Identifying disease-causing short tandem repeat expansions in massively parallel sequencing data, with a focus on ataxias

Rick Michael Tankard
ORCID 0000-0002-8847-9401

Doctor of Philosophy

January 2018

The Walter and Eliza Hall Institute of Medical Research
Department of Medical Biology – The University of Melbourne

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy
Abstract

Short tandem repeat (STR) expansions are responsible for over twenty neurological diseases in humans. This thesis explores the ability to identify disease-causing STR expansions in massively parallel sequencing (MPS) data (also known as next-generation sequencing (NGS) data). The focus of this thesis is on repeat expansions of spinocerebellar ataxias (SCAs) as these can be difficult to subtype clinically. Detection of repeat expansion disorder alleles is important for disease management and carrier screening. Well-developed methods for analysing STRs in MPS data were limited to the fragment size of reads and did not attempt to give calls for oversized repeats. At the start of my PhD there were no published methods for detecting repeat expansions in MPS data. During the course of my PhD I developed one of these methods, exSTRa (expanded STR algorithm).

exSTRa is based on the hypothesis that although MPS reads are too short to cover most repeat expansion alleles, they are still detectable because such alleles lead to an increased number of reads mapping to the STR in question. Firstly, exSTRa identifies as many reads as possible that could map into the highly repetitive STR region, recovering read pairs that are usually discarded, to achieve this. Secondly, exSTRa counts the repeat motif content to form a “repeat score” for each read that approximates the length of the repeat within the read. This is used to form a test statistic for each sample at each locus. We derived approximate distributions of the test statistic under the null distribution.

We tested exSTRa on the largest and currently most diverse cohort of repeat expansion individuals available. We showed that repeat expansions can be detected with polymerase chain reaction–free (PCR-free) whole-genome sequencing (WGS) protocol data (as used by the other two published methods) and with PCR-based WGS sequencing protocol data, as well as whole-exome sequencing (WES) data (for targeted loci).

We also present a family with an heritable, undiagnosed spinocerebellar ataxia with apparent anticipation that we mapped to four genomic locations, including one within the SCA25 locus (OMIM #608703). Anticipation suggested the causal mutation was a repeat expansion, but several heuristic and visual methods failed to find a causal expansion. We found two rare non-repeat variants in the genes STON1 (or read-through gene STON1-GTF2A1L) and PNPT1, but these were of unknown significance and did not explain the apparent anticipation.

This work leads the way to retrospective and prospective repeat expansion detection of known STR loci and, in the future, could be expanded to detecting novel STR loci. This will benefit patients worldwide who may have their genetic disorder pinpointed to a repeat expansion disorder.
Declaration of Authorship

I, Rick Michael Tankard, declare that this thesis titled, ‘Identifying disease-causing short tandem repeat expansions in massively parallel sequencing data, with a focus on ataxias’ and the work presented in it are my own. I confirm that:

■ This thesis comprises only my own original work towards the Doctor of Philosophy, except for where indicated in the preface.

■ Due acknowledgement has been made in the text to all other material used.

■ This thesis is within the maximum word limit of 100,000 in length, exclusive of tables, maps, bibliographies and appendices.

Signed:  

Rick Michael Tankard  12/01/2018
Preface

The work in this thesis was carried out in collaboration with Paul J Lockhart¹,², Katherine B Howell¹,³,⁴, Greta Gillies², Kate Pope², David J Amor³,⁵, Elsdon Storey⁶,⁷, R J McKinlay Gardner⁸, Richard J Leventer¹,³,⁴, Martin B Delatycki¹,²,⁵, and Melanie Bahlo⁹,¹⁰.

More than 90% of this thesis is my own original work, consisting of a literature review, development of methods, bioinformatic analyses, and interpretation of results. Except where acknowledged in the text: Chapters 1 and 5 are entirely my own work; Chapters 2 and 3 were performed in collaboration with PL, MBD and MB, who assisted with patient recruitment, laboratory work, and data generation; and Chapter 4 was performed in collaboration with all listed collaborators, who assisted with clinical phenotyping, patient recruitment, laboratory work, and data generation.

Some methods and results of Chapters 2 and 3 have been published as a preprint (submitted under review at The American Journal of Human Genetics):


During my candidature, work was performed towards the following published peer-reviewed works, but are not presented in this thesis as these were primarily the work of others:

¹Department of Paediatrics, The University of Melbourne, Melbourne, VIC 3010, Australia
²The Bruce Lefroy Centre, Murdoch Childrens Research Institute, Melbourne, VIC 3052, Australia
³Department of Neurology, The Royal Children’s Hospital, Parkville, VIC 3052, Australia
⁴Neuroscience Research Group, Murdoch Childrens Research Institute, Melbourne, VIC 3052, Australia
⁵Victorian Clinical Genetics Services, Melbourne, VIC 3052, Australia
⁶Department of Medicine and Van Cleef Roet Centre for Nervous Disease, Monash University (Alfred Hospital Campus), Melbourne, VIC 3004, Australia
⁷Department of Neuroscience, The Alfred Hospital, Melbourne, VIC 3004, Australia
⁸Clinical Genetics Group, Dunedin School of Medicine, University of Otago, Dunedin 9016, New Zealand
⁹Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC 3052, Australia
¹⁰Department of Medical Biology, The University of Melbourne, Melbourne, VIC 3010, Australia


I was supported by an Australian Postgraduate Award from the Australian Government, Edith Moffat funding, and funding from the Bahlo Laboratory. The Walter and Eliza Hall Institute of Medical Research covered my University of Melbourne Student Services and Amenities Fee each year when applicable.

Additionally for Chapter 4, KBH was supported by the Gustav Nossal National Health and Medical Research Council postgraduate scholarship and The Clifford PhD scholarship, ES was supported by Monash University and Alfred Healthcare, R JL was supported by a Melbourne Children’s Clinician-Scientist Fellowship, PJL was supported by an Australian National Health and Medical Research Council (NHMRC) Career Development Fellowship (1032364), MB was supported by an NHMRC Senior Research Fellowship (1002098) and NHMRC Program Grant (1054618). The Chapter 4 work was also supported by Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIISS funding.
Acknowledgements

First I’d like to thank my primary supervisor Melanie Bahlo for agreeing to supervise me in this project, and suggesting that I was ready to tackle a PhD. Melanie has greatly helped to improve my research skills over the last four years of my project, as well as another three beforehand at The Walter and Eliza Hall Institute of Medical Research (WEHI). Despite running a division and a laboratory of more than ten people, Melanie has always had time to talk about my project or give me feedback almost any day of the week.

I thank my second supervisor, Paul Lockhart for helping to understand the biology side of my project better, and helping to coordinate the collection of samples and sequencing experiments. I thank my third supervisor, Terence Speed, for guidance throughout my 7.5 years at WEHI, and also suggesting I was ready to take on a PhD project.

Martin Delatyki, while not an official supervisor but as an expert on ataxias, helped to better understand the disorders. Gordon Smyth and Matthew Ritchie were on my PhD committee, who contributed ideas and valuable critique during meetings and seminars.

Many more from WEHI contributed to my knowledge and enjoyment at WEHI, including, but not limited to Katherine Smith, Catherine Bromhead, Peter Hickey, Thomas Scerri, Vesna Lukic, Peter Diakumis, Daniel Cameron, Peter Degorski, Saskia Freytag, Karen Oliver, Jan Schröder, Roberto Bonelli, Rhiannon Jenkins, Dineika Chandrananda, and Lyndal Henden. In particular, many thanks go to Keith Satterley for the help in administrating and making use of WEHI systems.

