locomotion phenotype when expressed globally, RhoGAP54D RNAi 54051>*actin-Gal4*, appeared to have the highest percentage of knockdown at about 50 percent. However, RhoGAP54D RNAi 31144>*actin-Gal4* has a similar level of mRNA knockdown. It is possible that a certain degree of knockdown needs to be reached for a fly to display a phenotype, indeed the disease-causing variants in human patients are found on both alleles rather than in heterozygous state. It is also possible that DAM is not a sensitive enough method to detect more subtle phenotypes. However, another explanation for a movement phenotype in only one RNAi line is that the phenotype is linked to an off-target effect of the RNAi line used.

Therefore, to validate the model and loss of function of the protein in the fly, we established a collaboration with Yohanns Bellaiche who tagged all fly GAP proteins with GFP and agreed to share their RhoGAP54D:GFP fly line. Additionally, a RhoGAP54D null line was a kind gift from the group and also used for validation of the models in this study.

5.3.2.3 A global knockdown of RhoGAP54D using degrad:FP system results in a movement phenotype at 3-7days

To complement the data of RNAi KD model, we utilised degradFP, a genetic system that promotes degradation of GFP-tagged fusion proteins via the ubiquitin pathway therefore avoiding the potential off-target effects of shRNA KD. This KD system is harder to achieve as it requires expression of the GFP on both alleles in the experimental flies. After successful ubiquitous expression of degradFP components in a background homozygote for a RhoGAP54D::GFP knock-in allele, we observed reduced overall and peak movement relative to RhoGAP54D::GFP homozygote controls (fig.5.20). The KD flies showed reduced startle response (fig.5.20A) to lights-off and reduced locomotion in 24hours (fig.5.20B) as well as in one hour after lights-off (ZT12-13, fig.5.20C).

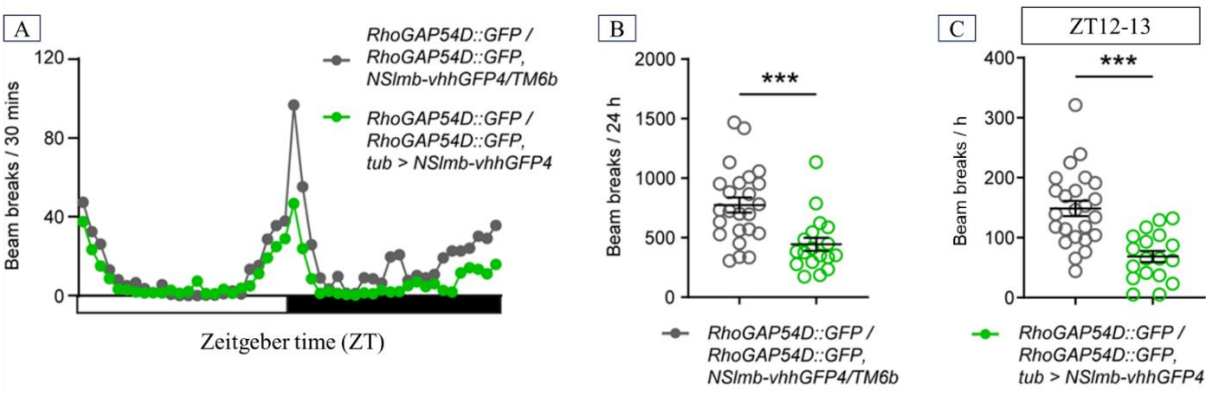


Figure 5.20 RhoGAP54D Drosophila activity monitor. A) sleep and wake periods of RhoGAP54D GFP KD and control in 24 hours B) total locomotion of adult fly in 24 hours C) Significant locomotion decrease at period immediately after lights off (ZT12-13). Data are shown as means ± SEM. One-way ANOVA, *** = p<0.001.

5.3.2.4 A knock out of RhoGAP54D using CRISPR/Cas9 null lines results in movement phenotype at 3-7 days

As a final confirmation of the results of RNAi and degradFP KD models, we subjected to behavioural studies, a *RhoGAP54D* null allele (*RhoGAP54D*^{KO}) generated through CRISPR/Cas9 gene editing by Yohanns Bellaiche, France. These flies express no *RhoGAP54D* therefore avoiding potential off-target effects of other models. In addition, since human patients with the variants in *ARHGAP19* carry the variants in homozygous state, the *RhoGAP54D* knock-out model best represents the loss-of-function of the gene products.

Comparison of *RhoGAP54D*^{KO} heterozygote and homozygote flies again revealed reduced overall and peak movement in *RhoGAP54D*^{KO} homozygotes relative to heterozygote controls (fig.5.21). The KO flies showed reduced startle response (fig.5.21A) to lights-off and reduced locomotion in 24hours (fig.5.21B) as well as in one hour after lights-off (ZT12-13, fig.5.21C).

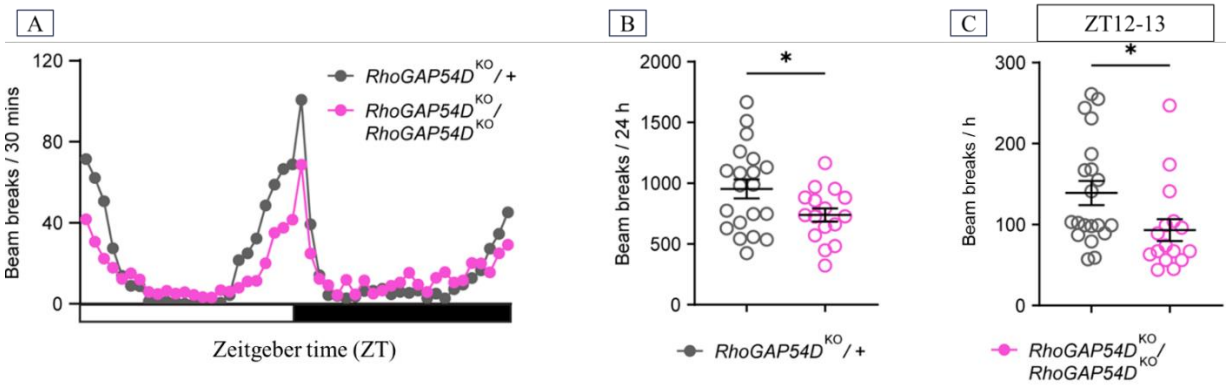


Figure 5.21 RhoGAP54D Drosophila activity monitor. A) sleep and wake periods of RhoGAP54D GFP KO and control in 24 hours B) total locomotion of adult fly in 24 hours C) Significant locomotion decrease at period immediately after lights off (ZT12-13). Data are shown as means ± SEM. One-way ANOVA, * = $p<0.05$.

5.3.2.5 Immunostaining suggests RhoGAP54D is expressed in a subset of glial cells

Expression of *RhoGAP54D* in fly is not well studied therefore we examined a reporter to determine where the gene is expressed in the adult fly brain. A CRIMIC line, expressing GAL4 in the endogenous pattern of *RhoGAP54D* was crossed with UAS-TdTomato to be able to visualise the endogenous *RhoGAP54D* expression pattern in the progeny. The adult fly brain and thoracic ganglion were dissected and visualised under confocal microscope.

The pattern of tdTomato fluorescence suggests that RhoGAP54D is expressed in subperineural and perineural glia in the adult fly brain (fig.5.22).

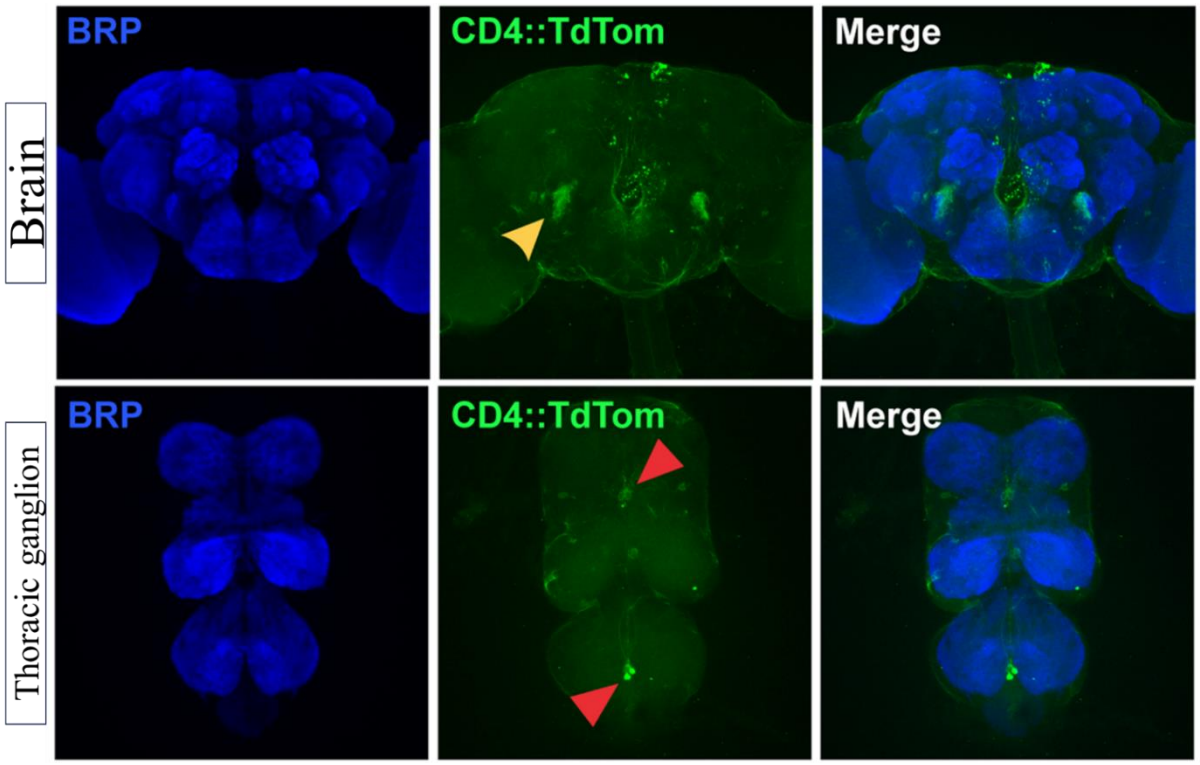


Figure 5.22 RhoGAP54D localisation in adult fly brain. Confocal images illustrating *RhoGAP54D*-driven membrane-tagged CD4::TdTomato expression in the adult male *Drosophila* brain and Ventral Nerve Cord (VNC). Yellow arrow points to projections close to the antennal mechanosensory motor centre. Red arrows point to sparse cell bodies in the VNC.

5.3.2.6 Conclusion to RhoGAP54D results

Through the use of three independent RhoGAP54D RNA interference lines controlled by Gal4-UAS system and expressed globally with *actin-Gal4* driver, I established one RhoGAP54D knockdown model and used qPCR to validate its efficacy to about 50% knockdown in adult head. In this knockdown model, we observed reduced locomotion ability of 3-7 day old fly using the *Drosophila* Activity Monitor, which recapitulated *ARHGAP19* patient movement phenotype.

Because 2 other RNAi lines tested resulted in no phenotypic changed when expressed globally, we sought to further validate this model. Using *Drosophila* activity monitor we observed significant movement decrease in both RhoGAP54D protein knockdown and *RhoGAP54D* knock-out flies compared to their respective controls.

5.4 Discussion

Drosophila melanogaster has served as an invaluable model organism for over a hundred years and during my thesis in collaboration with Prof James Jepson at UCL, I established *Drosophila* loss-of-function models for fly orthologs of human *RFC1* and *ARHGAP19*. These flies not only recapitulated the patient locomotion phenotypes but also allowed us to gain understanding of the protein localisation in the case of *RhoGAP54D* (*ARHGAP19* ortholog) and lifespan prediction in the case of *Gnf1* (*RFC1* ortholog). The fly models established in this thesis have a wider impact on the scientific community by helping to elucidate the mechanisms of two neurological diseases and opening up the possibility to further research using these models.

Below I discuss the results of the study as well as the limitations and possible future lines of investigation arising from the data collected.

5.4.1 Gnf1

To model loss-of-function of Gnf1, I used the RNA interference to knockdown expression of Gnf1 in neuronal tissue using *nsyb*-Gal4 driver at different ages of the fly. Interestingly, only older Gnf1 KD flies (40 day old) displayed perturbed locomotion compared to controls. This finding is in line with late onset of the human RFC1 disease where median age of onset is 54 years (Curro *et al*., 2023).

Since human RFC1 and fly Gnf1 are involved in DNA damage and repair (Liu *et al*., 2022; Tsuchiya *et al*., 2007), I investigated whether the knockdown fly experiences heightened DNA damage as opposed to the relevant controls. RFC1 is highly conserved throughout species and is specifically involved in binding gapped or nicked DNA and to date its dysfunction has mostly been implicated in various human cancers (Li *et al*., 2018) and in Hutchinson-Gilford progeria syndrome (HGPS) (Tang *et al*., 2012). Interestingly, in HGSP which is a genetic disease-causing premature aging, RFC1 has been found to be truncated and defective at loading proliferating cell nuclear antigen (PCNA) and pol δ onto DNA for replication. In the fly, we specifically interrogated non-replicating post-mitotic neurons for the consequences of Gnf1 KD. For this, we used an antibody to histone 2A variant (H2Av) which in fly combines the function of H2Ax and H2Az in humans. The antibody recognises the variant histone that is associated with DNA double strand breaks (Lake *et al*., 2013). We investigated the DNA damage at 40 days – thus age matched to the fly which showed movement phenotype. We observed significant H2Av staining increase in experimental fly at 40 days old. This could suggest that Gnf1 knockdown contributes to accumulation of DNA damage in neurons which cannot readily be repaired. Interestingly, interrogating the lifespan data, we observed that the

knockdown fly has a reduced lifespan and begins dying off at around 40 days old. There is a possibility therefore that knocking down Gnf1 may lead to significantly reduced ability of the fly to repair DNA damage which can in turn exacerbate death events, a phenotype also observed in HGPS but whether it is a direct consequence of RFC1 loss-of-function remains to be elucidated.

A substantial loss of Purkinje fibres has been observed in post-mortem brains of CANVAS patients, most severe in vermis (Cortese *et al*., 2019). This is in line with established knowledge that ataxia is marked with neuronal loss in cerebellum. Further to this, previous evidence from CAG repeat expansion disorders such as Huntingdon's disease and SCAs showed that cerebellum exhibits a high expression of DNA damage response genes (Kacher *et al*., 2024) where the CAG expansions might be controlled by these DNA repair mechanisms. In ataxia telangiectasia (A-T), loss-of-function mutations in a DNA repair gene ATM have been shown to cause cerebellar neurodegeneration leading to patient locomotor defects (Deacon *et al*., 2024). These highlight that cerebellum is particularly vulnerable to dysfunction in DNA damage repair machinery, and the data from our Gnf1 KD *Drosophila* model suggest that loss-of-function of Gnf1 in post-mitotic neurons may cause DNA damage accumulation leading to progressive neurological defects that impair motor function. However, whether DNA damage accumulation is observed in CANVAS patients, remains to be investigated.

Following on the results of DNA damage accumulation, we decided to stress the system further by using a known DNA damaging agent – cisplatin. Cisplatin treatment had detrimental effect on fly survival, this was true both for the experimental as well as the control flies. However, the experimental fly showed significantly reduced lifespan as opposed to the controls and only survived to about 20 days. With this in mind I used 12-day old flies for the DAM, age just when the flies started dying off. Surprisingly, after interrogating the DAM data from two replicated experiments, I observed hyperactivity in the experimental fly. It is of note that control flies appeared static when observed with naked eye and they showed increased grooming. It is possible that the experimental fly showed hyperactivity related to approaching death event. We also stained the fly brains treated with cisplatin with H2Av antibody to investigate any exacerbated DNA damage due to the damaging agent. We observed significant increase ($p<0.05$) in H2Av staining in the KD fly as opposed to the control. These data suggest that loss of Gnf1 might impair the ability of *Drosophila* to counteract neuronal DNA damage and in the context of reduced Gnf1 function, acute induction of DNA damage specifically perturbs the activity of pre-motor centres, and potentially sensory pathways, that collectively sculpt the

normal crepuscular pattern of locomotor activity exhibited by flies in oscillating light-dark conditions. It is not currently known whether circadian defects have been observed in CANVAS patients, however, sensory symptoms are well documented (Cortese *et al*., 2019; Curro *et al*., 2023).

5.4.1.1 Limitations and future horizons

The main limitation of the work described in this chapter is that the locomotion and lifespan phenotype was only observed in one RNAi model system. RNA interference relies on a complementary short harpin RNA (shRNA) that induces cleavage of the corresponding host mRNA. The shRNA, however, might in some instances bind to a different than expected target and cause a phenotype by an off-target effect. Therefore, a future investigation will seek to validate the established knockdown line by means of knocking down the expression of Gnf1 further concurrently with a heterozygous Gnf1 null allele. Another validation of the KD model described herein could seek alternative Gnf1 RNAi lines or neuron-specific CRISPR?Cas9 knock-out.

Moreover, our investigations for Gnf1 function explored loss-of-function mechanism only by knocking down the expression of the gene. In human patients, the disease is caused by biallelic repeat expansions in the second intronic region of *RFC1* gene (Cortese *et al*., 2019). Whilst some evidence stated in the introduction points towards the loss-of-function mechanism of the gene, there is no conclusive evidence that gain of function mechanism may not be present in the disease. Indeed, a recent study submitted in a manuscript depository (Todd *et al*., 2024) suggests that gain-of-function mechanism might be present in CANVAS disease as presence of repeat peptides was detected in two post-mortem brains of patients. Whilst this evidence is not peer-reviewed, it needs to be considered and establishing another *Drosophila melanogaster* model where repeat expansions in Gnf1 are expressed could shed light into these repeats and whether they would cause a phenotype in the fly model, and importantly whether the phenotype would be independent of Gnf1 gene function itself.

5.4.2 RhoGAP54D

In *Drosophila*, similarly to humans and indeed in all eukaryotes, due to their modulation of RhoGTPases, RhoGAPs and GEFs play important roles in cytoskeleton organisation, cell cycle control, cell division, and migration. The Rho/ROCK pathway (described in chapter 4) is an essential pathway, that has been extensively studied, with many GEFs and GAPs described and characterised. We found ourselves in a unique position of, for the first time, having characterised patients with neuropathy who carry biallelic mutations in *ARHGAP19* thus allowing us to form hypotheses about the function of the gene some of which could be tested in *Drosophila melanogaster*.