The support at WEHI was wonderful. Both Natalie Senzo and Maria Markovic were friendly and helpful division assistants. I thank the WEHI IT staff for promptly resolving IT issues. I thank Sue Hardy and Keely Bumsted-O’Brien for assistance with administration and thesis structure. I thank the Walter and Eliza Hall Student Association committees, during my PhD, that contributed to many enjoyable events thereby helping me to engage with my fellow student body.

I thank Thomas Scerri for being a great friend in my lab. It was great to have somebody to joke around with and understand my sense of humour.

I thank Nicola Armstrong for enabling me to finish my thesis while starting a post-doc position with her in Perth, Australia.

I thank Carol Dobson-Stone and John Kwok for providing whole-genome sequencing and microsatellite data, that was useful for understanding short tandem repeats but was not presented in this thesis. I thank Katie Ayers, from the Murdoch Childrens Research
Institute, for providing the PCR-free control data presented in this thesis. I thank Egor Dolzhenko of Illumina Inc. for assistance with the use of ExpansionHunter software.

My friends who greatly helped me stay sane during my PhD with pleasant distractions, including Kean Wong, Kelly Chiu, Sumi Oetomo, Vi Vu, Gimwah Sng, Justin Kerney, Nick Kocovski, Christopher & Stephanie Haggarty-Weir, Jess Kapuscinski-Evans, Kristian Wasilewski, Carol & Cameron Barr, Crystal & Wei Liao, Ting Ting Lee, Jonathan Methley, Jeanne-Clare Storace, Elizabeth Howes and Amy Toohey.

Many others contributed to my knowledge and enjoyment during my candidature that I have not listed here, but their contributions are appreciated nevertheless.

I thank Melanie Bahlo, Thomas Scerri, Terry Speed, Paul Lockhart and Nick Kocovski for helping to proof-read my thesis. This has no doubt helped remove many errors in this work. Additionally, I thank my two anonymous reviewers for their interest and careful reading of my thesis; their comments greatly improved the quality of this work.

I’d like to thank my brother Billy, for being a fun and lively housemate throughout my PhD, and often cooking up amazing meals. And also my mother, father and grandmother for their support throughout the years, including giving me and my brother a cat named George who moved in at the beginning of my PhD. I thank my late paternal-grandmother and maternal-grandfather, for their support when I was growing up that enabled me to grow to a level to take on a PhD project. I also thank my aunties and uncles who have helped me throughout the years in small but meaningful ways.

Finally, I’d like to thank my fiancée Nia. It has not been easy spending most of our studies on opposite sides of the country. Submitting this thesis marks the start of wedding preparations that I am sincerely looking forward to. Nia has always encouraged me to do the best that I can do.
Contents

Abstract i
Declaration of Authorship ii
Preface iii
Acknowledgements v
Contents vii
List of Figures x
List of Tables xiii
Third party copyright material xiv
Abbreviations xv

1 Introduction 1
1.1 Human genetics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1
1.2 Human genetic disorders . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
1.3 Massively parallel sequencing . . . . . . . . . . . . . . . . . . . . . . . . . . 4
  1.3.1 Mapping and alignment . . . . . . . . . . . . . . . . . . . . . . . . . 9
  1.3.2 PCR duplicates . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
  1.3.3 Variant calling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
  1.3.4 Effect of MPS on disease discovery . . . . . . . . . . . . . . . . . . . 13
1.4 Short tandem repeats . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14
1.5 Repeat expansion disorders and spinocerebellar ataxia (SCA) . . . . . . . . 17
  1.5.1 Spinocerebellar ataxias . . . . . . . . . . . . . . . . . . . . . . . . . . 22
  1.5.2 Discovery of repeat expansion disorders . . . . . . . . . . . . . . . . 23
1.6 Current methods for analysis of STRs and expansions in MPS data . . . . . . 24
1.7 Thesis outline . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

2 Methods to detect repeat expansions in MPS data 29
  2.1 exSTRa: detecting outlier read length data . . . . . . . . . . . . . . . . . . 29
    2.1.1 exSTRa: Detecting expansions by repeat motif score . . . . . . . . . . 30
    2.1.2 Visualisation of repeat scores . . . . . . . . . . . . . . . . . . . . . 31
    2.1.3 A simple one-sided, Mann-Whitney two-sample test . . . . . . . . . . 34
    2.1.4 exSTRa: an aggregated t-statistic . . . . . . . . . . . . . . . . . . . . 35
2.1.5 exSTRa: p-values for the aggregated t-statistic ......................... 37
2.1.6 Software implementations ................................................. 39
2.2 Data gathering and preparation .............................................. 39
2.2.1 Cohorts of MPS data for STR investigation ......................... 39
2.2.2 MPS data preparation ...................................................... 42
2.3 Repeat expansion detection with existing software .................. 46
2.3.1 lobSTR ................................................................. 46
2.3.2 ExpansionHunter ......................................................... 48
2.4 Comparing repeat expansion detection algorithms .................... 50

3 Results of detecting repeat expansions in MPS data 51
3.1 Quality and alignment summary ............................................. 51
3.2 Expansion detection in other software ..................................... 63
3.2.1 lobSTR results ........................................................ 63
3.2.2 ExpansionHunter results ................................................ 70
3.3 Assessing the repeat score procedure in exSTRa .................... 76
3.4 Classification of repeat expansions ....................................... 86
3.4.1 Visual classification ..................................................... 86
3.4.2 Classification with Mann-Whitney test .............................. 87
3.4.3 Classification with p-values of the aggregated t-statistic ......... 95
3.5 Results summary ............................................................ 103
3.6 Chapter summary ............................................................. 107

4 Massively parallel sequencing and clinical data of a family with spinocerebellar ataxia mapping to the SCA25 locus 108
4.1 Introduction ................................................................. 108
4.2 Methods ................................................................. 109
4.2.1 Recruitment and diagnosis ............................................ 109
4.2.2 SNP chip linkage analysis ............................................ 109
4.2.3 Exome/genome sequencing analysis ............................... 112
4.2.4 Copy-number variations .............................................. 114
4.2.5 Validation by Sanger sequencing .................................... 114
4.2.6 Analysis of repeats for putative repeat expansions ............. 114
4.3 Results ................................................................. 115
4.3.1 Clinical details ......................................................... 115
4.3.2 SNP chip linkage analysis ............................................ 117
4.3.3 Exome/genome sequencing .......................................... 123
4.3.4 Copy-number variations .............................................. 126
4.3.5 Validation by Sanger sequencing .................................... 126
4.3.6 Analysis of repeats for putative repeat expansions ............. 128
4.4 Discussion ............................................................... 132
4.5 Conclusion ............................................................... 134