Characterisation of GAPs and GEFs is non-trivial, with as many as 80 having been discovered in humans thus far. These take part in complex interplay not only between themselves, other proteins, but also in spatial and temporal manner (DeGeer and Lamarche-Vane., 2013). *Drosophila melanogaster* allowed for modelling loss of function of the *ARHGAP19* gene ortholog, *RhoGAP54D*, and visualising the consequences it may have on cellular and multi-system level. Here, I successfully established two global RhoGAP54D KD models (using RNAi and degrad:FP) and a KO model using CRISP/Cas9 all with locomotion phenotypes matching this of human patients. The experimental flies had a reduction of total movement at 3-7 days of age – and indeed human patients show movement phenotype in the first decades of life (Dominik *et al*., 2024). Interestingly, ability to establish global knockouts of RhoGAP54d suggests that the protein is not essential for survival in the fly or other proteins might have a compensatory role in those flies which remains to be investigated.

Collectively, these data demonstrate that RhoGAP54D promotes robust locomotor activity in *Drosophila*, supporting the genetic link between variants in the human ortholog *ARHGAP19* and disrupted movement.

5.4.2.1 Expression of RhoGAP54D

Establishing expression of protein is an important part of elucidating its function. We leveraged a reporter line to express a fluorescent protein in endogenous RhoGAP54D expression pattern and subsequent analysis revealed that the protein is expressed in subperineural and perineural glia around the central brain. These have role in blood brain barrier, and it is plausible that loss of function of RhoGAP54D could impair integrity of blood brain barrier and in turn cause detrimental effects on processes such as neuronal excitability or viability. We were unable to investigate the protein expression further in periphery, but previous evidence indicates that the protein is expressed in fly leg (Greenberg and Hatini, 2011) which is in line with expression in

humans where ARHGAP19 is most highly expressed in tibial nerve (GTEx Project). All these data have great impact on the field of neuroscience, translating to human data where ARHGAP19 is expressed in periphery and locomotion phenotype and foot deformities are observed in patients, but also the protein is expressed in lower extend it glial cells, implications of which are not yet known.

5.4.2.2 Limitations and future horizons

All the behavioural studies for the RhoGAP54D KD and KO models were performed using *Drosophila* Activity Monitor (DAM). The DAM is a robust system allowing for concurrent measurement of total activity of a large number of animals and extrapolating data such as for sleep and wake periods and activity linked to oscillating light-dark conditions. However, the DAM does not allow for monitoring more subtle movements such as leg jerks and alternatives such as Drosophila ARousal Tracking system (DART) could be used (Faville *et al*., 2015).

This would be of particular interest as the patients with biallelic pathogenic *ARHGAP19* mutations suffer from movement difficulties that are often on mild to moderate spectrum and therefore some of the movement behaviours might have been missed in the fly.

Our patient cohort displays biallelic mutations in various locations in the protein therefore we chose not to model any particular mutation but rather to use knock down and knock out models of the gene products. Nevertheless, we show that loss of function of RhoGAP54D recapitulates the patient phenotype providing further evidence to patient genotype phenotype observations.

Moreover, it is important note that although the neuropathy caused by biallelic mutations in *ARHGAP19* is predominantly axonal, some patients had demyelinating neuropathy and in the context of modelling demyelinating neuropathies caution should be taken when using *Drosophila* as a model organism as *Drosophila melanogaster* does not form myelin sheaths. However, it was not an aim of the study described in this thesis to provide evidence of axonal or demyelination changes but rather to investigate whether the loss of function of *ARHGAP19* and fly ortholog RhoGAP54D will cause phenotypic changes in the models explored.

Future lines of investigation in the *Drosophila melanogaster* KO and KD models of RhoGAP54D may include studying the models at earlier developmental stages such as 3rd instar larvae to give insight into any developmental consequences of loss-of-function of *RhoGAP54D*. Moreover, one of the key points to address would be to define the cell-types where *RhoGAP54D* knockdown causes locomotion phenotype and a range of tissue-specific drivers such as neuronal (*nsyb*-Gal4), glial (*repo*-Gal4) or neuroblast (*wor*-Gal4) or many

others could be employed. In addition, since we describe this gene for the first time as neuropathy causing, aging the fly models could show whether the phenotype may be progressive. Importantly, as mentioned in previous chapter 4, Rho/ROCK pathway is regulated by GAPs and GEFs and screening for any cognate GEF proteins to RhoGAP54D could be performed to elucidate their interactions and possible phenotypic rescue mechanism. Lastly, since Rho/ROCK pathway is well described, many small molecular drugs are available for inhibiting the pathway and these could be administered orally to the fly models, either acutely or chronically, to assess whether the locomotion phenotype could be rescued by these drugs.

5.5 Conclusions

Genetic knockdown using RNA interference (RNAi) is a robust method for revealing function of many genes in the genome, however, it may be insufficient. Indeed, RNAi may have off-target effect or gene expression may remain high even in presence of RNA interference. Therefore, even though RNAi remain one of the first and more common tools in gene knockdown, often, the models need to be validated. In many cases, if by using 2 or more RNAi lines against the given gene produce phenotypic changes, it is considered valid. However, in this thesis, only one RNAi line per gene showed a relevant phenotype and other validation methods needed to be employed. Even with these limitations, *Drosophila melanogaster* proved to be an invaluable model organism for our purposes with successful modelling of loss of function for both RhoGAP54D and Gnf1. These add an important contribution to our understanding of function of *ARHGAP19* and *RFC1* and they complement data from human studies and other models. *Drosophila melanogaster* has also allowed us to research data that would be difficult to collect in human subjects, for example, examining the impact of Gnf1 knockdown on lifespan.

It is important to note that in both genes studied, other methods of knockdown could have been used such as for example CRISPR/Cas9 in the Gnf1 work; or point mutations could have been introduced in the RhoGAP54D fly to elucidate their consequence. Similarly, looking at Gnf1 fly, a pathogenic repeat expansion seen in *RFC1* disease spectrum could have been introduced to study its effect. These methods could be used in possible further *Drosophila* modelling and the successfully established models may pave a way to further understanding of the function of *RFC1* and *ARHGAP19* which can benefit the patients and the scientific community.

CHAPTER 6. General conclusions

In this PhD thesis, I have used a combination of high-throughput genetics and molecular biology techniques to advance understanding and disease characterisation of rare neurological disorders.

I contributed to the work of Dr Andrea Cortese on Cerebellar Ataxia with Neuropathy and Vestibular Areflexia Syndrome (CANVAS) and the disease spectrum caused by repeat expansions in *RFC1* gene (Cortese *et al*., 2019). My work has helped to better characterise *RFC1* repeat expansion disorder where we further described the genetic heterogeneity underlying this disorder and we discovered novel pathogenic repeat expansions in the *RFC1* gene causing CANVAS (Dominik *et al*., 2023). Furthermore, we streamlined the screening workflow for the *RFC1* repeat expansions which since has been implemented in the NHS diagnostic settings. I described the process of optimisation of Southern blotting which had further advantages to our work on sizing *RFC1* repeat expansions and correlating the sizes with the disease onset, severity and progression which can in turn help to better counsel the patients. Additionally, we successfully implemented novel technology of analysing structural variation in the genome, Bionano Optical Genome Mapping in Queens Square Institute of Neurology, and we validated the method for repeat expansion testing by comparing to the traditional Southern blotting method (Facchini*, Dominik* *et al*., 2023). Importantly, we showed that even thought at discovery they were thought to be rare, repeat expansion in *RFC1* are a common cause of ataxia and sensory neuropathy, and growing evidence shows that *RFC1* disease spectrum may be underdiagnosed and indeed the carrier frequency of the pathogenic expansions in the healthy populations reaches up to 7% in current literature (Davies *et al*., 2022).

I also contributed to genetic and functional characterisation of a novel neuropathy gene, *ARHGAP19*, discovered by Dr Stephanie Efthymiou and Professor Henry Houlden. In the gene discovery chapter, I demonstrated the importance of independent validation of suspected pathogenic variants by Sanger sequencing for segregation within the families and affected individuals as well as functional validation for characterisation and interpretation of the consequences of the variants on the gene function. Additionally, I highlighted the importance of international collaborations that can aid discovery of new genes causing rare neurological conditions. Whole exome and whole genome sequencing together with deep phenotyping of

affected individuals are now well-established methods of gene discovery and these methods have and will continue to accelerate discovery of disease associated genes.

Importantly, I used *Drosophila melanogaster* in Professor James Jepson Lab at UCL as a model organism for both the genes described in this thesis – *RFC1* and *ARHGAP19*. I successfully established loss-of-function *Drosophila* models which recapitulate patient locomotion phenotypes and have further helped to elucidate the mechanisms of the gene dysfunctions in their respective diseases of CANVAS and CMT. The established animal models can have further impact on the study of the diseases with future work planned beyond the scope of this thesis.

6.1 *RFC1* repeat expansion sizing and disease correlations

In chapter 2, I described our work on Cerebellar Ataxia with Neuropathy and Vestibular Areflexia syndrome (CANVAS). The genetic cause of CANVAS was first described by Dr Cortese in 2019 as biallelic repeat expansions of AAGGG repeat motif in second intron of Replication Factor C subunit 1 (*RFC1*) gene. This repeat expansion differs from the normal reference allele not only in size, but also in the repeat sequence – the reference allele contains the repeat motif $AAAAG_{11}$.

Since the initial discovery, much attention has been given to CANVAS by the scientific community culminating in publications screening various populations (Scriba *et al*., 2020; Tsuchiya *et al*., 2020; Nakamura *et al*., 2020; Rafehi *et al*., 2019) and conclusions by us and others that this disease is underdiagnosed, and the carrier frequency is quite high (Davies *et al*., 2022).

In this thesis, as part of the wider study involving multicentre cohorts, I contributed to the exploration of the relationship between the size of AAGGG repeat and the age of onset and severity of the disease (Curro *et al*., 2023). It is well documented that other neurological diseases caused by repeat expansions such as for example Friedreich Ataxia or Huntington's disease are inversely correlated to the size of the repeats (Filla *et al*., 1996; Santoro *et al*., 2000; Gardiner *et al*., 2017) – the larger the repeat size, the lower the age of onset of the disease and the more severe the disease. Therefore, we aimed to measure the repeat sizes of a large cohort of individuals presenting with at least one of core CANVAS phenotypes, with PCR screening indicating likely positive biallelic AAGGG expansions.

Firstly, however, for this work, I explored the molecular biology screening techniques for *RFC1* repeat expansions and optimised the gold standard technique for measuring repeat expansions

– Southern blotting. Southern blotting is a cumbersome technique used for measuring sizes of repeat expansions and it relies on large quantities (over 5µg) of good quality DNA (Dominik *et al*., 2020). In this chapter, I described methods of optimising the Southern blotting which resulted in better quality images and improvement to the transfer of the DNA onto positively charged membrane used. This work increased our ability to measure the sizes of repeat expansions and we further correlated the sizes of AAGGG expansions with clinical variables. Here, we found that patients with isolated sensory neuropathy had smaller expansions compared to the other core CANVAS phenotypes and there was no significant difference of expansion sizes between patients with complex neuropathy and full-blown CANVAS. These findings suggests that the repeat expansion size can act as a modifier of the disease phenotype and sensory neurons are probably more susceptible to the AAGGG repeat expansion than other tissues. The repeat expansion size in *RFC1* also influences the onset of neurological symptoms and we found that the larger the repeat expansions is, the younger the age of onset is, especially when considering the smaller allele (Curro *et al*., 2023).

Finally, I helped to establish and validate a novel technology – Optical Genome Mapping (OGM) which is capable of analysing different types of structural variation within genome, including repeat expansions. Here, we analysed the sizes of expansions in 17 patients using both the gold standard Southern blotting and the novel OGM technology. We showed a very good linear correlation of the two techniques, however, OGM tended to better resolve alleles of similar sizes. Moreover, a distinct advantage of OGM is that this technology allows for analysis of the structural variants in the entire patient genome whereas Southern blotting is locus specific and dependent on the probe used (Facchini*, Dominik* *et al*., 2023).

This work has several limitations thoroughly described in chapter 2 which range from retrospective nature of the study to unavailability of large quantities of patient DNA for some of the patients. Regardless of the limitation, this work is highly impactful as for the first time we correlated the size of the repeat expansion with the disease onset and its severity which can help better counsel the patients. This is also the biggest cohort of *RFC1* patients tested so far which yet again highlights the importance of multicentre collaborations. Importantly, we also optimised the screening procedure for *RFC1* repeat expansions, mainly focusing on the Southern blotting methodology and we implemented a novel technology of Optical Genome Mapping that is likely to replace Southern blotting in the future.

6.2 Genetic heterogeneity of *RFC1* and discovery of novel pathogenic repeat expansion motifs

In chapter 3, I follow on our research from chapter 2 in *RFC1* by exploring the underlying genetic heterogeneity associated with RFC1 disease spectrum. A majority of the research into *RFC1* repeat expansions had been conducted on individuals of European ancestry (Cortese *et al*., 2019, Rafehi *et al*., 2019, Cortese *et al*., 2020). However, evidence from different populations such as Asia Pacific and Maori (Beecroft *et al*., 2020), had revealed that other pathogenic motifs exist apart from the most common AAGGG expansion motif. Indeed, between 3-18 % of individuals with clinical CANVAS do not test positive for biallelic AAGGG expansions (Cortese *et al*., 2019; Rafehi *et al*., 2019; Ronco *et al*., 2023). In this study we leveraged short read WGS from the Genomics England sequencing project to investigate the normal and pathologic variation of the *RFC1* repeat expansions and to identify additional pathogenic repeat configurations in *RFC1* causing CANVAS and disease spectrum. We discovered three novel pathogenic repeat configuration motifs that cause CANVAS: AGAGG, AGGGC and AAGGC either in homozygous or compound heterozygous state with the common pathogenic AAGGG repeat expansion. We further explored the full sequences of the motifs by means of long read sequencing, using both available technologies of Oxford Nanopore and Pacific Biosciences and concluded that long read sequencing is still challenging in highly repetitive GC-rich sequences. Very importantly, we observed that a previously thought non-pathogenic repeat motif AAAGG can become pathogenic when sufficiently expanded and in compound heterozygous state with AAGGG as evidenced by a significant enrichment of frequency of this motif in ataxia cohort of Genomics England as opposed to the non-neurological controls and further by Southern blotting size exploration which revealed that patients with this motif had neurological symptoms if the expansion size was larger than 600 repeats. However, it cannot be excluded that patients with smaller sizes of AAAGG motif than 600 repeats may develop CANVAS later in life (Dominik *et al*., 2023).

Further, because the standard PCR screening methodology is motif specific and therefore may yield false negative results in patients suspected of CANVAS if screening for AAGGG only is performed, we explored whether Oxford Nanopore sequencing using a barcoded PCR product for targeting the specific *RFC1* locus may be used to give an indication of a repeat motif present in the sample. However, at this time this method resulted in sequences with a very high error rate where motifs cannot be concluded.

Finally, I showed an algorithm for an expansion of screening methodology where clinical CANVAS is suspected but PCR suggests negative result for AAGGG expansion, or the expansion is detected in one allele only.

The above have implications for diagnostic testing where caution must be taken with interpretation of negative AAGGG screening if clinical CANVAS is suspected.

6.3 Novel gene discovery

In chapter 4, I described for the first time the role, function and potential mechanism of *ARHGAP19*, as a novel disease-causing gene for CMT. This discovery and functional validation were possible due to the integration of basic research in the field of high-throughput genome analysis with biochemistry, cell biology and animal models via the establishment of strong collaborations led by Dr Stephanie Efthymiou and Professor Henry Houlden.

In order to first identify the disease-causing gene, we initially performed whole exome sequencing in the neuropathy affected cases recruited by our lab and filtered the results for novel and very rare variants with high impact in-silico scores on protein function. Further using our established collaboration links, notably SYNAPS and ICGNMD, as well as Gene Matcher and connections with other researchers, we identified a large cohort of 25 patients from 20 families, with biallelic mutations in *ARHGAP19* with weakness in lower limbs, foot deformities and mixed axonal or demyelinating neuropathy (Dominik*, Efthymiou* *et al*., 2024).

To study the impact of these variants at cellular and organism levels, we employed in silico methods to predict the variant consequences, in vitro activity assay to assess the consequences of the variants in the functional domain and in vivo animal models to interrogate the predicted loss-of-function mechanism.

ARHGAP19 is a GTPase activating protein (GAP) acting in RhoA/ROCK pathway where is stimulates low intrinsic GTPase activity of RhoA therefore negatively regulating the pathway (David *et al*., 2014). In our cohort, we collected patients with missense and nonsense biallelic variants, and we predicted steric hindrance affecting the protein folding using AlphaFold. Further, we investigated the GAP activity of 3 chosen variants lying inside the functional GAP domain and we observed that the activity is abrogated. We also modelled the loss-of-function of the gene products as the mechanism of the disease using two animal models – *Drosophila melanogaster* and *Danio rerio*. The knockdown and knockout models of the respective gene

orthologs in these animals recapitulated the locomotion phenotype seen in the patients providing further evidence for the genotype phenotype correlation in the patients.

In vitro migration assays showed significantly reduced migration of the patient derived fibroblasts as compared to non-neurological controls however, no significant differences were observed in the *ARHGAP19* mRNA levels. Interestingly, we observed significant ($p<0.05$) reduction of ARHGAP19 in patient derived iPSC motor neurons as opposed to healthy controls, but no significant protein level changes in fibroblasts. This could suggest that the loss-of-function of ARHGAP19 may have more robust consequences in motor neurons rather than in other tissues.

Our findings do not fully elucidate the mechanism of axonal damage caused by the *ARHGAP19* mutations but nevertheless provide genetic and functional evidence for *ARHGAP19* to be added to the growing list of CMT-causing genes.