5 Discussion and summary ..................................................... 135
5.1 Use of MPS to determine the cause of genetic disorders .......... 135
5.2 Developments in this thesis ............................................... 136
5.3 Further directions ........................................................ 137
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Cost per genome</td>
<td>6</td>
</tr>
<tr>
<td>1.2</td>
<td>Illumina HiSeq X Ten system</td>
<td>7</td>
</tr>
<tr>
<td>1.3</td>
<td>Illumina sequencing by synthesis</td>
<td>8</td>
</tr>
<tr>
<td>1.4</td>
<td>FastQ format read pair example</td>
<td>9</td>
</tr>
<tr>
<td>1.5</td>
<td>Aligned MPS data viewed with the Integrative Genomics Viewer (IGV)</td>
<td>10</td>
</tr>
<tr>
<td>1.6</td>
<td>Mapping and alignment</td>
<td>11</td>
</tr>
<tr>
<td>1.7</td>
<td>OMIM entries by year with a phenotype from 1986 to 2016</td>
<td>13</td>
</tr>
<tr>
<td>1.8</td>
<td>(CA)_n repeat D6S282 (located on chr6)</td>
<td>15</td>
</tr>
<tr>
<td>1.9</td>
<td>Percentage of human genome repeats by class</td>
<td>15</td>
</tr>
<tr>
<td>1.10</td>
<td>Human UCSC Simple Repeats track 2-6 bp annotated by ANNOVAR to UCSC Known Genes bar chart</td>
<td>18</td>
</tr>
<tr>
<td>1.11</td>
<td>STRs by genomic region</td>
<td>19</td>
</tr>
<tr>
<td>1.12</td>
<td>Repeat expansion disorder allele sizes compared to HiSeq X data</td>
<td>20</td>
</tr>
<tr>
<td>1.13</td>
<td>Summary of the known genetic causes of spinocerebellar ataxia</td>
<td>22</td>
</tr>
<tr>
<td>2.1</td>
<td>Identifying reads for repeat scoring</td>
<td>32</td>
</tr>
<tr>
<td>2.2</td>
<td>ECDF visual classification</td>
<td>33</td>
</tr>
<tr>
<td>2.3</td>
<td>Sample quantiles across samples</td>
<td>36</td>
</tr>
<tr>
<td>2.4</td>
<td>Sample preparation and data analysis pipeline summary</td>
<td>43</td>
</tr>
<tr>
<td>3.1</td>
<td>FastQC module results</td>
<td>52</td>
</tr>
<tr>
<td>3.2</td>
<td>FastQC per base sequence quality module for WGSrpt_21 read 1</td>
<td>53</td>
</tr>
<tr>
<td>3.3</td>
<td>FastQC per base sequence content module for WGSrpt_21 read 1</td>
<td>53</td>
</tr>
<tr>
<td>3.4</td>
<td>Box plots of GC percent in each FASTQ file</td>
<td>53</td>
</tr>
<tr>
<td>3.5</td>
<td>Box plots of median coverage over capture regions</td>
<td>54</td>
</tr>
<tr>
<td>3.6</td>
<td>WES median and interquartile range (IQR) of coverage, and read count over capture regions</td>
<td>55</td>
</tr>
<tr>
<td>3.7</td>
<td>WGS_PCR_1 median and interquartile range (IQR) of coverage, and read count over WES capture regions for comparison to WES</td>
<td>56</td>
</tr>
<tr>
<td>3.8</td>
<td>WGS_PCR_2 median and interquartile range (IQR) of coverage, and read count over WES capture regions for comparison to WES</td>
<td>56</td>
</tr>
<tr>
<td>3.9</td>
<td>WGS_PF_3 (PCR-free) median and interquartile range (IQR) of coverage, and read count over WES capture regions for comparison to WES</td>
<td>56</td>
</tr>
<tr>
<td>3.10</td>
<td>Read depth of repeat expansion loci with coverage normalised to 30x</td>
<td>57</td>
</tr>
<tr>
<td>3.11</td>
<td>Box plots of duplicate percentages by cohort</td>
<td>58</td>
</tr>
<tr>
<td>3.12</td>
<td>Duplicate percentages by sample</td>
<td>59</td>
</tr>
<tr>
<td>3.13</td>
<td>Median coverage to inferred PCR duplicate rate</td>
<td>60</td>
</tr>
<tr>
<td>3.14</td>
<td>Insert size by cohort</td>
<td>60</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3.15</td>
<td>WES insert size by batch</td>
<td>61</td>
</tr>
<tr>
<td>3.16</td>
<td>Insert sizes by sample</td>
<td>62</td>
</tr>
<tr>
<td>3.17</td>
<td>All allele sizes called by lobSTR</td>
<td>63</td>
</tr>
<tr>
<td>3.18</td>
<td>Allele size difference to minimum expansion size with lobSTR</td>
<td>64</td>
</tr>
<tr>
<td>3.19</td>
<td>WGS allele sizes from lobSTR</td>
<td>66</td>
</tr>
<tr>
<td>3.20</td>
<td>WES_PCR case sample lobSTR allele calls</td>
<td>67</td>
</tr>
<tr>
<td>3.21</td>
<td>WGS_PCR_1 case sample lobSTR allele calls</td>
<td>67</td>
</tr>
<tr>
<td>3.22</td>
<td>WGS_PCR_2 case sample lobSTR allele calls</td>
<td>68</td>
</tr>
<tr>
<td>3.23</td>
<td>WGS_PF_3 PCR-free case sample lobSTR allele calls</td>
<td>69</td>
</tr>
<tr>
<td>3.24</td>
<td>All allele sizes called by ExpansionHunter</td>
<td>70</td>
</tr>
<tr>
<td>3.25</td>
<td>Allele size difference to minimum expansion size with ExpansionHunter</td>
<td>71</td>
</tr>
<tr>
<td>3.26</td>
<td>WGS allele sizes from ExpansionHunter</td>
<td>72</td>
</tr>
<tr>
<td>3.27</td>
<td>WES_PCR case sample ExpansionHunter allele calls</td>
<td>73</td>
</tr>
<tr>
<td>3.28</td>
<td>WGS_PCR_1 case sample ExpansionHunter allele calls</td>
<td>73</td>
</tr>
<tr>
<td>3.29</td>
<td>WGS_PCR_2 case sample ExpansionHunter allele calls</td>
<td>74</td>
</tr>
<tr>
<td>3.30</td>
<td>WGS_PF_3 PCR-free case sample ExpansionHunter allele calls</td>
<td>75</td>
</tr>
<tr>
<td>3.31</td>
<td>repeat score by cohort for select loci part 1</td>
<td>78</td>
</tr>
<tr>
<td>3.32</td>
<td>repeat score by cohort for select loci part 2</td>
<td>79</td>
</tr>
<tr>
<td>3.33</td>
<td>WES_PCR ECDF of select loci</td>
<td>80</td>
</tr>
<tr>
<td>3.34</td>
<td>WGS_PCR_1 ECDF of select loci</td>
<td>81</td>
</tr>
<tr>
<td>3.35</td>
<td>WGS_PCR_2 ECDF of select loci with select samples coloured part 1</td>
<td>82</td>
</tr>
<tr>
<td>3.36</td>
<td>WGS_PCR_2 ECDF of select loci with select samples coloured part 2</td>
<td>83</td>
</tr>
<tr>
<td>3.37</td>
<td>WGS_PCR_2 ECDF of X loci for males</td>
<td>84</td>
</tr>
<tr>
<td>3.38</td>
<td>WGS_PF_3 PCR-free ECDF of select loci</td>
<td>85</td>
</tr>
<tr>
<td>3.39</td>
<td>Mann-Whitney p-values in WES_PCR by locus</td>
<td>88</td>
</tr>
<tr>
<td>3.40</td>
<td>Mann-Whitney p-values in WGS_PCR_1 by locus</td>
<td>89</td>
</tr>
<tr>
<td>3.41</td>
<td>Mann-Whitney p-values in WGS_PCR_2 by locus</td>
<td>90</td>
</tr>
<tr>
<td>3.42</td>
<td>Mann-Whitney p-values in WGS_PF_3 by locus</td>
<td>91</td>
</tr>
<tr>
<td>3.43</td>
<td>Mann-Whitney p-values per cohort</td>
<td>92</td>
</tr>
<tr>
<td>3.44</td>
<td>ROC curves of p-values of Mann-Whitney test per cohort</td>
<td>93</td>
</tr>
<tr>
<td>3.45</td>
<td>Q-Q plots of $-\log_{10}$ transform p-values from Mann-Whitney per cohort</td>
<td>94</td>
</tr>
<tr>
<td>3.46</td>
<td>Aggregated t-statistic in WES_PCR by locus</td>
<td>96</td>
</tr>
<tr>
<td>3.47</td>
<td>Aggregated t-statistic in WGS_PCR_1 by locus</td>
<td>97</td>
</tr>
<tr>
<td>3.48</td>
<td>Aggregated t-statistic in WGS_PCR_2 by locus</td>
<td>98</td>
</tr>
<tr>
<td>3.49</td>
<td>Aggregated t-statistic in WGS_PF_3 by locus</td>
<td>99</td>
</tr>
<tr>
<td>3.50</td>
<td>Aggregate t-statistic p-values per cohort</td>
<td>100</td>
</tr>
<tr>
<td>3.51</td>
<td>ROC curves of p-values of aggregated t-statistic per cohort</td>
<td>101</td>
</tr>
<tr>
<td>3.52</td>
<td>Q-Q plots of $-\log_{10}$ transform p-values from T per cohort</td>
<td>102</td>
</tr>
<tr>
<td>4.1</td>
<td>Pedigree of family affected with unknown spinocerebellar ataxia</td>
<td>111</td>
</tr>
<tr>
<td>4.2</td>
<td>Linkage analysis LOD scores for genome</td>
<td>117</td>
</tr>
<tr>
<td>4.3</td>
<td>Chromosome 2 linkage analysis LOD scores</td>
<td>118</td>
</tr>
<tr>
<td>4.4</td>
<td>Haplotypes of chromosome 1 critical linkage region</td>
<td>119</td>
</tr>
<tr>
<td>4.5</td>
<td>Haplotypes of chromosome 2 critical linkage region</td>
<td>120</td>
</tr>
<tr>
<td>4.6</td>
<td>Haplotypes of chromosome 10 critical linkage region</td>
<td>121</td>
</tr>
<tr>
<td>4.7</td>
<td>Haplotypes of chromosome 15 critical linkage region</td>
<td>122</td>
</tr>
<tr>
<td>4.8</td>
<td>QDNAseq CNVs in UCSC Genome Browser</td>
<td>127</td>
</tr>
</tbody>
</table>
List of Figures

4.9  Sanger sequencing verification of WES and WGS variants . . . . . . . . . . 128
4.10 ExpansionHunter alleles for affected individuals sequenced with Illumina
    HiSeq X, by locus and sample . . . . . . . . . . . . . . . . . . . . . . . . . . 129
4.11 IGV snapshots of STR loci showing WGS reads . . . . . . . . . . . . . . . . . 130

A.1  WES_PCR cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . . 161
A.2  WES_PCR cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . . 162
A.3  WES_PCR cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . . 163
A.4  WES_PCR cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . . 164
A.5  WGS_PCR_1 cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . 165
A.6  WGS_PCR_1 cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . 166
A.7  WGS_PCR_1 cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . 167
A.8  WGS_PCR_1 cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . 168
A.9  WGS_PCR_2 cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . 169
A.10 WGS_PCR_2 cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . 170
A.11 WGS_PCR_2 cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . 171
A.12 WGS_PCR_2 cohort repeat score ECDFs . . . . . . . . . . . . . . . . . . . 172
A.13 WGS_PF_3 cohort (PCR-free) repeat score ECDFs . . . . . . . . . . . . . 173
A.14 WGS_PF_3 cohort (PCR-free) repeat score ECDFs . . . . . . . . . . . . . 174
A.15 WGS_PF_3 cohort (PCR-free) repeat score ECDFs . . . . . . . . . . . . . 175
A.16 WGS_PF_3 cohort (PCR-free) repeat score ECDFs . . . . . . . . . . . . . 176
List of Tables

1.1 Human UCSC Simple Repeats track 2–6 bp annotated by ANNOVAR to UCSC Known Genes .......................... 17
1.2 Human repeat expansion disorders .................................. 21
1.3 MPS STR software .................................................. 25

2.1 Sequencing cohort summary ........................................ 40
2.2 Sample names by diagnosis, individual and cohort .................. 40

3.1 Classification of expansions by visual inspection of ECDFs .............. 86
3.2 Comparing known expansions detected ................................. 104
3.3 Performance summary of algorithmic expansion calls ................ 105
3.4 Summary of algorithmic expansion calls on case samples ................ 106

4.1 Genomic data generated for ataxia family ............................... 112
4.2 Linkage analysis critical regions ...................................... 118
4.3 Number of variants after filtering steps ............................... 124
4.4 Variants after filtering ................................................ 125
4.5 CNV variants found in WGS with QDNAseq, DNAcopy and CGHcal .... 125
4.6 Variants after filtering and Sanger verification .......................... 129
4.7 Coding STRs of 2–12 bp within linkage regions ..................... 131
4.8 Annotated lobSTR STR calls ........................................ 131
4.9 Filtered STRs genotyped by lobSTR ................................... 132

B.1 Estimated sex and call rates from BeadStudio ....................... 177
B.2 Chi-squared values of SNP frequencies of samples ................... 177
B.3 FEstim estimates ..................................................... 178
B.4 WES alignment statistics ............................................. 178
B.5 WGS alignment statistics ............................................. 178
B.6 Markers with matching genotypes in SNP chip data .................. 178
B.7 WES and WGS samples matching at HapMap 2 locations ............... 178
Third party copyright material

Figure 1.1 is available in the public domain courtesy of the National Human Genome Research Institute, as stated in general for their website at https://www.genome.gov/copyright.cfm (accessed 17th August 2017).

Abbreviations

Note that genes and disease symbols are not included in this list, instead, please see Table 1.2 on page 21 for those implicated in repeat expansion disorders. Software tools with abbreviated names as standard are also excluded.