6.4 *Drosophila melanogaster* as a model organism

In chapter 5, I described how *Drosophila melanogaster* can aid research into neurogenetic conditions and in collaboration with Professor James Jepson at UCL, I used the animal to model the loss-of-function mechanisms of the fly orthologs of human *RFC1* and *ARHGAP19*.

To first investigate the consequences of loss-of-function of the gene products in *Drosophila*, we used RNA interference to knock out the expression of *Gnf1* (fly ortholog of *RFC1*) in post-mitotic neurons and *RhoGAP54D* (fly ortholog of *ARHGAP19*) in global pattern and used *Drosophila* Activity Monitor to assess the locomotion activity of the resulting flies as opposed to their respective controls. RNAi are quick and easy methods of knocking down expression of a gene at an RNA level, however, off-target effects are not uncommon, and it is commonly anticipated that more than one RNAi line should produce flies with a phenotype to deem it suitable model.

I successfully established a knockdown model for *Gnf1* and a knockdown model for *RhoGAP54D* which recapitulated the patient locomotor phenotype. I further validated the knockdown levels by means of quantitative PCR. Moreover as only one of the three tested RNAi lines showed a phenotype for RhoGAP54D knockdown model, we established a collaboration to obtain two further fly lines for validation of our fly model– a Green Fluorescent Tagged (GFP) RhoGAP54D knock in line which we used to knockout the expression of RhoGAP54D by GFP degradation which uses ubiquitin proteasome pathway; and a biallelic

null knockout line of RhoGAP54D; both of which were then used in locomotion assays and showed similar phenotype to the RNAi knockdown line. These data suggest that *RhoGAP54D* promotes robust locomotor activity in *Drosophila*, supporting the genetic link between mutations in the human ortholog *ARHGAP19* and disrupted movement.

For Gnf1, we successfully established one RNAi knockdown model in post-mitotic neurons which recapitulated key locomotion phenotype linked to ataxia in patients with CANVAS. Interestingly, this phenotype was observed when the fly was aged to 40 days but not in the younger 7 day and 21-day old fly which is in line with the older age of onset of walking difficulties in human patients.

Since human *RFC1* and fly Gnf1 both have an important role in DNA damage response, we measured the accrued H2Av levels which indicate DNA damage in the knockdown and control flies and observed significant increase in DNA damage at 40 days. Following this data, I administered a known DNA damage agent to the KD and control flies immediately after eclosion and observed that these flies have an advanced evening anticipation and reduced startle response measured in the *Drosophila* Activity Monitor which suggest that acute DNA damage in reduced Gnf1 background may have an effect on sensory pathways.

Taken together, these data highlight the validity of *Drosophila melanogaster* as a model organism in neurogenetic research and pave a way for future line of investigation using the fly which may include a search for therapeutic avenues for the two diseases described herein.

6.5 Final remarks

The research projects included in this thesis emphasise the need for international collaborations to discover novel genes and better characterise the known disease-causing genes. Inclusion of affected individuals from all around the world can lead to increased genetic diversity in the cohorts and provide better understanding of underlying genetic causes of the disease. Furthermore, the advancements in sequencing and genomic mapping technologies have allowed for increased discovery of disease-causing genes and hold great promise for diagnostics and management of patients with rare diseases.

References

ADAMS, M. D., CELNIKER, S. E., HOLT, R. A., EVANS, C. A., GOCAYNE, J. D., AMANATIDES, P. G., SCHERER, S. E., LI, P. W., HOSKINS, R. A., GALLE, R. F., GEORGE, R. A., LEWIS, S. E., RICHARDS, S., ASHBURNER, M., HENDERSON, S. N., SUTTON, G. G., WORTMAN, J. R., YANDELL, M. D., ZHANG, Q., CHEN, L. X., BRANDON, R. C., ROGERS, Y. H., BLAZEJ, R. G., CHAMPE, M., PFEIFFER, B. D., WAN, K. H., DOYLE, C., BAXTER, E. G., HELT, G., NELSON, C. R., GABOR, G. L., ABRIL, J. F., AGBAYANI, A., AN, H. J., ANDREWS-PFANNKOCH, C., BALDWIN, D., BALLEW, R. M., BASU, A., BAXENDALE, J., BAYRAKTAROGLU, L., BEASLEY, E. M., BEESON, K. Y., BENOS, P. V., BERMAN, B. P., BHANDARI, D., BOLSHAKOV, S., BORKOVA, D., BOTCHAN, M. R., BOUCK, J., BROKSTEIN, P., BROTTIER, P., BURTIS, K. C., BUSAM, D. A., BUTLER, H., CADIEU, E., CENTER, A., CHANDRA, I., CHERRY, J. M., CAWLEY, S., DAHLKE, C., DAVENPORT, L. B., DAVIES, P., DE PABLOS, B., DELCHER, A., DENG, Z., MAYS, A. D., DEW, I., DIETZ, S. M., DODSON, K., DOUP, L. E., DOWNES, M., DUGAN-ROCHA, S., DUNKOV, B. C., DUNN, P., DURBIN, K. J., EVANGELISTA, C. C., FERRAZ, C., FERRIERA, S., FLEISCHMANN, W., FOSLER, C., GABRIELIAN, A. E., GARG, N. S., GELBART, W. M., GLASSER, K., GLODEK, A., GONG, F., GORRELL, J. H., GU, Z., GUAN, P., HARRIS, M., HARRIS, N. L., HARVEY, D., HEIMAN, T. J., HERNANDEZ, J. R., HOUCK, J., HOSTIN, D., HOUSTON, K. A., HOWLAND, T. J., WEI, M. H., IBEGWAM, C., *et al*. 2000. The genome sequence of Drosophila melanogaster. *Science,* 287**,** 2185-95.

AKCIMEN, F., ROSS, J. P., BOURASSA, C. V., LIAO, C., ROCHEFORT, D., GAMA, M. T. D., DICAIRE, M. J., BARSOTTINI, O. G., BRAIS, B., PEDROSO, J. L., DION, P. A. & ROULEAU, G. A. 2019. Investigation of the RFC1 Repeat Expansion in a Canadian and a Brazilian Ataxia Cohort: Identification of Novel Conformations. *Front Genet,* 10**,** 1219.

AMIEL, J., LAUDIER, B., ATTIE-BITACH, T., TRANG, H., DE PONTUAL, L., GENER, B., TROCHET, D., ETCHEVERS, H., RAY, P., SIMONNEAU, M., VEKEMANS, M., MUNNICH, A., GAULTIER, C. & LYONNET, S. 2003. Polyalanine expansion and frameshift mutations of the paired-like homeobox gene PHOX2B in congenital central hypoventilation syndrome. *Nat Genet,* 33**,** 459-61.

AMIN, E., JAISWAL, M., DEREWENDA, U., REIS, K., NOURI, K., KOESSMEIER, K. T., ASPENSTROM, P., SOMLYO, A. V., DVORSKY, R. & AHMADIAN, M. R. 2016. Deciphering the Molecular and Functional Basis of RHOGAP Family Proteins: A SYSTEMATIC APPROACH TOWARD SELECTIVE INACTIVATION OF RHO FAMILY PROTEINS. *J Biol Chem,* 291**,** 20353-71.

ARIAS, A. M. 2008. Drosophila melanogaster and the development of biology in the 20th century. *Methods Mol Biol,* 420**,** 1-25.

AYMERIC RAVEL-CHAPUIS, E. D., BERNARD J JASMIN 2022. Pharmacological and exercise-induced activation of AMPK as emerging therapies for myotonic dystrophy type 1 patients.

BASSELL, G. J. & WARREN, S. T. 2008. Fragile X syndrome: loss of local mRNA regulation alters synaptic development and function. *Neuron,* 60**,** 201-14.

BASSETT, A. R. & LIU, J. L. 2014. CRISPR/Cas9 and genome editing in Drosophila. *J Genet Genomics,* 41**,** 7-19.

BEAUDIN, M., MANTO, M., SCHMAHMANN, J. D., PANDOLFO, M. & DUPRE, N. 2022. Recessive cerebellar and afferent ataxias - clinical challenges and future directions. *Nat Rev Neurol,* 18**,** 257-272.

BEAUDIN, M., MATILLA-DUENAS, A., SOONG, B. W., PEDROSO, J. L., BARSOTTINI, O. G., MITOMA, H., TSUJI, S., SCHMAHMANN, J. D., MANTO, M., ROULEAU, G. A., KLEIN, C. & DUPRE, N. 2019. The Classification of Autosomal Recessive Cerebellar Ataxias: a Consensus Statement from the Society for Research on the Cerebellum and Ataxias Task Force. *Cerebellum,* 18**,** 1098-1125.

BECK, S., FOTINOS, A., GAWAZ, M. & ELVERS, M. 2014. Nadrin GAP activity is isoform- and target-specific regulated by tyrosine phosphorylation. *Cell Signal,* 26**,** 1975-84.

BECK, S., FOTINOS, A., LANG, F., GAWAZ, M. & ELVERS, M. 2013. Isoform-specific roles of the GTPase activating protein Nadrin in cytoskeletal reorganization of platelets. *Cell Signal,* 25**,** 236-46.

BEDRAT, A., LACROIX, L. & MERGNY, J. L. 2016. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res,* 44**,** 1746-59.

BEECROFT, S. J., CORTESE, A., SULLIVAN, R., YAU, W. Y., DYER, Z., WU, T. Y., MULROY, E., PELOSI, L., RODRIGUES, M., TAYLOR, R., MOSSMAN, S., LEADBETTER, R., CLELAND, J., ANDERSON, T., RAVENSCROFT, G., LAING, N. G., HOULDEN, H., REILLY, M. M. & ROXBURGH, R. H. 2020. A Maori specific RFC1 pathogenic repeat configuration in CANVAS, likely due to a founder allele. *Brain,* 143**,** 2673-2680.

BEIJER, D., POLAVARAPU, K., PREETHISH-KUMAR, V., BARDHAN, M., DOHRN, M. F., REBELO, A., ZUCHNER, S. & NALINI, A. 2022. Homozygous N-terminal missense variant in PLEKHG5 associated with intermediate CMT: A case report. *J Neuromuscul Dis,* 9**,** 347-351.

BENSLIMANE, F. 2022. Commentary on: Single Incision Endoscope-Assisted Gastrocnemius Muscle Resection for Calf Hypertrophy: Analysis of 300 Cases. *Aesthet Surg J,* 42**,** 1041-1044.

BENSLIMANE, N., LORET, C., CHAZELAS, P., FAVREAU, F., FAYE, P. A., LEJEUNE, F. & LIA, A. S. 2024. Readthrough Activators and Nonsense-Mediated mRNA Decay Inhibitor Molecules: Real Potential in Many Genetic Diseases Harboring Premature Termination Codons. *Pharmaceuticals (Basel),* 17.

BIRD, T. D. 1993. Myotonic Dystrophy Type 1. *In:* ADAM, M. P., FELDMAN, J., MIRZAA, G. M., PAGON, R. A., WALLACE, S. E., BEAN, L. J. H., GRIPP, K. W. & AMEMIYA, A. (eds.) *GeneReviews((R)).* Seattle (WA).

BLAIR, I. P., NASH, J., GORDON, M. J. & NICHOLSON, G. A. 1996. Prevalence and origin of de novo duplications in Charcot-Marie-Tooth disease type 1A: first report of a de novo duplication with a maternal origin. *Am J Hum Genet,* 58**,** 472-6.

VAN BLITTERSWIJK, M., DEJESUS-HERNANDEZ, M., NIEMANTSVERDRIET, E., MURRAY, M. E., HECKMAN, M. G., DIEHL, N. N., BROWN, P. H., BAKER, M. C., FINCH, N. A., BAUER, P. O., SERRANO, G., BEACH, T. G., JOSEPHS, K. A., KNOPMAN, D. S., PETERSEN, R. C., BOEVE, B. F., GRAFF-RADFORD, N. R., BOYLAN, K. B., PETRUCELLI, L., DICKSON, D. W. & RADEMAKERS, R. 2013. Association between repeat sizes and clinical and pathological characteristics in carriers of C9ORF72 repeat expansions (Xpansize-72): a cross-sectional cohort study. *Lancet Neurol,* 12**,** 978-88.

BRAND, A. H. & PERRIMON, N. 1993. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development,* 118**,** 401-15.

BREUZARD, G., PAGANO, A., BASTONERO, S., MALESINSKI, S., PARAT, F., BARBIER, P., PEYROT, V. & KOVACIC, H. 2019. Tau regulates the microtubule-dependent migration of glioblastoma cells via the Rho-ROCK signaling pathway. *J Cell Sci,* 132.

BRONSTEIN, A. M., MOSSMAN, S. & LUXON, L. M. 1991. The neck-eye reflex in patients with reduced vestibular and optokinetic function. *Brain,* 114 (Pt 1A)**,** 1-11.

BROWN, M. P., VERMA, S., PALMER, I., GUERRERO ZUNIGA, A., MEHTA, A., ROSENSWEIG, C., KELES, M. F. & WU, M. N. 2024. A subclass of evening cells promotes the switch from arousal to sleep at dusk. *Curr Biol*.

CALAP-QUINTANA, P., NAVARRO, J. A., GONZALEZ-FERNANDEZ, J., MARTINEZ-SEBASTIAN, M. J., MOLTO, M. D. & LLORENS, J. V. 2018. Drosophila melanogaster Models of Friedreich's Ataxia. *Biomed Res Int,* 2018**,** 5065190.

CARLSON, B. R., LLOYD, K. E., KRUSZEWSKI, A., KIM, I. H., RODRIGUIZ, R. M., HEINDEL, C., FAYTELL, M., DUDEK, S. M., WETSEL, W. C. & SODERLING, S. H. 2011. WRP/srGAP3 facilitates the initiation of spine development by an inverse F-BAR domain, and its loss impairs long-term memory. *J Neurosci,* 31**,** 2447-60.

CARMICHAEL, J. B., PROVOST, P., EKWALL, K. & HOBMAN, T. C. 2004. ago1 and dcr1, two core components of the RNA interference pathway, functionally diverge from rdp1 in regulating cell cycle events in Schizosaccharomyces pombe. *Mol Biol Cell,* 15**,** 1425-35.

CAUSSINUS, E. & AFFOLTER, M. 2016. deGradFP: A System to Knockdown GFP-Tagged Proteins. *Methods Mol Biol,* 1478**,** 177-187.

CAUSSINUS, E., KANCA, O. & AFFOLTER, M. 2011. Fluorescent fusion protein knockout mediated by anti-GFP nanobody. *Nat Struct Mol Biol,* 19**,** 117-21.

CHAGNON, M. J., WU, C. L., NAKAZAWA, T., YAMAMOTO, T., NODA, M., BLANCHETOT, C. & TREMBLAY, M. L. 2010. Receptor tyrosine phosphatase sigma (RPTPsigma) regulates, p250GAP, a novel substrate that attenuates Rac signaling. *Cell Signal,* 22**,** 1626-33.

CHANG, S. & CAO, Y. 2021. The ROCK inhibitor Y-27632 ameliorates blood-spinal cord barrier disruption by reducing tight junction protein degradation via the MYPT1-MLC2 pathway after spinal cord injury in rats. *Brain Res,* 1773**,** 147684.

CHEN, H. C. 2005. Boyden chamber assay. *Methods Mol Biol,* 294**,** 15-22.

CHEN, K. Z., LIU, S. X., LI, Y. W., HE, T., ZHAO, J., WANG, T., QIU, X. X. & WU, H. F. 2023. Vimentin as a potential target for diverse nervous system diseases. *Neural Regen Res,* 18**,** 969-975.

CHEN, L., HADD, A., SAH, S., FILIPOVIC-SADIC, S., KROSTING, J., SEKINGER, E., PAN, R., HAGERMAN, P. J., STENZEL, T. T., TASSONE, F. & LATHAM, G. J. 2010. An information-rich CGG repeat primed PCR that detects the full range of fragile X expanded alleles and minimizes the need for southern blot analysis. *J Mol Diagn,* 12**,** 589-600.

CHEN, X., GAN, Y., AU, N. P. B. & MA, C. H. E. 2024. Current understanding of the molecular mechanisms of chemotherapy-induced peripheral neuropathy. *Front Mol Neurosci,* 17**,** 1345811.

CHEN, Z., MAROOFIAN, R., BASAK, A. N., SHINGAVI, L., KARAKAYA, M., EFTHYMIOU, S., GUSTAVSSON, E. K., MEIER, L., POLAVARAPU, K., VENGALIL, S., PREETHISH-KUMAR, V., NANDEESH, B. N., GOKCE GUNES, N., AKAN, O., CANDAN, F., SCHRANK, B., ZUCHNER, S., MURPHY, D., KAPOOR, M., RYTEN, M., WIRTH, B., REILLY, M. M., NALINI, A., HOULDEN, H. & SARRAF, P. 2021. Novel variants broaden the phenotypic spectrum of PLEKHG5-associated neuropathies. *Eur J Neurol,* 28**,** 1344-1355.

CHEN, Z., TUCCI, A., CIPRIANI, V., GUSTAVSSON, E. K., IBANEZ, K., REYNOLDS, R. H., ZHANG, D., VESTITO, L., GARCIA, A. C., SETHI, S., BRENTON, J. W., GARCIA-RUIZ, S., FAIRBROTHER-BROWNE, A., GIL-MARTINEZ, A. L., GENOMICS ENGLAND RESEARCH, C., WOOD, N., HARDY, J. A., SMEDLEY, D., HOULDEN, H., BOTIA, J. & RYTEN, M. 2023. Functional genomics provide key insights to improve the diagnostic yield of hereditary ataxia. *Brain,* 146**,** 2869-2884.

CHINTALAPHANI, S. R., PINEDA, S. S., DEVESON, I. W. & KUMAR, K. R. 2021. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. *Acta Neuropathol Commun,* 9**,** 98.

CIPRIANI, S., GUERRERO-VALERO, M., TOZZA, S., ZHAO, E., VOLLMER, V., BEIJER, D., DANZI, M., RIVELLINI, C., LAZAREVIC, D., PIPITONE, G. B., GROSZ, B. R., LAMPERTI, C., MARZOLI, S. B., CARRERA, P., DEVOTO, M., PISCIOTTA, C., PAREYSON, D., KENNERSON, M., PREVITALI, S. C., ZUCHNER, S., SCHERER, S. S., MANGANELLI, F., BAHLER, M. & BOLINO, A. 2023. Mutations in MYO9B are associated with Charcot-Marie-Tooth disease type 2 neuropathies and isolated optic atrophy. *Eur J Neurol,* 30**,** 511-526.