A  Adenine (nucleotide)
AGHA  Australian Genomics Health Alliance
AGRF  Australian Genome Research Facility
AUC  Area Under Curve (of ROC curve)
BAM  Binary version of SAM format
bp  Base Pair
C  Cytosine (nucleotide)
CI  Confidence Interval
cM  CentiMorgan
CNV  Copy-Number Variation
CRAM  Intelligently compressed version of BAM format
dbSNP  The Single Nucleotide Polymorphism Database
DGV  Database of Genomic Variants
DNA  Deoxyribonucleic Acid
ECDF  Empirical Cumulative Distribution Function
eSTR  Expression STR
ExAC  Exome Aggregation Consortium
FastQ  Format for storing sequence data and quality scores
G  Guanine (nucleotide)
GTex  Genotype-Tissue Expression (consortium)
indel  Insertion or Deletion
KCCCG  Kinghorn Centre for Clinical Genomics
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOD</td>
<td>Logarithm of the Odds</td>
</tr>
<tr>
<td>LOESS</td>
<td>Local Regression</td>
</tr>
<tr>
<td>MGHA</td>
<td>Melbourne Genomics Health Alliance</td>
</tr>
<tr>
<td>MPS</td>
<td>Massively Parallel Sequencing (synonym of NGS)</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>NGS</td>
<td>Next-Generation Sequencing (synonym of MPS)</td>
</tr>
<tr>
<td>OMIM</td>
<td>Online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>PF</td>
<td>PCR-free (non-standard acronym)</td>
</tr>
<tr>
<td>polyQ</td>
<td>Polyglutamine (usually with codon CAG in this thesis)</td>
</tr>
<tr>
<td>Q-Q plot</td>
<td>Quantile-Quantile Plot</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>ROC curve</td>
<td>Receiver Operating Characteristic Curve</td>
</tr>
<tr>
<td>SAM</td>
<td>Sequence Alignment/Map format</td>
</tr>
<tr>
<td>SCA</td>
<td>Spinocerebellar Ataxia</td>
</tr>
<tr>
<td>SCAR</td>
<td>Spinocerebellar Ataxia, Autosomal-Recessive forms</td>
</tr>
<tr>
<td>SNP</td>
<td>Single-Nucleotide Polymorphism</td>
</tr>
<tr>
<td>SNV</td>
<td>Single-Nucleotide Variant</td>
</tr>
<tr>
<td>STR</td>
<td>Short Tandem Repeat (also known as microsatellite)</td>
</tr>
<tr>
<td>T</td>
<td>Thymine (nucleotide)</td>
</tr>
<tr>
<td>TRF</td>
<td>Tandem Repeat Finder (software)</td>
</tr>
<tr>
<td>U</td>
<td>Uracil</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated Region</td>
</tr>
<tr>
<td>VCF</td>
<td>Variant Call Format</td>
</tr>
<tr>
<td>WEHI</td>
<td>The Walter and Eliza Hall Institute of Medical Research</td>
</tr>
<tr>
<td>WES</td>
<td>Whole-Exome Sequencing</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole-Genome Sequencing</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

This thesis focuses on the detection of repeat expansions using massively parallel sequencing (MPS) data. We concentrate on repeat expansion disorders that are Mendelian. Repeat expansions are known to cause over 20 human diseases, which are mainly neurological (La Spada and Taylor, 2010). To set the scene for the algorithm to detect repeat expansions, I first will describe some basic human genetics.

In this chapter, we introduce important terms in bold font, along with their definitions.

1.1 Human genetics

Deoxyribonucleic acid (DNA) is a molecule that carries the instructions for the development and functioning of an organism. DNA is a long polymer of nucleotides that stores information in one of four bases: adenine (A), guanine (G), cytosine (C) and thymine (T). Most DNA molecules consist of two strands in a double helix structure, with nucleotides held together between strands by hydrogen bonds according to Watson-Crick rules, where A pairs to T with two hydrogen bonds, and G pairs to C with three hydrogen bonds (Strachan and Read, 2011, pp. 2–7). Lengths of DNA sequences are measured in base pairs (bp) or bases, giving the number of nucleotides on one strand, and known as the physical distance. For large numbers of bases, we often use the terms kilobases (kb) for one thousand bp and megabases (Mb) for one million bp.

In most human cells, there are multiple DNA molecules called chromosomes. In humans these are numbered from 1 to 22 (in approximate order of size) and are known
as **autosomes**. Additionally there are the sex-determining chromosomes, X and Y (**sex chromosomes**), and mitochondrial DNA. In most humans, the majority of cells contain two copies (**diploid**) of each autosome in the cell nucleus, with one copy inherited from each parent. Diploid cells also contain two sex chromosomes: either two copies of the X chromosome making them female or one copy each of the X and Y chromosomes making them male (Strachan and Read, 2011, pp.30–31). The gametes (ova and sperm) only contain a single copy (**haploid**) of the autosomes, plus the X chromosome in ova, and either a single X or Y chromosome in sperm. Additionally, some cells have no cell nucleus and therefore contain no chromosomes; examples include mature red blood cells that destroy the nucleus to carry more oxygen and be small enough to fit through capillaries, and cornified cells such as hair, nails and outer skin that have been filled with the structural protein keratin and undergone apoptosis (programmed cell death) (Eckhart et al., 2013). Some humans will have exceptions to these rules, known as aneuploidy, such as only a single chromosome (monosomy) or an extra chromosome (3 chromosomes, trisomy). Aneuploidy often does not create a viable organism, but otherwise will usually cause a disorder, such as Turner syndrome (single X chromosome, no Y chromosome) and Down syndrome (trisomy 21) (Strachan and Read, 2011, pp.50–51). The collection of a single set of autosomes, sex chromosomes and mitochondrial DNA is a **human genome**, which consists of approximately 3 billion nucleotides (Haines and Pericak-Vance, 1998, p.24). 

A position or region on a chromosome is known as a **locus** (plural **loci**), and may be many base pairs long. **Genes** are loci that are the basic unit of heredity. Many genes are **protein coding**, where the nucleotide sequence is transcribed into ribonucleic acid (**RNA**), more specifically messenger RNA (**mRNA**), where nucleotides are complemented from A to U (uracil), T to A, C to G and G to C. The mRNA undergoes modifications including the systematic splicing of sequences within it, with the removed sequences known as **introns** and the remaining sequences known as **exons**. The resulting mRNA is then translated into a **protein**. A protein consists of one or more long chains of amino acids. **Codons** are sequences of three nucleotides that code one of twenty amino acids, or start and end codons. Translation begins at a start codon (AUG), with this codon and subsequent codons giving a sequence of amino acids until a stop codon (UAA, UAG or UGA) is reached. Sequence in the mRNA before the start codon is known as the 5-prime untranslated region (**5'UTR**), and sequence after the stop codon as the 3-prime untranslated region (**3'UTR**). For a gene, splicing may occur in multiple ways and alternative start codons being used, with
each known as an alternative transcript. Some genes produce non-coding RNA that is not translated into a protein, with the RNA functional itself (Strachan and Read, 2011, pp. 16–26). The set of all exons is known as the exome.

1.2 Human genetic disorders

The genome of different human individuals usually differs, with these differences known as genetic variation. Differences at a single locus are known as variants. For a single locus, each variant is known as an allele. Sometimes during DNA replication, mutations occur that are newly created variants that did not exist in the originating DNA. Many disorders in humans have a genetic component that may be hereditary (passed down from parents DNA (germline)), or somatic where mutations occur during cell development that are not germline and passed onto a restricted set of daughter cells.

Human genetic disorders are on a spectrum from complex disorders, where multiple loci (and possibly environmental factors) contribute to a disorder; to Mendelian, where a single locus is responsible for the disease (Haines and Pericak-Vance, 1998, pp. 53–76).