COOK, A. & GIUNTI, P. 2017. Friedreich's ataxia: clinical features, pathogenesis and management. *Br Med Bull,* 124**,** 19-30.

CORTESE, A., CURRO, R., VEGEZZI, E., YAU, W. Y., HOULDEN, H. & REILLY, M. M. 2022. Cerebellar ataxia, neuropathy and vestibular areflexia syndrome (CANVAS): genetic and clinical aspects. *Pract Neurol,* 22**,** 14-18.

CORTESE, A., SIMONE, R., SULLIVAN, R., VANDROVCOVA, J., TARIQ, H., YAU, W. Y., HUMPHREY, J., JAUNMUKTANE, Z., SIVAKUMAR, P., POLKE, J., ILYAS, M., TRIBOLLET, E., TOMASELLI, P. J., DEVIGILI, G., CALLEGARI, I., VERSINO, M., SALPIETRO, V., EFTHYMIOU, S., KASKI, D., WOOD, N. W., ANDRADE, N. S., BUGLO, E., REBELO, A., ROSSOR, A. M., BRONSTEIN, A., FRATTA, P., MARQUES, W. J., ZUCHNER, S., REILLY, M. M. & HOULDEN, H. 2019. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat Genet,* 51**,** 649-658.

CORTESE, A., TOZZA, S., YAU, W. Y., ROSSI, S., BEECROFT, S. J., JAUNMUKTANE, Z., DYER, Z., RAVENSCROFT, G., LAMONT, P. J., MOSSMAN, S., CHANCELLOR, A., MAISONOBE, T., PEREON, Y., CAUQUIL, C., COLNAGHI, S., MALLUCCI, G., CURRO, R., TOMASELLI, P. J., THOMAS-BLACK, G., SULLIVAN, R., EFTHYMIOU, S., ROSSOR, A. M., LAURA, M., PIPIS, M., HORGA, A., POLKE, J., KASKI, D., HORVATH, R., CHINNERY, P. F., MARQUES, W., TASSORELLI, C., DEVIGILI, G., LEONARDIS, L., WOOD, N. W., BRONSTEIN, A., GIUNTI, P., ZUCHNER, S., STOJKOVIC, T., LAING, N., ROXBURGH, R. H., HOULDEN, H. & REILLY, M. M. 2020. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome due to RFC1 repeat expansion. *Brain,* 143**,** 480-490.

CORTESE, A., ZHU, Y., REBELO, A. P., NEGRI, S., COUREL, S., ABREU, L., BACON, C. J., BAI, Y., BIS-BREWER, D. M., BUGIARDINI, E., BUGLO, E., DANZI, M. C., FEELY, S. M. E., ATHANASIOU-FRAGKOULI, A., HARIDY, N. A., INHERITED NEUROPATHY, C., ISASI, R., KHAN, A., LAURA, M., MAGRI, S., PIPIS, M., PISCIOTTA, C., POWELL, E., ROSSOR, A. M., SAVERI, P., SOWDEN, J. E., TOZZA, S., VANDROVCOVA, J., DALLMAN, J., GRIGNANI, E., MARCHIONI, E., SCHERER, S. S., TANG, B., LIN, Z., AL-AJMI, A., SCHULE, R., SYNOFZIK, M., MAISONOBE, T., STOJKOVIC, T., AUER-GRUMBACH, M., ABDELHAMED, M. A., HAMED, S. A., ZHANG, R.,

MANGANELLI, F., SANTORO, L., TARONI, F., PAREYSON, D., HOULDEN, H., HERRMANN, D. N., REILLY, M. M., SHY, M. E., ZHAI, R. G. & ZUCHNER, S. 2020. Biallelic mutations in SORD cause a common and potentially treatable hereditary neuropathy with implications for diabetes. *Nat Genet,* 52**,** 473-481.

CORY, G. 2011. Scratch-wound assay. *Methods Mol Biol,* 769**,** 25-30.

COUTELLE, O., BLAGDEN, C. S., HAMPSON, R., HALAI, C., RIGBY, P. W. & HUGHES, S. M. 2001. Hedgehog signalling is required for maintenance of myf5 and myoD expression and timely terminal differentiation in zebrafish adaxial myogenesis. *Dev Biol,* 236**,** 136-50.

CURRO, R., DOMINIK, N., FACCHINI, S., VEGEZZI, E., SULLIVAN, R., GALASSI DEFORIE, V., FERNANDEZ-EULATE, G., TRASCHUTZ, A., ROSSI, S., GARIBALDI, M., KWARCIANY, M., TARONI, F., BRUSCO, A., GOOD, J. M., CAVALCANTI, F., HAMMANS, S., RAVENSCROFT, G., ROXBURGH, R. H., GROUP, R. F. C. R. E. S., PAROLIN SCHNEKENBERG, R., RUGGININI, B., ABATI, E., MANINI, A., QUARTESAN, I., GHIA, A., LOPEZ DE MUNAIN, A., MANGANELLI, F., KENNERSON, M., SANTORELLI, F. M., INFANTE, J., MARQUES, W., JOKELA, M., MURPHY, S. M., MANDICH, P., FABRIZI, G. M., BRIANI, C., GOSAL, D., PAREYSON, D., FERRARI, A., PRADOS, F., YOUSRY, T., KHURANA, V., KUO, S. H., MILLER, J., TROAKES, C., JAUNMUKTANE, Z., GIUNTI, P., HARTMANN, A., BASAK, N., SYNOFZIK, M., STOJKOVIC, T., HADJIVASSILIOU, M., REILLY, M. M., HOULDEN, H. & CORTESE, A. 2024. Role of the repeat expansion size in predicting age of onset and severity in RFC1 disease. *Brain,* 147**,** 1887-1898.

CURTIS-CIOFFI, K. M., RODRIGUEIRO, D. A., RODRIGUES, V. C., CICARELLI, R. M. & SCAREL-CAMINAGA, R. M. 2012. Comparison between the polymerase chain reaction-based screening and the Southern blot methods for identification of fragile X syndrome. *Genet Test Mol Biomarkers,* 16**,** 1303-8.

D'ANGELO, C. S., HERMES, A., MCMASTER, C. R., PRICHEP, E., RICHER, E., VAN DER WESTHUIZEN, F. H., REPETTO, G. M., MENGCHUN, G., MALHERBE, H., REICHARDT, J. K. V., ARBOUR, L., HUDSON, M., DU PLESSIS, K., HAENDEL, M., WILCOX, P., LYNCH, S. A., RIND, S., EASTEAL, S., ESTIVILL, X., THOMAS, Y. & BAYNAM, G. 2020. Barriers and Considerations for Diagnosing Rare Diseases in Indigenous Populations. *Front Pediatr,* 8**,** 579924.

DAVID, M. D., PETIT, D. & BERTOGLIO, J. 2014. The RhoGAP ARHGAP19 controls cytokinesis and chromosome segregation in T lymphocytes. *J Cell Sci,* 127**,** 400-10.

DAVIE, K., JANSSENS, J., KOLDERE, D., DE WAEGENEER, M., PECH, U., KREFT, L., AIBAR, S., MAKHZAMI, S., CHRISTIAENS, V., BRAVO GONZALEZ-BLAS, C., POOVATHINGAL, S., HULSELMANS, G., SPANIER, K. I., MOERMAN, T., VANSPAUWEN, B., GEURS, S., VOET, T., LAMMERTYN, J., THIENPONT, B., LIU, S., KONSTANTINIDES, N., FIERS, M., VERSTREKEN, P. & AERTS, S. 2018. A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. *Cell,* 174**,** 982-998 e20.

DAVIES, K., SZMULEWICZ, D. J., CORBEN, L. A., DELATYCKI, M. & LOCKHART, P. J. 2022. RFC1-Related Disease: Molecular and Clinical Insights. *Neurol Genet,* 8**,** e200016.

DEACON, S., DALLEYWATER, W., PEAT, C., PAINE, S. M. L. & DINEEN, R. A. 2024. Disproportionate Expression of ATM in Cerebellar Cortex During Human Neurodevelopment. *Cerebellum,* 23**,** 502-511.

DEGEER, J. & LAMARCHE-VANE, N. 2013. Rho GTPases in neurodegeneration diseases. *Exp Cell Res,* 319**,** 2384-94.

DEJESUS-HERNANDEZ, M., MACKENZIE, I. R., BOEVE, B. F., BOXER, A. L., BAKER, M., RUTHERFORD, N. J., NICHOLSON, A. M., FINCH, N. A., FLYNN, H., ADAMSON, J., KOURI, N., WOJTAS, A., SENGDY, P., HSIUNG, G. Y., KARYDAS, A., SEELEY, W. W., JOSEPHS, K. A., COPPOLA, G., GESCHWIND, D. H., WSZOLEK, Z. K., FELDMAN, H., KNOPMAN, D. S., PETERSEN, R. C., MILLER, B. L., DICKSON, D. W., BOYLAN, K. B., GRAFF-RADFORD, N. R. & RADEMAKERS, R. 2011. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron,* 72**,** 245-56.

DELATYCKI, M. B. & BIDICHANDANI, S. I. 2019. Friedreich ataxia- pathogenesis and implications for therapies. *Neurobiol Dis,* 132**,** 104606.

DELATYCKI, M. B. & CORBEN, L. A. 2012. Clinical features of Friedreich ataxia. *J Child Neurol,* 27**,** 1133-7.

DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M. & GINGERAS, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

DOLZHENKO, E., ENGLISH, A., DASHNOW, H., DE SENA BRANDINE, G., MOKVELD, T., ROWELL, W. J., KARNISKI, C., KRONENBERG, Z., DANZI, M. C., CHEUNG, W. A., BI, C., FARROW, E., WENGER, A., CHUA, K. P., MARTINEZ-CERDENO, V., BARTLEY, T. D., JIN, P., NELSON, D. L., ZUCHNER, S., PASTINEN, T., QUINLAN, A. R., SEDLAZECK, F. J. & EBERLE, M. A. 2024. Characterization and visualization of tandem repeats at genome scale. *Nat Biotechnol*.

DOMINIK, N., MAGRI, S., CURRO, R., ABATI, E., FACCHINI, S., CORBETTA, M., MACPHERSON, H., DI BELLA, D., SARTO, E., STEVANOVSKI, I., CHINTALAPHANI, S. R., AKCIMEN, F., MANINI, A., VEGEZZI, E., QUARTESAN, I., MONTGOMERY, K. A., PIROTA, V., CRESPAN, E., PERINI, C., GRUPELLI, G. P., TOMASELLI, P. J., MARQUES, W., GENOMICS ENGLAND RESEARCH, C., SHAW, J., POLKE, J., SALSANO, E., FENU, S., PAREYSON, D., PISCIOTTA, C., TOFARIS, G. K., NEMETH, A. H., EALING, J., RADUNOVIC, A., KEARNEY, S., KUMAR, K. R., VUCIC, S., KENNERSON, M., REILLY, M. M., HOULDEN, H., DEVESON, I., TUCCI, A., TARONI, F. & CORTESE, A. 2023. Normal and pathogenic variation of RFC1 repeat expansions: implications for clinical diagnosis. *Brain,* 146**,** 5060-5069.

EFTHYMIOU, S., MANOLE, A. & HOULDEN, H. 2016. Next-generation sequencing in neuromuscular diseases. *Curr Opin Neurol,* 29**,** 527-36.

ELBASHIR, S. M., MARTINEZ, J., PATKANIOWSKA, A., LENDECKEL, W. & TUSCHL, T. 2001. Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. *EMBO J,* 20**,** 6877-88.

EL-BORAIE, A., CHENOWETH, M. J., POUGET, J. G., BENOWITZ, N. L., FUKUNAGA, K., MUSHIRODA, T., KUBO, M., NOLLEN, N. L., SANDERSON COX, L., LERMAN, C., KNIGHT, J. & TYNDALE, R. F. 2021. Transferability of Ancestry-Specific and Cross-Ancestry CYP2A6 Activity Genetic Risk Scores in African and European Populations. *Clin Pharmacol Ther,* 110**,** 975-985.

ENGERT, J. C., BERUBE, P., MERCIER, J., DORE, C., LEPAGE, P., GE, B., BOUCHARD, J. P., MATHIEU, J., MELANCON, S. B., SCHALLING, M., LANDER, E. S., MORGAN, K., HUDSON, T. J. & RICHTER, A. 2000. ARSACS, a spastic ataxia common in northeastern Quebec, is caused by mutations in a new gene encoding an 11.5-kb ORF. *Nat Genet,* 24**,** 120-5.

ERDMANN, H., SCHOBERL, F., GIURGIU, M., LEAL SILVA, R. M., SCHOLZ, V., SCHARF, F., WENDLANDT, M., KLEINLE, S., DESCHAUER, M., NUBLING, G., HEIDE, W., BABACAN, S. S., SCHNEIDER, C., NEUHANN, T., HAHN, K., SCHOSER, B., HOLINSKI-FEDER, E., WOLF, D. A. & ABICHT, A. 2023. Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing. *Brain,* 146**,** 1831-1843.

FAN, Y., ZHANG, S., YANG, J., MAO, C. Y., YANG, Z. H., HU, Z. W., WANG, Y. L., LIU, Y. T., LIU, H., YUAN, Y. P., SHI, C. H. & XU, Y. M. 2020. No biallelic intronic AAGGG repeat expansion in RFC1 was found in patients with late-onset ataxia and MSA. *Parkinsonism Relat Disord,* 73**,** 1-2.

FAUTSCH, M. P., WIEBEN, E. D., BARATZ, K. H., BHATTACHARYYA, N., SADAN, A. N., HAFFORD-TEAR, N. J., TUFT, S. J. & DAVIDSON, A. E. 2021. TCF4-mediated Fuchs endothelial corneal dystrophy: Insights into a common trinucleotide repeat-associated disease. *Prog Retin Eye Res,* 81**,** 100883.

FAVILLE, R., KOTTLER, B., GOODHILL, G. J., SHAW, P. J. & VAN SWINDEREN, B. 2015. How deeply does your mutant sleep? Probing arousal to better understand sleep defects in Drosophila. *Sci Rep,* 5**,** 8454.

FEELY, S. M., LAURA, M., SISKIND, C. E., SOTTILE, S., DAVIS, M., GIBBONS, V. S., REILLY, M. M. & SHY, M. E. 2011. MFN2 mutations cause severe phenotypes in most patients with CMT2A. *Neurology,* 76**,** 1690-6.

FERNANDEZ, G., YUBERO, D., PALAU, F. & ARMSTRONG, J. 2022. Molecular Modelling Hurdle in the Next-Generation Sequencing Era. *Int J Mol Sci,* 23.

FERREIRA, C. R. 2019. The burden of rare diseases. *Am J Med Genet A,* 179**,** 885-892.

FILLA, A., DE MICHELE, G., CAVALCANTI, F., PIANESE, L., MONTICELLI, A., CAMPANELLA, G. & COCOZZA, S. 1996. The relationship between trinucleotide (GAA) repeat length and clinical features in Friedreich ataxia. *Am J Hum Genet,* 59**,** 554-60.

FORTUNATO, F., TONELLI, L., FARNE, M., SELVATICI, R. & FERLINI, A. 2023. DMD deletions underlining mild dystrophinopathies: literature review highlights phenotype-related mutation clusters and provides insights about genetic mechanisms and prognosis. *Front Neurol,* 14**,** 1288721.

FOSSATI, M., PIZZARELLI, R., SCHMIDT, E. R., KUPFERMAN, J. V., STROEBEL, D., POLLEUX, F. & CHARRIER, C. 2016. SRGAP2 and Its Human-Specific Paralog Co-Regulate the Development of Excitatory and Inhibitory Synapses. *Neuron,* 91**,** 356-69.

FRANCASTEL, C. & MAGDINIER, F. 2019. DNA methylation in satellite repeats disorders. *Essays Biochem,* 63**,** 757-771.

FRASSON, I., PIROTA, V., RICHTER, S. N. & DORIA, F. 2022. Multimeric G-quadruplexes: A review on their biological roles and targeting. *Int J Biol Macromol,* 204**,** 89-102.

FREDERIKSEN, S. D., AVRAMOVIC, V., MAROILLEY, T., LEHMAN, A., ARBOUR, L. & TARAILO-GRAOVAC, M. 2022. Rare disorders have many faces: in silico characterization of rare disorder spectrum. *Orphanet J Rare Dis,* 17**,** 76.

GAHL, W. A. 2012. The battlefield of rare diseases: where uncommon insights are common. *Sci Transl Med,* 4**,** 154ed7.

GARDINER, S. L., VAN BELZEN, M. J., BOOGAARD, M. W., VAN ROON-MOM, W. M. C., ROZING, M. P., VAN HEMERT, A. M., SMIT, J. H., BEEKMAN, A. T. F., VAN GROOTHEEST, G., SCHOEVERS, R. A., OUDE VOSHAAR, R. C., ROOS, R. A. C., COMIJS, H. C., PENNINX, B., VAN DER MAST, R. C. & AZIZ, N. A.

2017. Huntingtin gene repeat size variations affect risk of lifetime depression. *Transl Psychiatry,* 7**,** 1277.
GANDOLFO, L. C., BAHLO, M. & SPEED, T. P. 2014. Dating rare mutations from small samples with dense marker data. *Genetics,* 197**,** 1315-27.
GARRETT, M. D., SELF, A. J., VAN OERS, C. & HALL, A. 1989. Identification of distinct cytoplasmic targets for ras/R-ras and rho regulatory proteins. *J Biol Chem,* 264**,** 10-3.
GEBUS, O., MONTAUT, S., MONGA, B., WIRTH, T., CHERAUD, C., ALVES DO REGO, C., ZINCHENKO, I., CARRE, G., HAMDAOUI, M., HAUTECLOQUE, G., NGUYEN-THEM, L., LANNES, B., CHANSON, J. B., LAGHA-BOUKBIZA, O., FLEURY, M. C., DEVYS, D., NICOLAS, G., RUDOLF, G., BEREAU, M., MALLARET, M., RENAUD, M., ACQUAVIVA, C., KOENIG, M., KOOB, M., KREMER, S., NAMER, I. J., CAZENEUVE, C., ECHANIZ-LAGUNA, A., TRANCHANT, C. & ANHEIM, M. 2017. Deciphering the causes of sporadic late-onset cerebellar ataxias: a prospective study with implications for diagnostic work. *J Neurol,* 264**,** 1118-1126.
GHORBANI, F., DE BOER-BERGSMA, J., VERSCHUUREN-BEMELMANS, C. C., PENNINGS, M., DE BOER, E. N., KREMER, B., VANHOUTTE, E. K., DE VRIES, J. J., VAN DE BERG, R., KAMSTEEG, E. J., VAN DIEMEN, C. C., WESTERS, H., VAN DE WARRENBURG, B. P. & VERBEEK, D. S. 2022. Prevalence of intronic repeat expansions in RFC1 in Dutch patients with CANVAS and adult-onset ataxia. *J Neurol,* 269**,** 6086-6093.
GRASSO, M., BOON, E. M., FILIPOVIC-SADIC, S., VAN BUNDEREN, P. A., GENNARO, E., CAO, R., LATHAM, G. J., HADD, A. G. & COVIELLO, D. A. 2014. A novel methylation PCR that offers standardized determination of FMR1 methylation and CGG repeat length without southern blot analysis. *J Mol Diagn,* 16**,** 23-31.
GREEN, E. D. & DONOHUE, C. R. 2018. Special Issue Editors' Introduction: "Genomics and the Human Genome Project". *J Hist Biol,* 51**,** 625-629.
GREENBERG, L. & HATINI, V. 2011. Systematic expression and loss-of-function analysis defines spatially restricted requirements for Drosophila RhoGEFs and RhoGAPs in leg morphogenesis. *Mech Dev,* 128**,** 5-17.
GREENSPAN, G. & GEIGER, D. 2004. Model-based inference of haplotype block variation. *J Comput Biol,* 11**,** 493-504.
GROH, J., KLEIN, I., HOLLMANN, C., WETTMARSHAUSEN, J., KLEIN, D. & MARTINI, R. 2015. CSF-1-activated macrophages are target-directed and essential mediators of Schwann cell dedifferentiation and dysfunction in Cx32-deficient mice. *Glia,* 63**,** 977-86.
GROSZ, B. R., GOLOVCHENKO, N. B., ELLIS, M., KUMAR, K., NICHOLSON, G. A., ANTONELLIS, A. & KENNERSON, M. L. 2019. A de novo EGR2 variant, c.1232A > G p.Asp411Gly, causes severe early-onset Charcot-Marie-Tooth Neuropathy Type 3 (Dejerine-Sottas Neuropathy). *Sci Rep,* 9**,** 19336.
HADJIVASSILIOU, M., CURRO, R., BEAUCHAMP, N., DOMINIK, N., GRUNEWALD, R. A., SHANMUGARAJAH, P., ZIS, P., HOGGARD, N. & CORTESE, A. 2024. Can CANVAS due to RFC1 biallelic expansions present with pure ataxia? *J Neurol Neurosurg Psychiatry,* 95**,** 171-174.
HAEUSLER, A. R., DONNELLY, C. J., PERIZ, G., SIMKO, E. A., SHAW, P. G., KIM, M. S., MARAGAKIS, N. J., TRONCOSO, J. C., PANDEY, A., SATTLER, R., ROTHSTEIN, J. D. & WANG, J. 2014. C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature,* 507**,** 195-200.

HE, Y., GOYETTE, M. A., CHAPELLE, J., BOUFAIED, N., AL RAHBANI, J., SCHONEWOLFF, M., DANEK, E. I., MULLER, W. J., LABBE, D. P., COTE, J. F. & LAMARCHE-VANE, N. 2023. CdGAP is a talin-binding protein and a target of TGF-beta signaling that promotes HER2-positive breast cancer growth and metastasis. *Cell Rep,* 42**,** 112936.

HERAUD, C., PINAULT, M., LAGREE, V. & MOREAU, V. 2019. p190RhoGAPs, the ARHGAP35- and ARHGAP5-Encoded Proteins, in Health and Disease. *Cells,* 8.

HERNANDEZ, D., TEWHEY, R., VEYRIERAS, J. B., FARINELLI, L., OSTERAS, M., FRANCOIS, P. & SCHRENZEL, J. 2014. De novo finished 2.8 Mbp Staphylococcus aureus genome assembly from 100 bp short and long range paired-end reads. *Bioinformatics,* 30**,** 40-9.

HERNANDEZ, D. G., REED, X. & SINGLETON, A. B. 2016. Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. *J Neurochem,* 139 Suppl 1**,** 59-74.

HOLBROOK, S., FINLEY, J. K., LYONS, E. L. & HERMAN, T. G. 2012. Loss of syd-1 from R7 neurons disrupts two distinct phases of presynaptic development. *J Neurosci,* 32**,** 18101-11.

HOULDEN, H., LAURA, M., WAVRANT-DE VRIEZE, F., BLAKE, J., WOOD, N. & REILLY, M. M. 2008. Mutations in the HSP27 (HSPB1) gene cause dominant, recessive, and sporadic distal HMN/CMT type 2. *Neurology,* 71**,** 1660-8.

HUANG, G. H., SUN, Z. L., LI, H. J. & FENG, D. F. 2017. Rho GTPase-activating proteins: Regulators of Rho GTPase activity in neuronal development and CNS diseases. *Mol Cell Neurosci,* 80**,** 18-31.

HUANG, H., WANG, S., GUAN, Y., REN, J. & LIU, X. 2024. Molecular basis and current insights of atypical Rho small GTPase in cancer. *Mol Biol Rep,* 51**,** 141.

HUANG, L., FANG, L., LIU, Q., TORSHIZI, A. D. & WANG, K. 2022. Integrated analysis on transcriptome and behaviors defines HTT repeat-dependent network modules in Huntington's disease. *Genes Dis,* 9**,** 479-493.

INFANTE, J., GARCIA, A., SERRANO-CARDENAS, K. M., GONZALEZ-AGUADO, R., GAZULLA, J., DE LUCAS, E. M. & BERCIANO, J. 2018. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS) with chronic cough and preserved muscle stretch reflexes: evidence for selective sparing of afferent Ia fibres. *J Neurol,* 265**,** 1454-1462.

INVESTIGATORS, G. P. P., SMEDLEY, D., SMITH, K. R., MARTIN, A., THOMAS, E. A., MCDONAGH, E. M., CIPRIANI, V., ELLINGFORD, J. M., ARNO, G., TUCCI, A., VANDROVCOVA, J., CHAN, G., WILLIAMS, H. J., RATNAIKE, T., WEI, W., STIRRUPS, K., IBANEZ, K., MOUTSIANAS, L., WIELSCHER, M., NEED, A., BARNES, M. R., VESTITO, L., BUCHANAN, J., WORDSWORTH, S., ASHFORD, S., REHMSTROM, K., LI, E., FULLER, G., TWISS, P., SPASIC-BOSKOVIC, O., HALSALL, S., FLOTO, R. A., POOLE, K., WAGNER, A., MEHTA, S. G., GURNELL, M., BURROWS, N., JAMES, R., PENKETT, C., DEWHURST, E., GRAF, S., MAPETA, R., KASANICKI, M., HAWORTH, A., SAVAGE, H., BABCOCK, M., REESE, M. G., BALE, M., BAPLE, E., BOUSTRED, C., BRITTAIN, H., DE BURCA, A., BLEDA, M., DEVEREAU, A., HALAI, D., HARALDSDOTTIR, E., HYDER, Z., KASPERAVICIUTE, D., PATCH, C., POLYCHRONOPOULOS, D., MATCHAN, A., SULTANA, R., RYTEN, M., TAVARES, A. L. T., TREGIDGO, C., TURNBULL, C., WELLAND, M., WOOD, S., SNOW, C., WILLIAMS, E., LEIGH, S., FOULGER, R. E., DAUGHERTY, L. C., NIBLOCK, O., LEONG, I. U. S., WRIGHT, C. F., DAVIES, J., CRICHTON, C., WELCH, J., WOODS, K., ABULHOUL, L., AURORA, P., BOCKENHAUER, D.,

BROOMFIELD, A., CLEARY, M. A., LAM, T., DATTANI, M., FOOTITT, E., GANESAN, V., GRUNEWALD, S., COMPEYROT-LACASSAGNE, S., MUNTONI, F., PILKINGTON, C., QUINLIVAN, R., THAPAR, N., WALLIS, C., WEDDERBURN, L. R., WORTH, A., BUESER, T., COMPTON, C., *et al*. 2021. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl J Med,* 385**,** 1868-1880.

ISHIURA, H., SHIBATA, S., YOSHIMURA, J., SUZUKI, Y., QU, W., DOI, K., ALMANSOUR, M. A., KIKUCHI, J. K., TAIRA, M., MITSUI, J., TAKAHASHI, Y., ICHIKAWA, Y., MANO, T., IWATA, A., HARIGAYA, Y., MATSUKAWA, M. K., MATSUKAWA, T., TANAKA, M., SHIROTA, Y., OHTOMO, R., KOWA, H., DATE, H., MITSUE, A., HATSUTA, H., MORIMOTO, S., MURAYAMA, S., SHIIO, Y., SAITO, Y., MITSUTAKE, A., KAWAI, M., SASAKI, T., SUGIYAMA, Y., HAMADA, M., OHTOMO, G., TERAO, Y., NAKAZATO, Y., TAKEDA, A., SAKIYAMA, Y., UMEDA-KAMEYAMA, Y., SHINMI, J., OGATA, K., KOHNO, Y., LIM, S. Y., TAN, A. H., SHIMIZU, J., GOTO, J., NISHINO, I., TODA, T., MORISHITA, S. & TSUJI, S. 2019. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nat Genet,* 51**,** 1222-1232.

JARIUS, S., WANDINGER, K. P., HORN, S., HEUER, H. & WILDEMANN, B. 2010. A new Purkinje cell antibody (anti-Ca) associated with subacute cerebellar ataxia: immunological characterization. *J Neuroinflammation,* 7**,** 21.

JOOSTEN, I. B. T., HELLEBREKERS, D., DE GREEF, B. T. A., SMEETS, H. J. M., DE DIE-SMULDERS, C. E. M., FABER, C. G. & GERRITS, M. M. 2020. Parental repeat length instability in myotonic dystrophy type 1 pre- and protomutations. *Eur J Hum Genet,* 28**,** 956-962.

KAARTINEN, V., GONZALEZ-GOMEZ, I., VONCKEN, J. W., HAATAJA, L., FAURE, E., NAGY, A., GROFFEN, J. & HEISTERKAMP, N. 2001. Abnormal function of astroglia lacking Abr and Bcr RacGAPs. *Development,* 128**,** 4217-27.

KACHER, R., LEJEUNE, F. X., DAVID, I., BOLUDA, S., COARELLI, G., LECLERE-TURBANT, S., HEINZMANN, A., MARELLI, C., CHARLES, P., GOIZET, C., KABIR, N., HILAB, R., JORNEA, L., SIX, J., DOMMERGUES, M., FAURET, A. L., BRICE, A., HUMBERT, S. & DURR, A. 2024. CAG repeat mosaicism is gene specific in spinocerebellar ataxias. *Am J Hum Genet,* 111**,** 913-926.

KAMIZA, A. B., TOURE, S. M., VUJKOVIC, M., MACHIPISA, T., SOREMEKUN, O. S., KINTU, C., CORPAS, M., PIRIE, F., YOUNG, E., GILL, D., SANDHU, M. S., KALEEBU, P., NYIRENDA, M., MOTALA, A. A., CHIKOWORE, T. & FATUMO, S. 2022. Transferability of genetic risk scores in African populations. *Nat Med,* 28**,** 1163-1166.

KERN, J. V., ZHANG, Y. V., KRAMER, S., BRENMAN, J. E. & RASSE, T. M. 2013. The kinesin-3, unc-104 regulates dendrite morphogenesis and synaptic development in Drosophila. *Genetics,* 195**,** 59-72.

KHELFAOUI, M., DENIS, C., VAN GALEN, E., DE BOCK, F., SCHMITT, A., HOUBRON, C., MORICE, E., GIROS, B., RAMAKERS, G., FAGNI, L., CHELLY, J., NOSTEN-BERTRAND, M. & BILLUART, P. 2007. Loss of X-linked mental retardation gene oligophrenin1 in mice impairs spatial memory and leads to ventricular enlargement and dendritic spine immaturity. *J Neurosci,* 27**,** 9439-50.

KIKIN, O., D'ANTONIO, L. & BAGGA, P. S. 2006. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res,* 34**,** W676-82.

KIM, B., JOO LEE, C., WON, H. H. & LEE, S. H. 2023. Genetic Variants Associated with Supernormal Coronary Arteries. *J Atheroscler Thromb,* 30**,** 467-480.

KIM, M. J., KIM, S. C. & KIM, Y. J. 2018. A Universal Analysis Pipeline of Hybrid Capture-Based Targeted Sequencing Data with Unique Molecular Indexes (UMIs). *Genomics Inform,* 16**,** e29.
KLEIN, C. J. 2020. Charcot-Marie-Tooth Disease and Other Hereditary Neuropathies. *Continuum (Minneap Minn),* 26**,** 1224-1256.
KLOCKGETHER, T. 2010. Sporadic ataxia with adult onset: classification and diagnostic criteria. *Lancet Neurol,* 9**,** 94-104.
KOVALENKO, M., DRAGILEVA, E., ST CLAIRE, J., GILLIS, T., GUIDE, J. R., NEW, J., DONG, H., KUCHERLAPATI, R., KUCHERLAPATI, M. H., EHRLICH, M. E., LEE, J. M. & WHEELER, V. C. 2012. Msh2 acts in medium-spiny striatal neurons as an enhancer of CAG instability and mutant huntingtin phenotypes in Huntington's disease knock-in mice. *PLoS One,* 7**,** e44273.
KRATTER, I. H. & FINKBEINER, S. 2010. PolyQ disease: too many Qs, too much function? *Neuron,* 67**,** 897-9.
KROLL, F., POWELL, G. T., GHOSH, M., GESTRI, G., ANTINUCCI, P., HEARN, T. J., TUNBAK, H., LIM, S., DENNIS, H. W., FERNANDEZ, J. M., WHITMORE, D., DREOSTI, E., WILSON, S. W., HOFFMAN, E. J. & RIHEL, J. 2021. A simple and effective F0 knockout method for rapid screening of behaviour and other complex phenotypes. *Elife,* 10.
LAKE, C. M., HOLSCLAW, J. K., BELLENDIR, S. P., SEKELSKY, J. & HAWLEY, R. S. 2013. The development of a monoclonal antibody recognizing the Drosophila melanogaster phosphorylated histone H2A variant (gamma-H2AV). *G3 (Bethesda),* 3**,** 1539-43.
LAM, C., GILLIAM, K. M., RODDEN, L. N., SCHADT, K. A., LYNCH, D. R. & BIDICHANDANI, S. 2023. FXN gene methylation determines carrier status in Friedreich ataxia. *J Med Genet,* 60**,** 797-800.
LAURA, M., PIPIS, M., ROSSOR, A. M. & REILLY, M. M. 2019. Charcot-Marie-Tooth disease and related disorders: an evolving landscape. *Curr Opin Neurol,* 32**,** 641-650.
LEE, P. T., ZIRIN, J., KANCA, O., LIN, W. W., SCHULZE, K. L., LI-KROEGER, D., TAO, R., DEVEREAUX, C., HU, Y., CHUNG, V., FANG, Y., HE, Y., PAN, H., GE, M., ZUO, Z., HOUSDEN, B. E., MOHR, S. E., YAMAMOTO, S., LEVIS, R. W., SPRADLING, A. C., PERRIMON, N. & BELLEN, H. J. 2018. A gene-specific T2A-GAL4 library for Drosophila. *Elife,* 7.
LEITAO, E., SCHRODER, C. & DEPIENNE, C. 2024. Identification and characterization of repeat expansions in neurological disorders: Methodologies, tools, and strategies. *Rev Neurol (Paris)*.
LI, M., YE, L., YE, X., WANG, S., ZHANG, H., LIU, J. & HONG, H. 2019. Hypoxia-induced ARHGAP26 deficiency inhibits the proliferation and migration of human ductus arteriosus smooth muscle cell through activating RhoA-ROCK-PTEN pathway. *J Cell Biochem,* 120**,** 10106-10117.
LI, Y., GAN, S., REN, L., YUAN, L., LIU, J., WANG, W., WANG, X., ZHANG, Y., JIANG, J., ZHANG, F. & QI, X. 2018. Multifaceted regulation and functions of replication factor C family in human cancers. *Am J Cancer Res,* 8**,** 1343-1355.
LI, H., JANSSENS, J., DE WAEGENEER, M., KOLLURU, S. S., DAVIE, K., GARDEUX, V., SAELENS, W., DAVID, F. P. A., BRBIC, M., SPANIER, K., LESKOVEC, J., MCLAUGHLIN, C. N., XIE, Q., JONES, R. C., BRUECKNER, K., SHIM, J., TATTIKOTA, S. G., SCHNORRER, F., RUST, K., NYSTUL, T. G., CARVALHO-SANTOS, Z., RIBEIRO, C., PAL, S., MAHADEVARAJU, S., PRZYTYCKA, T. M., ALLEN, A. M., GOODWIN, S. F., BERRY, C. W., FULLER, M. T., WHITE-COOPER, H., MATUNIS, E. L., DINARDO, S., GALENZA, A., O'BRIEN, L. E.,

DOW, J. A. T., SIGN, F. C. A. C. S., JASPER, H., OLIVER, B., PERRIMON, N., DEPLANCKE, B., QUAKE, S. R., LUO, L., AERTS, S., AGARWAL, D., AHMED-BRAIMAH, Y., ARBEITMAN, M., ARISS, M. M., AUGSBURGER, J., AYUSH, K., BAKER, C. C., BANISCH, T., BIRKER, K., BODMER, R., BOLIVAL, B., BRANTLEY, S. E., BRILL, J. A., BROWN, N. C., BUEHNER, N. A., CAI, X. T., CARDOSO-FIGUEIREDO, R., CASARES, F., CHANG, A., CLANDININ, T. R., CRASTA, S., DESPLAN, C., DETWEILER, A. M., DHAKAN, D. B., DONA, E., ENGERT, S., FLOC'HLAY, S., GEORGE, N., GONZALEZ-SEGARRA, A. J., GROVES, A. K., GUMBIN, S., GUO, Y., HARRIS, D. E., HEIFETZ, Y., HOLTZ, S. L., HORNS, F., HUDRY, B., HUNG, R. J., JAN, Y. N., JASZCZAK, J. S., JEFFERIS, G., KARKANIAS, J., KARR, T. L., KATHEDER, N. S., KEZOS, J., KIM, A. A., KIM, S. K., KOCKEL, L., KONSTANTINIDES, N., KORNBERG, T. B., KRAUSE, H. M., LABOTT, A. T., LATURNEY, M., LEHMANN, R., LEINWAND, S., LI, J., LI, J. S. S., *et al*. 2022. Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science,* 375**,** eabk2432.