Mendelian disorders have various modes of inheritance, depending on the location of the mutation and how many copies of a mutation are required to manifest disease. When the causal mutation is on an autosome (chromosomes 1–22), the disease is autosomal, whereas if it is on the X chromosome, the disease is an X-linked disorder (Haines and Pericak-Vance, 1998, pp. 39–43). Autosomal disorders can be classified as dominant, where a single copy of the mutation is sufficient to causes disease; homozygous recessive, where a copy of the same mutation on each chromosome are required to cause disease (these are usually inherited from a common ancestor due to the parents being consanguineous) and compound recessive, where two different mutations in the same gene, one on each chromosome, are required to cause the disorder. X-linked disorders may be X-linked recessive, where all males (being hemizygous (a single copy) for the X chromosome) with the causative allele are affected. In an X-linked recessive disorder females that carry one copy of the causative allele are carriers, and may either pass this onto their sons who become affected or daughters as new carriers (assuming the father did not also have the causative allele); males can only pass the disease allele onto all their daughters. X-linked disorders may also be X-linked dominant, where one copy of the
disease-causing allele is sufficient to cause disease, such that both heterozygous (two different alleles) females and hemizygous males are affected. For the following, the probability of being affected by a given disease $A$ is $P(A)$. The allele causing a given autosomal dominant disease is $D$ (common population allele $d$), and an allele causing a given autosomal recessive disease is $r$ (common population allele $R$). A genotype is written as the combination of two alleles (unordered), such as $Dd$, $dd$, or $Rr$.

**Penetrance** is the proportion of individuals with the underlying disease-causing genotype that exhibit the disease (binary phenotype) (Ott, 1999, p. 4). Formally this is $P(A \mid DD)$ or $P(A \mid Dd)$ for an autosomal dominant disorder and $P(A \mid rr)$ for an autosomal recessive disorder. For many Mendelian disorders the penetrance is less than 1, so that not all individuals who carry the causative alleles, according to the mode of inheritance, will develop the disorder; this may be due to environmental or other genetic factors. Additionally, when focusing on an allele that causes a phenotype, that phenotype may appear without the corresponding risk genotype and is known as a phenocopy; this can result from other genetic or environmental factors. Formally the phenocopy rates are $P(A \mid dd)$ for an autosomal dominant disorder or $P(A \mid RR)$ and $P(A \mid Rr)$ for autosomal recessive disorders. The possibility of phenocopies should be considered more highly in Mendelian disorders when the phenotype is non-specific such as ataxia (loss of full control of movement) or deafness (Kopp et al., 1999).

There are many types of variation that can cause Mendelian disease, including single-nucleotide variants (SNVs), insertions/deletions (indels), structural variation (including copy-number variations (CNVs) and inversions) and expansion of short tandem repeat (STR; also known as microsatellites) sequences. The focus of this thesis is on the expansion of STRs, that we refer to as repeat expansions. Disorders caused by repeat expansions are discussed in further detail in Section 1.5.

### 1.3 Massively parallel sequencing

Sanger sequencing was developed by Frederick Sanger, Nicklen S and Coulson AR in 1977 (Sanger et al., 1977) that allowed determining of the underlying sequence of interest. Sanger sequencing can produce sequences slightly less than 1 kb (1,000 bp) of high quality
Massively parallel sequencing (MPS) (also known as next-generation sequencing (NGS)) technologies have been developed that allow interrogation of genomes at a much higher throughput and cheaper than Sanger sequencing, making sequencing of an entire human genome affordable for small research groups (Park and Kim, 2016). MPS technologies can generate millions of short sequences known as reads. The read base calls have error rates generally higher than that of Sanger sequencing. There are multiple vendors of MPS platforms with varying processes and chemistries. For this thesis we focus on the Illumina platform because of its high-throughput, ease of interpretation, low polymerase stutter and its high usage worldwide. One of the major shortcomings of Illumina data are the short read lengths (limited to 150 bp for the highest-throughput systems and 300 bp for the lower-throughput systems) that cannot span many expanded alleles, while newer competitors such as Oxford Nanopore (reads of 5 kb to hundreds of kb) or PacBio (10 kb to 60 kb) are more able to span large STR alleles (Weirather et al., 2017). However, these newer technologies are currently more expensive and in lower use than Illumina sequencing.

The first human genome was sequenced, from 1990 to 2003, at a cost of approximately three-billion US dollars by the Human Genome Project (Lander et al., 2001). Figure 1.1 shows that the cost of sequencing a human genome has decreased over time, and with the introduction of MPS the cost decreased much more rapidly than the rate of Moore’s law\(^1\) from 2008 (sequencing cost is halving more rapidly than once every two years).

As the cost of sequencing a whole genome could still be very expensive until recently, most early MPS was performed with capture technologies to enrich regions likely to contribute to disease or the feature of interest. This is performed by either a custom design for a project, or ready-made designs to enrich the exome (whole-exome sequencing, WES) (García-García et al., 2016), or to enrich genes in a panel related to a disease of interest (e.g. Illumina TruSight Sequencing Panels). Exome enrichment kits have been produced by Agilent (SureSelect), Roche/Nimblegen (SeqCap) and Illumina (TruSeq). Later iterations of these kits have generally covered more of the exome. Enrichment kits use DNA (Nimblegen, Illumina) or RNA (Agilent) baits complementary to the regions, that bind

\(^1\)We have applied Moore’s law, that the number of transistors in a dense integrated circuit would double every two years, to sequencing, it’s analogue that the number of bases sequenced would double every two years for the same cost. The main text gives an equivalent definition.
to DNA fragments of interest; these fragments are enriched by pull-down with magnetic beads (Clark et al., 2011) for the final library. This enrichment is imperfect such that these libraries will contain many off-target fragments, sometimes more than 50%, that can be useful for some analyses (Smith et al., 2011; Samuels et al., 2013).

![Figure 1.1. Cost per genome](https://www.genome.gov/sequencingcostsdata/updated 24th May 2016, accessed 29th March 2017)

The Illumina HiSeq X Ten platform (Figure 1.2) is notable for being the first system to be able to sequence an entire human genome at 30x coverage for US$1000 (assuming that the system is running at full-capacity, but not including the cost of extraction of DNA and bioinformatic analysis)\(^2\). In many use cases, this often made performing whole-genome sequencing (WGS) preferred over WES. Illumina HiSeq X Ten platforms for WGS either require 100 ng of DNA for use with the TruSeq Nano DNA Library Preparation Kit or 1 µg with the TruSeq DNA PCR-Free Library Preparation Kit.\(^3\) The TruSeq Nano has the advantage of being able to work with a smaller sample of DNA by including a polymerase chain reaction (PCR) step to amplify fragments ligated with adapters, but will suffer multiple PCR-duplicates (see Section 1.3.2) and increase polymerase stutter in the reads. Both of these preparation kits were released in 2013 (Rhodes et al., 2014). **PCR-free** library preparation protocols exclude the PCR step, but often require higher quality and

---


greater quantity DNA samples than PCR methods. Other companies also produce library preparation kits for Illumina sequencing, such as the KAPA Hyper Prep Kit by Kapa Biosystems\(^4\). The KAPA Hyper Prep Kit has multiple possible workflows, including with PCR and PCR-free protocols.

![Illumina HiSeq X Ten system](image)

**Figure 1.2. Illumina HiSeq X Ten system**


Library preparation of DNA fragments for Illumina sequencing involves adding adapters, sequencing binding sites, indices and sequences complementary to the flow cell oligonucleotides (short DNA or RNA fragments). The bases that contain the original DNA fragment is termed the **insert**, and its number of bases as the **insert size**. In **single-end** sequencing, the insert is sequenced once from one end, where as in **paired-end** sequencing the insert is sequenced twice from each end of the insert towards the centre. In paired-end sequencing, the insert size is often large enough that there there is no overlap in the bases sequenced from each end of the insert, while bases in the middle of the insert are not sequenced.

Libraries are loaded to lanes of a flow cell. On the Illumina HiSeq X, a patterned flow cell of nanowells is used to increase throughput. This flow cell has eight lanes each allowing a different library to be sequenced simultaneously. Each flow cell has multiple tiles that contain nanowells, where fragments are captured by surface-bound oligonucleotides complementary to the library adapters, then amplified into clonal clusters with PCR (bridge amplification). In a process named “sequencing by synthesis” (Figure 1.3), each **forward**
read (read 1) is sequenced from the start of the insert simultaneously in cycles that correspond to the position of each base in the read. In each cycle, the next nucleotide in the sequence is attached with a fluorescent tag, which are then imaged four times (once for each nucleotide) each cycle for each tile. In paired-end sequencing, the reverse read (read 2) is obtained after the forward read by repeating the sequencing by synthesis process from the other end of the insert. As the reverse read is sequenced after the first, its quality is lower as there is more chance of errors being introduced. The images are analysed by Illumina software to produce base calls with associated quality scores for each sequence (Illumina Inc, 2017).

MPS data from Illumina sequencing is encoded in FastQ format, which for each read provides: base calls, quality scores for each base, and sequence identifiers that include information on the sequencer, sequencing run, flow cell identifier, lane number and position within the flow cell. Each sequence identifier is designed to be unique across all experiments in the world, but this may not be true of all sequencing providers as the format of the identifier can be modified. Paired-end reads are provided in two FastQ files, with one each for the forward (read 1) and reverse (read 2) reads respectively. An example
of a paired-end read represented in FastQ format is given in Figure 1.4, with sequence ID `@ST-E00132:44:H08PPALXX:5:1101:2441:1010 1:N:0:0` indicating machine identifier ST-E00132, the 44th run of the machine, flow cell ID H08PPALXX, lane number 5, within-lane tile number 1101, (x,y) coordinates of (2441,1010) within the tile, and an indication of whether the read is the first or second in pair. Each base is given a probability estimate $P$ that the base call is incorrect. Base quality scores are Phred-scaled, $Q = -10 \log_{10} P$, and for efficient representation encoded with the ASCII character with number $Q + 33$ (alternatively, some older Illumina data are encoded as $Q + 64$).

<table>
<thead>
<tr>
<th>Forward read (R1)</th>
<th>FastQ format text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence ID</td>
<td>@ST-E00132:44:H08PPALXX:5:1101:2441:1010 1:N:0:0</td>
</tr>
<tr>
<td>Base calls (Uninformative)</td>
<td>GCATCAAGGTCTTCTTTTGAGAAGTATCTGTTCATATCCTTCACTCTCTTGTTGATGGGGTTGTTTGT</td>
</tr>
<tr>
<td>Quality scores</td>
<td>AFFFFJFAJFJJJJFJJJJ7FJFJJFJJJJ-FAJJJAJFJJJJFJJJJAFJFFJJJJFFJFJ&lt;AF</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reverse read (R2)</th>
<th>FastQ format text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence ID</td>
<td>@ST-E00132:44:H08PPALXX:5:1101:2441:1010 2:N:0:0</td>
</tr>
<tr>
<td>Base calls (Uninformative)</td>
<td>AAGACTTAAATGTTAGACCTAAAACCATAAAAACCCTAGAAGAAAAGCTAGGCAATACCTTTCAGGAC</td>
</tr>
<tr>
<td>Quality scores</td>
<td>&lt;77FFJJJJJJJJJJJJJ-A-FJJFJJJJJJcJJJJJJJJAA-FAJJ-JFJFFJJJJJAFAJF&lt;-&gt;FJJ</td>
</tr>
</tbody>
</table>

**Figure 1.4. FastQ format read pair example**

**GC content bias** (or GC bias) is a problem of Illumina sequencing where regions of low or high proportion of G or C bases are less likely to be observed in the final data set than moderate GC content, as sequencing temperatures are optimised for the average GC content observed in the genome (with higher GC regions requiring higher temperatures to break bonds). In humans, the highest coverages are observed in regions of 40 to 55% GC content (Benjamini and Speed, 2012). When the coverage of reads is important for the analysis, such as in copy-number variation detection, then correcting for GC content bias is an important step for accurate results. GC bias may additionally make the detection of variation in some genomic regions difficult due to poor coverage.

### 1.3.1 Mapping and alignment

A reference genome provides an approximation to the DNA sequence of a member of that species. An **alignment** describes the location of each base of a read with respect to the reference genome. In MPS, the correct alignment is not known *a priori*. Instead, various algorithms infer alignment of reads to a reference genome, returning the position of each base relative to the reference. A example of aligned reads is shown for one locus...
with the Integrative Genomics Viewer (IGV) (Thorvalsdóttir et al., 2013) in Figure 1.5. Alignments are usually stored in either a human-readable Sequence Alignment/Map format (SAM); **BAM format**, a binary version of SAM that allows compression and indexing for fast retrieval of alignments; or **CRAM format**, that allows lossless compression of BAM by a further 40–50% by taking advantage of the fact that most bases will match the reference sequence, while still allowing indexing. Further, as lossless CRAM files may still be at least 100 GB for each human sample, the format allows lossy compression for more compact file sizes (e.g. lossy compression of base quality scores). Other formats for aligned sequence data exist, but are not in general use.

Unless stated otherwise, all genome coordinates in this thesis will be with respect to the human reference genome hg19, released in February 2009 (equivalent to GRCh37).

![Figure 1.5. Aligned MPS data viewed with the Integrative Genomics Viewer (IGV)](image)

This figure shows the alignment of about 50 reads across the complement factor H related 3 (CFHR2) gene. Each individual read is represented by a grey horizontal bar with the underlying sequence (A, C, G or T) colour coded. The reference sequence for this locus is shown at the bottom as continuous lines of sequence, together with corresponding amino acid codes on the forward strand for codons at the three frame start locations. Also shown is a SNV that is visible in the centre of the figure (bounded by dashed vertical lines). It appears as a SNV because approximately half of the reads have a T and the other half a C at this position.

The difference between **read mapping** and **read alignment** is subtle and therefore are often used interchangeably, but in this thesis the difference is important. The **alignment** of a read to a reference describes the location of each base of the read with respect to the reference, whereas the **mapping** of the read is the locus from the first to last reference bases of the alignment (Figure 1.6). An alignment is considered correct if all the bases
match correctly, whereas mapping is considered correct if it overlaps the true mapping (Li, 2011b). In this way, the information contained in a read mapping is a subset of the information conveyed by the alignment. Therefore a read with the correct alignment always has a correct mapping, but a read with correct mapping does not necessarily have the correct alignment.

![Figure 1.6. Mapping and alignment](image)

Example of two read sequence’s alignment and mapping. The mapping with respect to the reference is shown as grey bars, that is, the first to last reference bases. The alignment specifies which bases match, where bases have either been deleted or inserted, or soft-clipped (bases not aligned).