LIETO, M., ROCA, A., SANTORELLI, F. M., FICO, T., DE MICHELE, G., BELLOFATTO, M., SACCA, F., DE MICHELE, G. & FILLA, A. 2019. Degenerative and acquired sporadic adult onset ataxia. *Neurol Sci,* 40**,** 1335-1342.

LIU, F., GUO, H., OU, M., HOU, X., SUN, G., GONG, W., JING, H., TAN, Q., XUE, W., DAI, Y. & SUI, W. 2016. ARHGAP4 mutated in a Chinese intellectually challenged family. *Gene,* 578**,** 205-9.

LIU, X., GAUBITZ, C., PAJAK, J. & KELCH, B. A. 2022. A second DNA binding site on RFC facilitates clamp loading at gapped or nicked DNA. *Elife,* 11.

LOPEZ DEL AMO, V., PALOMINO-SCHATZLEIN, M., SECO-CERVERA, M., GARCIA-GIMENEZ, J. L., PALLARDO, F. V., PINEDA-LUCENA, A. & GALINDO, M. I. 2017. A Drosophila model of GDAP1 function reveals the involvement of insulin signalling in the mitochondria-dependent neuromuscular degeneration. *Biochim Biophys Acta Mol Basis Dis,* 1863**,** 801-809.

LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol,* 15**,** 550.

LUNARDI, P., SACHSER, R. M., SIERRA, R. O., PEDRAZA, L. K., MEDINA, C., DE LA FUENTE, V., ROMANO, A., QUILLFELDT, J. A. & DE OLIVEIRA ALVARES, L. 2018. Effects of Hippocampal LIMK Inhibition on Memory Acquisition, Consolidation, Retrieval, Reconsolidation, and Extinction. *Mol Neurobiol,* 55**,** 958-967.

MAJKA, J. & BURGERS, P. M. 2004. The PCNA-RFC families of DNA clamps and clamp loaders. *Prog Nucleic Acid Res Mol Biol,* 78**,** 227-60.

MALTBY, C. J., KRANS, A., GRUDZIEN, S. J., PALACIOS, Y., MUINOS, J., SUAREZ, A., ASHER, M., WILLEY, S., VAN DEYNZE, K., MUMM, C., BOYLE, A. P., CORTESE, A., NDAYISABA, A., KHURANA, V., BARMADA, S. J., DIJKSTRA, A. A. & TODD, P. K. 2024. AAGGG repeat expansions trigger RFC1-independent synaptic dysregulation in human CANVAS neurons. *Sci Adv,* 10**,** eadn2321.

MANTO, M., GANDINI, J., FEIL, K. & STRUPP, M. 2020. Cerebellar ataxias: an update. *Curr Opin Neurol,* 33**,** 150-160.

MARCEAUX, C., PETIT, D., BERTOGLIO, J. & DAVID, M. D. 2018. Phosphorylation of ARHGAP19 by CDK1 and ROCK regulates its subcellular localization and function during mitosis. *J Cell Sci,* 131.

MATSELL, E., ANDERSEN, J. P. & MOLDAY, R. S. 2024. Functional and in silico analysis of ATP8A2 and other P4-ATPase variants associated with human genetic diseases. *Dis Model Mech,* 17.

MCGUIRE, S. E., ROMAN, G. & DAVIS, R. L. 2004. Gene expression systems in Drosophila: a synthesis of time and space. *Trends Genet,* 20**,** 384-91.
MENG, Y., TAKAHASHI, H., MENG, J., ZHANG, Y., LU, G., ASRAR, S., NAKAMURA, T. & JIA, Z. 2004. Regulation of ADF/cofilin phosphorylation and synaptic function by LIM-kinase. *Neuropharmacology,* 47**,** 746-54.
MIGLIACCIO, A. A., HALMAGYI, G. M., MCGARVIE, L. A. & CREMER, P. D. 2004. Cerebellar ataxia with bilateral vestibulopathy: description of a syndrome and its characteristic clinical sign. *Brain,* 127**,** 280-93.
MISHRA, A., FERRARI, R., HEUTINK, P., HARDY, J., PIJNENBURG, Y., POSTHUMA, D. & INTERNATIONAL, F. T. D. G. C. 2017. Gene-based association studies report genetic links for clinical subtypes of frontotemporal dementia. *Brain,* 140**,** 1437-1446.
MITCHEL, M. W., MORENO-DE-LUCA, D., MYERS, S. M., LEVY, R. V., TURNER, S., LEDBETTER, D. H. & MARTIN, C. L. 1993. 17q12 Recurrent Deletion Syndrome. *In:* ADAM, M. P., FELDMAN, J., MIRZAA, G. M., PAGON, R. A., WALLACE, S. E., BEAN, L. J. H., GRIPP, K. W. & AMEMIYA, A. (eds.) *GeneReviews((R)).* Seattle (WA).
MITSUHASHI, S. & MATSUMOTO, N. 2020. Long-read sequencing for rare human genetic diseases. *J Hum Genet,* 65**,** 11-19.
MORAES, D. B. V., CORADINE, T. L. C., SILVA, E. V. L., SOBREIRA-NETO, M. A., MARQUES, W., JR., GITAI, L. L. G. & TUMAS, V. 2024. Genetic Epidemiology and Clinical Characteristics of Patients with Spinocerebellar Ataxias in an Unexplored Brazilian State, Using Strategies for Resource-Limited Settings. *Cerebellum,* 23**,** 609-619.
MORATO TORRES, C. A., ZAFAR, F., TSAI, Y. C., VAZQUEZ, J. P., GALLAGHER, M. D., MCLAUGHLIN, I., HONG, K., LAI, J., LEE, J., CHIRINO-PEREZ, A., ROMERO-MOLINA, A. O., TORRES, F., FERNANDEZ-RUIZ, J., ASHIZAWA, T., ZIEGLE, J., JIMENEZ GIL, F. J. & SCHULE, B. 2022. ATTCT and ATTCC repeat expansions in the ATXN10 gene affect disease penetrance of spinocerebellar ataxia type 10. *HGG Adv,* 3**,** 100137.
MORENA, J., GUPTA, A. & HOYLE, J. C. 2019. Charcot-Marie-Tooth: From Molecules to Therapy. *Int J Mol Sci,* 20.
MORGAN, T. H. 1910. Sex Limited Inheritance in Drosophila. *Science,* 32**,** 120-2.
MUZAIMI, M. B., THOMAS, J., PALMER-SMITH, S., ROSSER, L., HARPER, P. S., WILES, C. M., RAVINE, D. & ROBERTSON, N. P. 2004. Population based study of late onset cerebellar ataxia in south east Wales. *J Neurol Neurosurg Psychiatry,* 75**,** 1129-34.
NAKAMURA, H., DOI, H., MITSUHASHI, S., MIYATAKE, S., KATOH, K., FRITH, M. C., ASANO, T., KUDO, Y., IKEDA, T., KUBOTA, S., KUNII, M., KITAZAWA, Y., TADA, M., OKAMOTO, M., JOKI, H., TAKEUCHI, H., MATSUMOTO, N. & TANAKA, F. 2020. Long-read sequencing identifies the pathogenic nucleotide repeat expansion in RFC1 in a Japanese case of CANVAS. *J Hum Genet,* 65**,** 475-480.
NAKAMURA, T., ARIMA-YOSHIDA, F., SAKAUE, F., NASU-NISHIMURA, Y., TAKEDA, Y., MATSUURA, K., AKSHOOMOFF, N., MATTSON, S. N., GROSSFELD, P. D., MANABE, T. & AKIYAMA, T. 2016. PX-RICS-deficient mice mimic autism spectrum disorder in Jacobsen syndrome through impaired GABAA receptor trafficking. *Nat Commun,* 7**,** 10861.
NAKAZAWA, T., HASHIMOTO, R., SAKOORI, K., SUGAYA, Y., TANIMURA, A., HASHIMOTODANI, Y., OHI, K., YAMAMORI, H., YASUDA, Y., UMEDA-YANO, S., KIYAMA, Y., KONNO, K., INOUE, T., YOKOYAMA, K., INOUE, T., NUMATA, S., OHNUMA, T., IWATA, N., OZAKI, N., HASHIMOTO, H.,

WATANABE, M., MANABE, T., YAMAMOTO, T., TAKEDA, M. & KANO, M. 2016. Emerging roles of ARHGAP33 in intracellular trafficking of TrkB and pathophysiology of neuropsychiatric disorders. *Nat Commun,* 7**,** 10594.

NAM, S. H., HONG, Y. B., HYUN, Y. S., NAM DA, E., KWAK, G., HWANG, S. H., CHOI, B. O. & CHUNG, K. W. 2016. Identification of Genetic Causes of Inherited Peripheral Neuropathies by Targeted Gene Panel Sequencing. *Mol Cells,* 39**,** 382-8.

NGUENGANG WAKAP, S., LAMBERT, D. M., OLRY, A., RODWELL, C., GUEYDAN, C., LANNEAU, V., MURPHY, D., LE CAM, Y. & RATH, A. 2020. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet,* 28**,** 165-173.

NOWAK, F. V. 2018. Porf-2 = Arhgap39 = Vilse: A Pivotal Role in Neurodevelopment, Learning and Memory. *eNeuro,* 5.

OFER, N., WEISMAN-SHOMER, P., SHKLOVER, J. & FRY, M. 2009. The quadruplex r(CGG)n destabilizing cationic porphyrin TMPyP4 cooperates with hnRNPs to increase the translation efficiency of fragile X premutation mRNA. *Nucleic Acids Res,* 37**,** 2712-22.

OH, S. B., CHO, S., KIM, H. J. & KIM, S. J. 2024. Differential expression of the enzymes regulating myosin light chain phosphorylation are responsible for the slower relaxation of pulmonary artery than mesenteric artery in rats. *Korean J Physiol Pharmacol,* 28**,** 49-57.

OTERO, B. A., POUKALOV, K., HILDEBRANDT, R. P., THORNTON, C. A., JINNAI, K., FUJIMURA, H., KIMURA, T., HAGERMAN, K. A., SAMPSON, J. B., DAY, J. W. & WANG, E. T. 2021. Transcriptome alterations in myotonic dystrophy frontal cortex. *Cell Rep,* 34**,** 108634.

PAGNAMENTA, A. T., KAIYRZHANOV, R., ZOU, Y., DA'AS, S. I., MAROOFIAN, R., DONKERVOORT, S., DOMINIK, N., LAUFFER, M., FERLA, M. P., ORIOLI, A., GIESS, A., TUCCI, A., BEETZ, C., SEDGHI, M., ANSARI, B., BARRESI, R., BASIRI, K., CORTESE, A., ELGAR, G., FERNANDEZ-GARCIA, M. A., YIP, J., FOLEY, A. R., GUTOWSKI, N., JUNGBLUTH, H., LASSCHE, S., LAVIN, T., MARCELIS, C., MARKS, P., MARINI-BETTOLO, C., MEDNE, L., MOSLEMI, A. R., SARKOZY, A., REILLY, M. M., MUNTONI, F., MILLAN, F., MURARESKU, C. C., NEED, A. C., NEMETH, A. H., NEUHAUS, S. B., NORWOOD, F., O'DONNELL, M., O'DRISCOLL, M., RANKIN, J., YUM, S. W., ZOLKIPLI-CUNNINGHAM, Z., BRUSIUS, I., WUNDERLICH, G., GENOMICS ENGLAND RESEARCH, C., KARAKAYA, M., WIRTH, B., FAKHRO, K. A., TAJSHARGHI, H., BONNEMANN, C. G., TAYLOR, J. C. & HOULDEN, H. 2021. An ancestral 10-bp repeat expansion in VWA1 causes recessive hereditary motor neuropathy. *Brain,* 144**,** 584-600.

PALMA, J. A., NORCLIFFE-KAUFMANN, L. & KAUFMANN, H. 2018. Diagnosis of multiple system atrophy. *Auton Neurosci,* 211**,** 15-25.

PANOSYAN, F. B., LAURA, M., ROSSOR, A. M., PISCIOTTA, C., PISCOSQUITO, G., BURNS, J., LI, J., YUM, S. W., LEWIS, R. A., DAY, J., HORVATH, R., HERRMANN, D. N., SHY, M. E., PAREYSON, D., REILLY, M. M., SCHERER, S. S. & INHERITED NEUROPATHIES CONSORTIUM-RARE DISEASES CLINICAL RESEARCH, N. 2017. Cross-sectional analysis of a large cohort with X-linked Charcot-Marie-Tooth disease (CMTX1). *Neurology,* 89**,** 927-935.

PAYNE, A., HOLMES, N., CLARKE, T., MUNRO, R., DEBEBE, B. J. & LOOSE, M. 2021. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat Biotechnol,* 39**,** 442-450.

PELLERIN, D., DANZI, M. C., WILKE, C., RENAUD, M., FAZAL, S., DICAIRE, M. J., SCRIBA, C. K., ASHTON, C., YANICK, C., BEIJER, D., REBELO, A., ROCCA, C., JAUNMUKTANE, Z., SONNEN, J. A., LARIVIERE, R., GENIS, D., MOLINA PORCEL, L., CHOQUET, K., SAKALLA, R., PROVOST, S., ROBERTSON, R., ALLARD-CHAMARD, X., TETREAULT, M., REILING, S. J., NAGY, S., NISHADHAM, V., PURUSHOTTAM, M., VENGALIL, S., BARDHAN, M., NALINI, A., CHEN, Z., MATHIEU, J., MASSIE, R., CHALK, C. H., LAFONTAINE, A. L., EVOY, F., RIOUX, M. F., RAGOUSSIS, J., BOYCOTT, K. M., DUBE, M. P., DUQUETTE, A., HOULDEN, H., RAVENSCROFT, G., LAING, N. G., LAMONT, P. J., SAPORTA, M. A., SCHULE, R., SCHOLS, L., LA PIANA, R., SYNOFZIK, M., ZUCHNER, S. & BRAIS, B. 2023. Deep Intronic FGF14 GAA Repeat Expansion in Late-Onset Cerebellar Ataxia. *N Engl J Med,* 388**,** 128-141.

PELLERIN, D., DANZI, M. C., WILKE, C., RENAUD, M., FAZAL, S., DICAIRE, M. J., SCRIBA, C. K., ASHTON, C., YANICK, C., BEIJER, D., REBELO, A., ROCCA, C., JAUNMUKTANE, Z., SONNEN, J. A., LARIVIERE, R., GENIS, D., MOLINA PORCEL, L., CHOQUET, K., SAKALLA, R., PROVOST, S., ROBERTSON, R., ALLARD-CHAMARD, X., TETREAULT, M., REILING, S. J., NAGY, S., NISHADHAM, V., PURUSHOTTAM, M., VENGALIL, S., BARDHAN, M., NALINI, A., CHEN, Z., MATHIEU, J., MASSIE, R., CHALK, C. H., LAFONTAINE, A. L., EVOY, F., RIOUX, M. F., RAGOUSSIS, J., BOYCOTT, K. M., DUBE, M. P., DUQUETTE, A., HOULDEN, H., RAVENSCROFT, G., LAING, N. G., LAMONT, P. J., SAPORTA, M. A., SCHULE, R., SCHOLS, L., LA PIANA, R., SYNOFZIK, M., ZUCHNER, S. & BRAIS, B. 2023. Deep Intronic FGF14 GAA Repeat Expansion in Late-Onset Cerebellar Ataxia. *N Engl J Med,* 388**,** 128-141.

PEREIRA, L., MUTESA, L., TINDANA, P. & RAMSAY, M. 2021. African genetic diversity and adaptation inform a precision medicine agenda. *Nat Rev Genet,* 22**,** 284-306.

PERIC, S., PESOVIC, J., SAVIC-PAVICEVIC, D., RAKOCEVIC STOJANOVIC, V. & MEOLA, G. 2021. Molecular and Clinical Implications of Variant Repeats in Myotonic Dystrophy Type 1. *Int J Mol Sci,* 23.

DI PIETRO, F., OSSWALD, M., DE LAS HERAS, J. M., CRISTO, I., LOPEZ-GAY, J., WANG, Z., PELLETIER, S., GAUGUE, I., LEROY, A., MARTIN, C., MORAIS-DE-SA, E. & BELLAICHE, Y. 2023. Systematic analysis of RhoGEF/GAP localizations uncovers regulators of mechanosensing and junction formation during epithelial cell division. *Curr Biol,* 33**,** 858-874 e7.

PIPIS, M., ROSSOR, A. M., LAURA, M. & REILLY, M. M. 2019. Next-generation sequencing in Charcot-Marie-Tooth disease: opportunities and challenges. *Nat Rev Neurol,* 15**,** 644-656.

PISCIOTTA, C. & SHY, M. E. 2018. Neuropathy. *Handb Clin Neurol,* 148**,** 653-665.

PISCIOTTA, C. & SHY, M. E. 2023. Hereditary neuropathy. *Handb Clin Neurol,* 195**,** 609-617.

POPAT, K. C., LEARY SWAN, E. E., MUKHATYAR, V., CHATVANICHKUL, K. I., MOR, G. K., GRIMES, C. A. & DESAI, T. A. 2005. Influence of nanoporous alumina membranes on long-term osteoblast response. *Biomaterials,* 26**,** 4516-22.

PORUBSKY, D. & EICHLER, E. E. 2024. A 25-year odyssey of genomic technology advances and structural variant discovery. *Cell,* 187**,** 1024-1037.

QUINODOZ, M., PETER, V. G., BEDONI, N., ROYER BERTRAND, B., CISAROVA, K., SALMANINEJAD, A., SEPAHI, N., RODRIGUES, R., PIRAN, M., MOJARRAD, M., PASDAR, A., GHANBARI ASAD, A., SOUSA, A. B., COUTINHO SANTOS, L., SUPERTI-FURGA, A. & RIVOLTA, C. 2021. AutoMap is a high performance

homozygosity mapping tool using next-generation sequencing data. *Nat Commun,* 12**,** 518.

RAFEHI, H., READ, J., SZMULEWICZ, D. J., DAVIES, K. C., SNELL, P., FEARNLEY, L. G., SCOTT, L., THOMSEN, M., GILLIES, G., POPE, K., BENNETT, M. F., MUNRO, J. E., NGO, K. J., CHEN, L., WALLIS, M. J., BUTLER, E. G., KUMAR, K. R., WU, K. H., TOMLINSON, S. E., TISCH, S., MALHOTRA, A., LEE-ARCHER, M., DOLZHENKO, E., EBERLE, M. A., ROBERTS, L. J., FOGEL, B. L., BRUGGEMANN, N., LOHMANN, K., DELATYCKI, M. B., BAHLO, M. & LOCKHART, P. J. 2023. An intronic GAA repeat expansion in FGF14 causes the autosomal-dominant adult-onset ataxia SCA27B/ATX-FGF14. *Am J Hum Genet,* 110**,** 1018.

RAFEHI, H., SZMULEWICZ, D. J., BENNETT, M. F., SOBREIRA, N. L. M., POPE, K., SMITH, K. R., GILLIES, G., DIAKUMIS, P., DOLZHENKO, E., EBERLE, M. A., BARCINA, M. G., BREEN, D. P., CHANCELLOR, A. M., CREMER, P. D., DELATYCKI, M. B., FOGEL, B. L., HACKETT, A., HALMAGYI, G. M., KAPETANOVIC, S., LANG, A., MOSSMAN, S., MU, W., PATRIKIOS, P., PERLMAN, S. L., ROSEMERGY, I., STOREY, E., WATSON, S. R. D., WILSON, M. A., ZEE, D. S., VALLE, D., AMOR, D. J., BAHLO, M. & LOCKHART, P. J. 2019. Bioinformatics-Based Identification of Expanded Repeats: A Non-reference Intronic Pentamer Expansion in RFC1 Causes CANVAS. *Am J Hum Genet,* 105**,** 151-165.

RAVEL-CHAPUIS, A., DUCHESNE, E. & JASMIN, B. J. 2022. Pharmacological and exercise-induced activation of AMPK as emerging therapies for myotonic dystrophy type 1 patients. *J Physiol,* 600**,** 3249-3264.

REITER, L. T. & BIER, E. 2002. Using Drosophila melanogaster to uncover human disease gene function and potential drug target proteins. *Expert Opin Ther Targets,* 6**,** 387-99.

RICHARDS, S., AZIZ, N., BALE, S., BICK, D., DAS, S., GASTIER-FOSTER, J., GRODY, W. W., HEGDE, M., LYON, E., SPECTOR, E., VOELKERDING, K., REHM, H. L. & COMMITTEE, A. L. Q. A. 2015. 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. *Genet Med,* 17**,** 405-24.

RIDLEY, R. M., FRITH, C. D., FARRER, L. A. & CONNEALLY, P. M. 1991. Patterns of inheritance of the symptoms of Huntington's disease suggestive of an effect of genomic imprinting. *J Med Genet,* 28**,** 224-31.

RIZIG, M., BANDRES-CIGA, S., MAKARIOUS, M. B., OJO, O. O., CREA, P. W., ABIODUN, O. V., LEVINE, K. S., ABUBAKAR, S. A., ACHORU, C. O., VITALE, D., ADENIJI, O. A., AGABI, O. P., KORETSKY, M. J., AGULANNA, U., HALL, D. A., AKINYEMI, R. O., XIE, T., ALI, M. W., SHAMIM, E. A., ANI-OSHEKU, I., PADMANABAN, M., ARIGBODI, O. M., STANDAERT, D. G., BELLO, A. H., DEAN, M. N., ERAMEH, C. O., ELSAYED, I., FAROMBI, T. H., OKUNOYE, O., FAWALE, M. B., BILLINGSLEY, K. J., IMARHIAGBE, F. A., JEREZ, P. A., IWUOZO, E. U., BAKER, B., KOMOLAFE, M. A., MALIK, L., NWANI, P. O., DAIDA, K., NWAZOR, E. O., MIANO-BURKHARDT, A., NYANDAITI, Y. W., FANG, Z. H., OBIABO, Y. O., KLUSS, J. H., ODENIYI, O. A., HERNANDEZ, D. G., ODIASE, F. E., TAYEBI, N., OJINI, F. I., SIDRANKSY, E., ONWUEGBUZIE, G. A., D'SOUZA, A. M., OSAIGBOVO, G. O., BERHE, B., OSEMWEGIE, N., REED, X., OSHINAIKE, O. O., LEONARD, H. L., OTUBOGUN, F. M., ALVARADO, C. X., OYAKHIRE, S. I., OZOMMA, S. I., SAMUEL, S. C., TAIWO, F. T., WAHAB, K. W., ZUBAIR, Y. A., IWAKI, H., KIM, J. J., MORRIS, H. R.,

HARDY, J., NALLS, M. A., HEILBRON, K., NORCLIFFE-KAUFMANN, L., NIGERIA PARKINSON DISEASE RESEARCH, N., INTERNATIONAL PARKINSON'S DISEASE GENOMICS CONSORTIUM, A., BLACK, AFRICAN AMERICAN CONNECTIONS TO PARKINSON'S DISEASE STUDY, G., ANDME RESEARCH, T., BLAUWENDRAAT, C., HOULDEN, H., SINGLETON, A., OKUBADEJO, N. U. & GLOBAL PARKINSON'S GENETICS, P. 2023. Identification of genetic risk loci and causal insights associated with Parkinson's disease in African and African admixed populations: a genome-wide association study. *Lancet Neurol,* 22**,** 1015-1025.

RIZZO, F., BONO, S., RUEPP, M. D., SALANI, S., OTTOBONI, L., ABATI, E., MELZI, V., CORDIGLIERI, C., PAGLIARANI, S., DE GIOIA, R., ANASTASIA, A., TAIANA, M., GARBELLINI, M., LODATO, S., KUNDERFRANCO, P., CAZZATO, D., CARTELLI, D., LONATI, C., BRESOLIN, N., COMI, G., NIZZARDO, M. & CORTI, S. 2023. Combined RNA interference and gene replacement therapy targeting MFN2 as proof of principle for the treatment of Charcot-Marie-Tooth type 2A. *Cell Mol Life Sci,* 80**,** 373.

RODDEN, L. N., RUMMEY, C., KESSLER, S., WILSON, R. B. & LYNCH, D. R. 2023. A Novel Metric for Predicting Severity of Disease Features in Friedreich's Ataxia. *Mov Disord,* 38**,** 970-977.

RONCO, R., PERINI, C., CURRO, R., DOMINIK, N., FACCHINI, S., GENNARI, A., SIMONE, R., STUART, S., NAGY, S., VEGEZZI, E., QUARTESAN, I., EL-SADDIG, A., LAVIN, T., TUCCI, A., SZYMURA, A., NOVIS DE FARIAS, L. E., GARY, A., DELFELD, M., KANDIKATLA, P., NIU, N., TAWDE, S., SHAW, J., POLKE, J., REILLY, M. M., WOOD, N. W., CRESPAN, E., GOMEZ, C., CHEN, J. Y. H., SCHMAHMANN, J. D., GOSAL, D., HOULDEN, H., DAS, S. & CORTESE, A. 2023. Truncating Variants in RFC1 in Cerebellar Ataxia, Neuropathy, and Vestibular Areflexia Syndrome. *Neurology,* 100**,** e543-e554.

ROOSTAEI, T., NAZERI, A., SAHRAIAN, M. A. & MINAGAR, A. 2014. The human cerebellum: a review of physiologic neuroanatomy. *Neurol Clin,* 32**,** 859-69.

RUANO, L., MELO, C., SILVA, M. C. & COUTINHO, P. 2014. The global epidemiology of hereditary ataxia and spastic paraplegia: a systematic review of prevalence studies. *Neuroepidemiology,* 42**,** 174-83.

RUBIN, G. M., YANDELL, M. D., WORTMAN, J. R., GABOR MIKLOS, G. L., NELSON, C. R., HARIHARAN, I. K., FORTINI, M. E., LI, P. W., APWEILER, R., FLEISCHMANN, W., CHERRY, J. M., HENIKOFF, S., SKUPSKI, M. P., MISRA, S., ASHBURNER, M., BIRNEY, E., BOGUSKI, M. S., BRODY, T., BROKSTEIN, P., CELNIKER, S. E., CHERVITZ, S. A., COATES, D., CRAVCHIK, A., GABRIELIAN, A., GALLE, R. F., GELBART, W. M., GEORGE, R. A., GOLDSTEIN, L. S., GONG, F., GUAN, P., HARRIS, N. L., HAY, B. A., HOSKINS, R. A., LI, J., LI, Z., HYNES, R. O., JONES, S. J., KUEHL, P. M., LEMAITRE, B., LITTLETON, J. T., MORRISON, D. K., MUNGALL, C., O'FARRELL, P. H., PICKERAL, O. K., SHUE, C., VOSSHALL, L. B., ZHANG, J., ZHAO, Q., ZHENG, X. H. & LEWIS, S. 2000. Comparative genomics of the eukaryotes. *Science,* 287**,** 2204-15.

RUIZ DE SABANDO, A., CIOSI, M., GALBETE, A., CUMMING, S. A., SPANISH, H. D. C. G., MONCKTON, D. G. & RAMOS-ARROYO, M. A. 2024. Somatic CAG repeat instability in intermediate alleles of the HTT gene and its potential association with a clinical phenotype. *Eur J Hum Genet,* 32**,** 770-778.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74**,** 5463-7.

SANTORO, L., PERRETTI, A., LANZILLO, B., COPPOLA, G., DE JOANNA, G., MANGANELLI, F., COCOZZA, S., DE MICHELE, G., FILLA, A. & CARUSO, G. 2000. Influence of GAA expansion size and disease duration on central nervous system impairment in Friedreich's ataxia: contribution to the understanding of the pathophysiology of the disease. *Clin Neurophysiol,* 111**,** 1023-30.

SCRIBA, C. K., BEECROFT, S. J., CLAYTON, J. S., CORTESE, A., SULLIVAN, R., YAU, W. Y., DOMINIK, N., RODRIGUES, M., WALKER, E., DYER, Z., WU, T. Y., DAVIS, M. R., CHANDLER, D. C., WEISBURD, B., HOULDEN, H., REILLY, M. M., LAING, N. G., LAMONT, P. J., ROXBURGH, R. H. & RAVENSCROFT, G. 2020. A novel RFC1 repeat motif (ACAGG) in two Asia-Pacific CANVAS families. *Brain,* 143**,** 2904-2910.

SEKIGUCHI, M., SOBUE, A., KUSHIMA, I., WANG, C., ARIOKA, Y., KATO, H., KODAMA, A., KUBO, H., ITO, N., SAWAHATA, M., HADA, K., IKEDA, R., SHINNO, M., MIZUKOSHI, C., TSUJIMURA, K., YOSHIMI, A., ISHIZUKA, K., TAKASAKI, Y., KIMURA, H., XING, J., YU, Y., YAMAMOTO, M., OKADA, T., SHISHIDO, E., INADA, T., NAKATOCHI, M., TAKANO, T., KURODA, K., AMANO, M., ALEKSIC, B., YAMOMOTO, T., SAKUMA, T., AIDA, T., TANAKA, K., HASHIMOTO, R., ARAI, M., IKEDA, M., IWATA, N., SHIMAMURA, T., NAGAI, T., NABESHIMA, T., KAIBUCHI, K., YAMADA, K., MORI, D. & OZAKI, N. 2020. ARHGAP10, which encodes Rho GTPase-activating protein 10, is a novel gene for schizophrenia risk. *Transl Psychiatry,* 10**,** 247.

SHARPE, J. L., HARPER, N. S., GARNER, D. R. & WEST, R. J. H. 2021. Modeling C9orf72-Related Frontotemporal Dementia and Amyotrophic Lateral Sclerosis in Drosophila. *Front Cell Neurosci,* 15**,** 770937.

SHINSATO, R. N., CORREA, C. G. & HERAI, R. H. 2024. Genetic network analysis indicate that individuals affected by neurodevelopmental conditions have genetic variations associated with ophthalmologic alterations: A critical review of literature. *Gene,* 908**,** 148246.

SHY, M. E., JANI, A., KRAJEWSKI, K., GRANDIS, M., LEWIS, R. A., LI, J., SHY, R. R., BALSAMO, J., LILIEN, J., GARBERN, J. Y. & KAMHOLZ, J. 2004. Phenotypic clustering in MPZ mutations. *Brain,* 127**,** 371-84.

SOBREIRA, N., SCHIETTECATTE, F., VALLE, D. & HAMOSH, A. 2015. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat,* 36**,** 928-30.

STAUS, D. P., BLAKER, A. L., MEDLIN, M. D., TAYLOR, J. M. & MACK, C. P. 2011. Formin homology domain-containing protein 1 regulates smooth muscle cell phenotype. *Arterioscler Thromb Vasc Biol,* 31**,** 360-7.

STEVANOVSKI, I., CHINTALAPHANI, S. R., GAMAARACHCHI, H., FERGUSON, J. M., PINEDA, S. S., SCRIBA, C. K., TCHAN, M., FUNG, V., NG, K., CORTESE, A., HOULDEN, H., DOBSON-STONE, C., FITZPATRICK, L., HALLIDAY, G., RAVENSCROFT, G., DAVIS, M. R., LAING, N. G., FELLNER, A., KENNERSON, M., KUMAR, K. R. & DEVESON, I. W. 2022. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Sci Adv,* 8**,** eabm5386.

STOCKER, H. & GALLANT, P. 2008. Getting started : an overview on raising and handling Drosophila. *Methods Mol Biol,* 420**,** 27-44.

SULEK, A., HOFFMAN-ZACHARSKA, D., BEDNARSKA-MAKARUK, M., SZIRKOWIEC, W. & ZAREMBA, J. 2004. Polymorphism of trinucleotide repeats in non-translated regions of SCA8 and SCA12 genes: allele distribution in a Polish control group. *J Appl Genet,* 45**,** 101-5.

SUN, M., JOHNSON, A. K., NELAKUDITI, V., GUIDUGLI, L., FISCHER, D., ARNDT, K., MA, L., SANDFORD, E., SHAKKOTTAI, V., BOYCOTT, K., WARMAN-CHARDON, J., LI, Z., DEL GAUDIO, D., BURMEISTER, M., GOMEZ, C. M., WAGGONER, D. J. & DAS, S. 2019. Targeted exome analysis identifies the genetic basis of disease in over 50% of patients with a wide range of ataxia-related phenotypes. *Genet Med,* 21**,** 195-206.

SWINNEN, B., ROBBERECHT, W. & VAN DEN BOSCH, L. 2020. RNA toxicity in non-coding repeat expansion disorders. *EMBO J,* 39**,** e101112.

SYNOFZIK, M. & NEMETH, A. H. 2018. Recessive ataxias. *Handb Clin Neurol,* 155**,** 73-89.

SZMULEWICZ, D. J., WATERSTON, J. A., MACDOUGALL, H. G., MOSSMAN, S., CHANCELLOR, A. M., MCLEAN, C. A., MERCHANT, S., PATRIKIOS, P., HALMAGYI, G. M. & STOREY, E. 2011. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome (CANVAS): a review of the clinical features and video-oculographic diagnosis. *Ann N Y Acad Sci,* 1233**,** 139-47.

TANAKA, R., LIAO, J., HADA, K., MORI, D., NAGAI, T., MATSUZAKI, T., NABESHIMA, T., KAIBUCHI, K., OZAKI, N., MIZOGUCHI, H. & YAMADA, K. 2023. Inhibition of Rho-kinase ameliorates decreased spine density in the medial prefrontal cortex and methamphetamine-induced cognitive dysfunction in mice carrying schizophrenia-associated mutations of the Arhgap10 gene. *Pharmacol Res,* 187**,** 106589.

TSUCHIYA, M., NAN, H., KOH, K., ICHINOSE, Y., GAO, L., SHIMOZONO, K., HATA, T., KIM, Y. J., OHTSUKA, T., CORTESE, A. & TAKIYAMA, Y. 2020. RFC1 repeat expansion in Japanese patients with late-onset cerebellar ataxia. *J Hum Genet,* 65**,** 1143-1147.

TOMASELLI, P. J., ROSSOR, A. M., HORGA, A., JAUNMUKTANE, Z., CARR, A., SAVERI, P., PISCOSQUITO, G., PAREYSON, D., LAURA, M., BLAKE, J. C., POH, R., POLKE, J., HOULDEN, H. & REILLY, M. M. 2017. Mutations in noncoding regions of GJB1 are a major cause of X-linked CMT. *Neurology,* 88**,** 1445-1453.

TRASCHUTZ, A., ADARMES-GOMEZ, A. D., ANHEIM, M., BAETS, J., FALKENBURGER, B. H., GBUREK-AUGUSTAT, J., DOSS, S., KAMM, C., KLIVENYI, P., GROBE-EINSLER, M., KLOPSTOCK, T., MINNEROP, M., MUNCHAU, A., PANE, C., RENAUD, M., SANTORELLI, F. M., SCHOLS, L., TIMMANN, D., VIELHABER, S., HAACK, T. B., VAN DE WARRENBURG, B. P., ZANNI, G. & SYNOFZIK, M. 2023. Autosomal Recessive Cerebellar Ataxias in Europe: Frequency, Onset, and Severity in 677 Patients. *Mov Disord,* 38**,** 1109-1112.

TRIOLO, D., DINA, G., TAVEGGIA, C., VACCARI, I., PORRELLO, E., RIVELLINI, C., DOMI, T., LA MARCA, R., CERRI, F., BOLINO, A., QUATTRINI, A. & PREVITALI, S. C. 2012. Vimentin regulates peripheral nerve myelination. *Development,* 139**,** 1359-67.

TSUCHIYA, M., NAN, H., KOH, K., ICHINOSE, Y., GAO, L., SHIMOZONO, K., HATA, T., KIM, Y. J., OHTSUKA, T., CORTESE, A. & TAKIYAMA, Y. 2020. RFC1 repeat expansion in Japanese patients with late-onset cerebellar ataxia. *J Hum Genet,* 65**,** 1143-1147.

ULGEN, E., OZISIK, O. & SEZERMAN, O. U. 2019. pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks. *Front Genet,* 10**,** 858.

VANKAN, P. 2013. Prevalence gradients of Friedreich's ataxia and R1b haplotype in Europe co-localize, suggesting a common Palaeolithic origin in the Franco-Cantabrian ice age refuge. *J Neurochem,* 126 Suppl 1**,** 11-20.

VARSHNEY, D., SPIEGEL, J., ZYNER, K., TANNAHILL, D. & BALASUBRAMANIAN, S. 2020. The regulation and functions of DNA and RNA G-quadruplexes. *Nat Rev Mol Cell Biol,* 21**,** 459-474.

VERHOEVEN, K., DE JONGHE, P., VAN DE PUTTE, T., NELIS, E., ZWIJSEN, A., VERPOORTEN, N., DE VRIENDT, E., JACOBS, A., VAN GERWEN, V., FRANCIS, A., CEUTERICK, C., HUYLEBROECK, D. & TIMMERMAN, V. 2003. Slowed conduction and thin myelination of peripheral nerves associated with mutant rho Guanine-nucleotide exchange factor 10. *Am J Hum Genet,* 73**,** 926-32.

WADA, T., DOI, H., OKUBO, M., TADA, M., UEDA, N., SUZUKI, H., TOMINAGA, W., KOIKE, H., KOMIYA, H., KUBOTA, S., HASHIGUCHI, S., NAKAMURA, H., TAKAHASHI, K., KUNII, M., TANAKA, K., MIYAJI, Y., HIGASHIYAMA, Y., KOSHIMIZU, E., MIYATAKE, S., KATSUNO, M., FUJII, S., TAKAHASHI, H., MATSUMOTO, N., TAKEUCHI, H. & TANAKA, F. 2024. RNA Foci in Two bi-Allelic RFC1 Expansion Carriers. *Ann Neurol,* 95**,** 607-613.

WANG, E., THOMBRE, R., SHAH, Y., LATANICH, R. & WANG, J. 2021. G-Quadruplexes as pathogenic drivers in neurodegenerative disorders. *Nucleic Acids Res,* 49**,** 4816-4830.

WILSON, L. A., MACKEN, W. L., PERRY, L. D., RECORD, C. J., SCHON, K. R., FREZATTI, R. S. S., RAGA, S., NAIDU, K., KOKEN, O. Y., POLAT, I., KAPAPA, M. M., DOMINIK, N., EFTHYMIOU, S., MORSY, H., NEL, M., FASSAD, M. R., GAO, F., PATEL, K., SCHOONEN, M., BISSCHOFF, M., VORSTER, A., JONVIK, H., HUMAN, R., LUBBE, E., NONYANE, M., VENGALIL, S., NASHI, S., SRIVASTAVA, K., LEMMERS, R., REYAZ, A., MISHRA, R., TOPF, A., TRAINOR, C. I., STEYN, E. C., MAHUNGU, A. C., VAN DER VLIET, P. J., CEYLAN, A. C., HIZ, A. S., CAVDARLI, B., SEMERCI GUNDUZ, C. N., CEYLAN, G. G., NAGAPPA, M., TALLAPAKA, K. B., GOVINDARAJ, P., VAN DER MAAREL, S. M., NARAYANAPPA, G., NANDEESH, B. N., WA SOMWE, S., BEARDEN, D. R., KVALSUND, M. P., RAMDHARRY, G. M., OKTAY, Y., YIS, U., TOPALOGLU, H., SARKOZY, A., BUGIARDINI, E., HENNING, F., WILMSHURST, J. M., HECKMANN, J. M., MCFARLAND, R., TAYLOR, R. W., SMUTS, I., VAN DER WESTHUIZEN, F. H., SOBREIRA, C., TOMASELLI, P. J., MARQUES, W., BHATIA, R., DALAL, A., SRIVASTAVA, M. V. P., YAREEDA, S., NALINI, A., VISHNU, V. Y., THANGARAJ, K., STRAUB, V., HORVATH, R., CHINNERY, P. F., PITCEATHLY, R. D. S., MUNTONI, F., HOULDEN, H., VANDROVCOVA, J., REILLY, M. M. & HANNA, M. G. 2023. Neuromuscular disease genetics in underrepresented populations: increasing data diversity. *Brain*.

WU, T. Y., TAYLOR, J. M., KILFOYLE, D. H., SMITH, A. D., MCGUINNESS, B. J., SIMPSON, M. P., WALKER, E. B., BERGIN, P. S., CLELAND, J. C., HUTCHINSON, D. O., ANDERSON, N. E., SNOW, B. J., ANDERSON, T. J., PAERMENTIER, L. A., CUTFIELD, N. J., CHANCELLOR, A. M., MOSSMAN, S. S. & ROXBURGH, R. H. 2014. Autonomic dysfunction is a major feature of cerebellar ataxia, neuropathy, vestibular areflexia 'CANVAS' syndrome. *Brain,* 137**,** 2649-56.

YU, J., LIUFU, T., ZHENG, Y., XU, J., MENG, L., ZHANG, W., YUAN, Y., HONG, D., CHARLET-BERGUERAND, N., WANG, Z. & DENG, J. 2022. CGG repeat expansion in NOTCH2NLC causes mitochondrial dysfunction and progressive

neurodegeneration in Drosophila model. *Proc Natl Acad Sci U S A,* 119**,** e2208649119.
ZHANG, N. & ASHIZAWA, T. 2017. RNA toxicity and foci formation in microsatellite expansion diseases. *Curr Opin Genet Dev,* 44**,** 17-29.
ZSCHOCKE, J., BYERS, P. H. & WILKIE, A. O. M. 2023. Mendelian inheritance revisited: dominance and recessiveness in medical genetics. *Nat Rev Genet,* 24**,** 442-463.

Appendix 1

From Dominik *et al*., 2023

Clinical and demographic features of patients carrying novel pathogenic repeat configurations in *RFC1*

AOO = age of onset; CANVAS = cerebellar ataxia, neuropathy and vestibular areflexia syndrome; DD = disease duration; F = female; M = male; RBD = REM sleep behaviour disorder.

Clinical and demographic features of patients carrying novel pathogenic repeat configurations in *RFC1*

	RFC1 genotype	Sex	Ethnicity	Phenotype	AOO	DD, y	Chronic cough	Cerebellar syndrome	Sensory neuropathy	Bilateral vestibular areflexia	Dysautonomia	Walking aid use (age, y)	Additional features
AAGGC expansion													
Case I-1	Allele 1: $(AAGGG)_{510}(AAGGC)_{880}$ Allele 2: $(AAGGG)_{940}(AAGGC)_{900}$	F	Caucasian (Indian)	CANVAS	24	17	Yes	Yes	Yes	Yes	No	Stick (36)	Cramps, pyramidal signs
Case I-2	Allele 1: $(AAGGG)_n(AAGGC)_n$ Allele 2: $(AAGGG)_n(AAGGC)_n$	F	Caucasian (Indian)	Sensory neuropathy + cough	34	8	Yes	N/A	Yes	N/A	N/A	No	–
AGGGC expansion													
Case II-1	Allele 1: $(AGGGC)_{1240}$ Allele 2: $(AAGGG)_{930}$	M	Mixed (Lebanese)	Sensory neuropathy + vestibular dysfunction	53	11	Yes	No	Yes	Yes	Yes	No	Cramps
Case III-1	Allele 1: $(AGGGC)_{3200}$ Allele 2: $(AAGGG)_{1000}$	M	Caucasian (British)	CANVAS	71	12	Yes	Yes	Yes	N/A	Yes	Wheelchair (81)	Cramps, cognitive/behavioural abnormalities after age 80
Case IV-1	Allele 1: $(AGGGC)_{1875}$/ Allele 2: $(AAGGG)_{500}$	M	Caucasian (Italian)	CANVAS	41	34	No	Yes	Yes	Yes	Yes	Wheelchair (72)	Cramps
Case V-1	Allele 1: $(AGGGC)_n$/ Allele 2: $(AAGGG)_n$	F	Caucasian (Italian)	Sensory neuropathy + cough	60	13	Yes	No	Yes	No	No	No	–
Case V-2	Allele 1: $(AGGGC)_n$/ Allele 2: $(AAGGG)_n$	F	Caucasian (Italian)	Sensory neuropathy	40	20	No	No	Yes	No	No	No	–
Case VI-1	Allele 1: $(AGGGC)_n$/ Allele 2: $(AAGGG)_n$	F	Caucasian (Italian)	Sensory ganglionopathy + cough	62	23	Yes	No	Yes	N/A	Yes	No	Voice and hand tremor, urinary incontinence

AGAGG expansion													
Case VII-1	Allele 1: $(AAAGG)_{470}(AGAGG)_{470}$/ Allele 2: $(AAGGG)_{1140}$	F	Caucasian (British)	CANVAS	50	24	Yes	Yes	Yes	Yes	No	Walker (69), wheelchair (74)	–
AAAGG expansion													
Case VIII-1	Allele 1: $(AAAGG)_{610}(AAGGG)_{390}$/ Allele 2: $(AAGGG)_{1100}$	M	Caucasian (British)	CANVAS	55	20	Yes	Yes	Yes	N/A	Yes	Walker and wheelchair (74)	Cognitive impairment since age 72
Case IX-1	Allele 1: $(AAGGG)_{700}(AAAGG)_{200}$/ Allele 2: $(AAGGG)_{1170}$	M	Caucasian (British)	CANVAS	45	31	Yes	Yes	Yes	Yes	Yes	Walker (75)	RBD, positive DatScan
Case X-1	Allele 1: $(AAAGG)_{980}$/ Allele 2: $(AAGGG)_{1010}$	M	Caucasian (Australian)	CANVAS	58	15	Yes	Yes	Yes	Yes	N/A	N/A	–
Case XI-1	Allele 1: $(AAAGG)_{800}$/ Allele 2: $(AAGGG)_{500}$	F	Caucasian (Italian)	Sensory ganglionopathy + cough	73	3	Yes	No	Yes	No	No	Stick (77)	–
Case XII-1	Allele 1: $(AAAGG)_{600}$/ Allele 2: $(AAGGG)_{390}$	M	Caucasian (Italian)	Sensory ganglionopathy + cough	56	10	Yes	No	Yes	No	No	No	–

Appendix 2

From Dominik*, Efthymiou* *et al*., 2024

Clinical and demographic features of patients carrying biallelic variants in *ARHGAP19*

MP motor predominant * neurophysiology data not seen +/- report available only ** very limited study, Clinical sensory involvement refers to symptoms and/or signs. Δ assessed approximately 2 years after onset of acute left-hand weakness. CS conduction slowing LL lower limb UL upper limb KJ knee jerk, AJ ankle jerk, MRC grade ADF/APF = medical research council power grading of power in ankle dorsiflexion/ankle plantar flexion. ~= approximate (where known). If there is discrepancy between limbs, a comma separates right and left. > greater than, LD length-dependent , M male, F female, Unk unknown, dist distal, prox proximal.

Family	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Case	PT1 (F1-II:1)	PT2 (F2-II:5)	PT3 (F3-II:1)	PT4 (F4-II:1)	PT5 (F5-II:2)	PT6 (F6-II:7)	PT7 (F7-II:1)	PT8 (F8-II:2)	PT9 (F9-II:3)	PT11 (F10-II:1)	PT12 (F11-II:2)	PT13 (F12-II:1)	PT14 (F13-II:3)	PT15 (F14-II:2)	PT17 (F15-II:2)	PT18 (F16-II:1)	PT19 (F17-II:1)	PT20 (F18-II:1)	PT21 (F19-II:1)	PT25 (F21-IV:I)
Nucleotide change	c.261dup	c.585dupA	c.1243C>T	c.1219C>T	c.419G>A	c.203T>C	c.451C>A	c.1A>G	c.452del	c.85A>G	c.932C>G	c.683T>A	c.585dupA	c.451C>A	c.203T>C	c.422T>G	c.717T>A	c.563del	c.683T>A	c.818C>T
Amino acid change	p.Pro88Alafs*43	p.His196Glnfs*9	p.Gln415*	p.Arg407*	p.Gly140Asp	p.Leu68Pro	p.Gln151Lys	p.Met1?	p.Asn160Metfs*21	p.Asn29Asp	p.Pro311Arg	p.Leu228His	p.His196Glnfs*9	p.Gln151Lys	p.Leu68Pro	p.Leu141Trp	p.Asn239Lys	p.Pro188Argfs*5	p.Leu228His	p.Pro273Leu
Sex	M	M	F	F	M	F	F	F	M	M	M	M	F	M	F	F	F	F	M	F
Ethnic Origin	Brazilian	Egypt	Pakistani	n/a	Spain	Turkish	Turkish	Brazilian	Syria	Pakistan	Australia (Italian)	Bangladeshi	Unk	Turkish	Turkish	Iran	Dubai	Pakistan	Afghanistan	Latin
Phenotype	CMT-MP**	CMT2 with CS	dHMN with CS	CMT2-MP	dHMN	dHMN with CS	CMT with CS*	CMT2-MP	CMT2 with CS**	CMT2-MP	CMT2-MP with CS	Unk*	CMT**	dHMN*	CMTi-MP	dHMN**	CMT with CS*	HMN with CS**	CMT2-MP	dHMN*
Consanguinity	Yes	No	Yes	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes
Age of onset (years)	6	15	9	10	10-20	infant	11	11	15	30	13	11m	n/a	11	13	12	8	9	11	2
Age at assessment (years)	8	29	16	19	40s	17	13	48	20	63	15	11m	n/a	n/a	14$^{\Delta}$	15	10	9	16	3.5
Symptom at onset	Falls	Walking difficulty	Falls	Dist LL weakness	Dist LL weakness	Walking difficulty	Dist LL weakness	Walking difficulty	Dist right LL weakness	Leg cramps	Dist LL weakness	Acute UL weakness	Unk	Falls	Acute left-hand weakness	Dist LL weakness	Right leg weakness	Dist right LL weakness	Dist left LL weakness	Dist LL weakness
Asymmetry	Yes	Yes	Yes	No	No	Yes	No	Yes	Yes	No	No	No	Unk	No	Yes	Yes	Yes	Yes	Yes	Yes
Clinical sensory involvement	No	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Unk	No	Yes	Yes	Unk	No	No	Yes
LL Areflexia	No	Brisk KJ, AJ	Yes	Yes	Yes	KJ present	Yes	KJ present	Yes	KJ brisk, AJ present, L Babinski	KJ present	Unk	Unk	Yes	Brisk KJ, AJ present	Yes	Left AJ present	Yes	No	KJ present

Family	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Case	PT1 (F1-II:1)	PT2 (F2-II:5)	PT3 (F3-II:1)	PT4 (F4-II:1)	PT5 (F5-II:2)	PT6 (F6-II:7)	PT7 (F7-II:1)	PT8 (F8-II:2)	PT9 (F9-II:3)	PT11 (F10-II:1)	PT12 (F11-II:2)	PT13 (F12-II:1)	PT14 (F13-II:3)	PT15 (F14-II:2)	PT17 (F15-II:2)	PT18 (F16-II:1)	PT19 (F17-II:1)	PT20 (F18-II:1)	PT21 (F19-II:1)	PT25 (F21-IV:I)
Pattern of muscle weakness (MRC grade ADF/APF)	Dist LL L>R (4,3/4, 4)	Distal LL L>R (1,2)	LD (~3,0/4)	LD ('severe')	LL predominant dist > prox (0/0)	LD (~2,3/4)	LD (4/4)	LD (0/4)	LD (~1,3/5)	LL predominant dist > prox (4/5)	LD (0/3)	UL > LL (Unk)	Unk	?LD (Unk)	Dist UL> LL (5/5)	LD (0,1/1)	Right dist LL only (Unk)	LL predominant dist > prox (~1/?)	Dist LL L>R, (4,0/5)	LD (2/4)
Foot deformity	Pes cavus	Pes planus, hammer toe	Pes cavus	Pes cavus	Nil	Pes cavus	Pes cavus	Pes planus	Pes planus	Nil	Pes cavus	Nil	Unk	Pes cavus	Hammer toe	Pes cavus	Unk	Unk	No	Yes
Gait	Left foot drop	Unk	Steppage	Unk	Steppage	Steppage	Steppage	Steppage	Steppage	Steppage and waddle	Steppage	Unclear	Unk	Steppage	Normal	Steppage	Unk	Steppage	Steppage	Steppage
Muscle atrophy	Dist LL L>R	Distal LL (calf hypertrophy)	Calves down	Dist LL	Calves down	Feet	Feet	Calves down	None	Calves down	Dist LL	Unk	Unk	Dist LL	Hands and dist LL	Dist UL and LL	Unk	Dist LL R>L	Dist LL	Dist LL > UL
Neurophysiology	Limited study: LL motor only. Asymmetric LD motor neuropathy	LL predominant axonal neuropathy with some conduction slowing	LD motor neuropathy with significant conduction slowing and block	LD MP axonal neuropathy	MP LL predominant axonal neuropathy	MP LL predominant axonal neuropathy with some conduction slowing	*Mixed axonal and demyelinating neuropathy (mainly axonal) UL and LL	LD MP axonal neuropathy with upper limb asymmetry	Very limited study; LD sensory and motor neuropathy	LD axonal neuropathy, some acute EMG changes	LD MP axonal neuropathy with conduction slowing	Not performed	Very limited only single nerve studied	*motor axonal neuropathy	LD MP demyelinating neuropathy	Limited study; MPLD neuropathy	*Mixed motor >sensory neuropathy (axonal >demyelinating) involving both LL, R>L	**Motor neuropathy with conduction slowing	LD MP axonal neuropathy with minimal conduction slowing	*Axonal motor neuropathy