A full alignment can be performed with the dynamic programming algorithms Needleman-Wunsch (Needleman and Wunsch, 1970) or Smith-Waterman (Smith and Waterman, 1981), that finds the optimal alignment between two sequences. However, as the human genome consists of approximately 3 billion bp, MPS data for a human patient with 30x coverage can have tens of million reads in WES and hundreds of millions of reads in WGS. Dynamic programming algorithms would not be able to perform a full alignment for each MPS read within a reasonable time frame. Instead, MPS alignment software uses heuristics to speed up alignment, sacrificing the guarantee that the given alignment is the best global alignment. Most software aligns each read independently, where multiple potential mapping locations are found (by matching part of the read sequence to the reference genome). The reads then undergo local alignment with the Needleman-Wunsch or Smith-Waterman algorithms at each potential mapping, with the final alignment the one with the best alignment score that these algorithms provide. Due to the repetitive nature of much of the human genome, many reads have multiple locations with the same, or best possible (exact match) alignment score, that are marked as multi-mapping reads. Some reads cannot be mapped or have insufficient alignment scores due to variation, difficult to sequence areas of the human genome, and sequencing errors. These reads are not assigned to any locus and known as unmapped reads.
The quality of a mapping is commonly given as a mapping quality score, that is also Phred-scaled, defined as
\[ MAPQ = -10 \log_{10} p, \]
rounded to the nearest integer, where \( p \) is the probability that the mapping is incorrect. Note that the calculation of the mapping quality score varies between software implementations, and may not accurately reflect the probability of an incorrect mapping (Ruffalo et al., 2011).

The insert size is approximated by the number of base pairs from the start of the forward read to the start of the reverse read. This is approximate as it unknown whether the original DNA fragment contained any indels. It is possible for the forward and reverse reads to overlap when the insert size is smaller than the total base pairs of the forward and reverse reads, and care must be taken that these sequences are not counted twice in variant detection.

### 1.3.2 PCR duplicates

Generating a library for Illumina platforms may involve a polymerase chain reaction (PCR) step to amplify DNA fragments with adapters. This is required to make sure these ligated fragments make up the majority of the library. The downside to this is that the same DNA fragments may be sequenced multiple times, with each subsequent sequence adding little information. This is particularly problematic for calling heterozygous SNVs or single-nucleotide polymorphisms (SNPs), where a confident call will have approximately half of the reads from each allele. PCR duplicates lead to alleles disproportionately represented and thus an overall increase in the variability of the allele proportions, which can cause downstream callers to erroneously classify the allele as homozygotes if not properly accounted for. For this reason, we mark duplicates with the software Picard (Broad Institute of MIT and Harvard, 2016) which finds read pairs with matching start locations for the two reads (the full alignment or sequence content is not taken into account). From these potential duplicates, Picard marks all but the highest-quality sequence as duplicates. Well written software will either ignore the duplicate marked reads or interpret them in a meaningful way such as merging the read information. Picard also marks optical duplicates, whereby a single cluster is mistakenly interpreted as multiple clusters by Illumina software.

Alternatively, using a PCR-free protocol can avoid PCR-duplicates, but will require a greater quantity of high-quality DNA.
1.3.3 Variant calling

Various software attempt to determine variants, such as SNVs and indels, in samples compared to the reference genome. These variants are commonly stored in Variant Call Format (VCF), that can include information for multiple samples. VCF may also encode information without sample genotypes, such as for databases giving allele frequencies (e.g. The Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al., 2001)) or summarising the allele frequencies in a large cohort (e.g. Exome Aggregation Consortium (ExAC) (Lek et al., 2016)). There are other alternatives to VCF, but these have not gained widespread use across software tools. Most software will call variants from aligned reads in BAM/CRAM format, but some may interrogate unaligned FastQ data instead, or from another format that stores MPS data.

1.3.4 Effect of MPS on disease discovery

Online Mendelian Inheritance in Man (OMIM) is a knowledge base of human genes and genetic disorders (Hamosh et al., 2005). Each entry in OMIM, either for a phenotype, gene, or both, summarises publications related to that topic. Due to the high throughput of MPS compared to Sanger sequencing, MPS has had a significant impact on disease discovery, with the number of new entries in OMIM with a phenotype increasing greatly from 2011 as shown in Figure 1.7.

![Figure 1.7. OMIM entries by year with a phenotype from 1986 to 2016](image_url)
MPS is increasingly being used as a tool in clinical genomics, including genetic diagnoses of inherited diseases. Sequencing can tell us about an individual’s risk of developing a disease or their likely response to a specific treatment. DNA from cancer cells can be sequenced to profile the tumour and guide treatment, known as precision medicine. Most clinical sequencing is currently performed on the Illumina HiSeq X Ten (or Five) system with 30x coverage that can sequence over 18,000 human genomes per year. Unfortunately, methods to detect disease-causing repeat expansions on the HiSeq X are limited and could be improved upon. We discuss these further in subsequent sections.

1.4 Short tandem repeats

Repetitive loci consist of sequences that occur multiple times in the genome (allowing for small differences). The majority of repeat sequence is transposable element–derived, that is, a sequence from one locus that has been duplicated at another locus (noting that the ancestral locus may have since been deleted). Estimating the percentage of repetitive sequence of a genome relies on computational methods to find similar sequences, allowing for mismatches and indels that have occurred over time. The RepeatMasker (Smit et al., 2013) software estimates that approximately 50% of the human sequence is repetitive, while estimates using P-clouds (oligonucleotide excess probability clouds (Gu et al., 2008)) have put this as high as 66–69% (de Koning et al., 2011). Approximately 8% of the genome consists of tandemly repeated sequence, where the same nucleotide sequence is repeated multiple times next to each other. The repeated sequence is known as the repeat unit or motif. Tandem repeats can be divided into classes by size. Satellites are the largest class that have repeat units hundreds of base pairs long, followed by minisatellites with repeat units 10–50 bp long and short tandem repeats (STRs) (microsatellites) 1–6 bp long, although the exact definitions vary in the literature. An example of an STR is given in Figure 1.8 together with nomenclature to describe the repeats. A breakdown of some of the repeat classes as annotated by RepeatMasker is given in Figure 1.9.

Algorithms to find STRs in DNA sequences vary in their tolerance for mismatches and length, with shorter repeats more likely to be false-positives without any biological significance (Schaper et al., 2012). Many STRs are polymorphic, that is, alleles of various sizes are common in the population. Therefore an STR that is polymorphic is sufficient for biological significance (Willems et al., 2014), but being non-polymorphic does
Figure 1.8. (CA)$_n$ repeat D6S282 (located on chr6).

A dinucleotide repeat, with repeat unit ‘CA’. Sequence (a) represents the tandem repeat in the hg19 human genome reference sequence with the repeat itself in red. Sequence (b) is the same sequence but with the repeat written concisely with a subscript indicating the repeat number.

not exclude biological significance. Schaper et al. (2012) suggest that multiple tandem-repeat prediction algorithms are used to find the most complete set of genuine repeats with subsequent filtering of the results.

Figure 1.9. Percentage of human genome repeats by class

As annotated by RepeatMasker. SINE are short interspersed nuclear elements. LINE are long interspersed nuclear elements. LTR are long terminal repeats. Simple repeats are STRs. Low complexity repeats are sequences that are rich for one or two bases. RNA includes all RNA-associated repeats not associated with SINEs, that is, RNA, rRNA, scRNA, snRNA, srpRNA and tRNA. Some repeats cannot be easily classified in the other categories and are left unknown. (Bao et al., 2015)

STRs will sometimes undergo polymerase stutter during DNA replication, where the number of repeats changes due to slippage. Backwards slippage in DNA replication causes an insertion of repeats while forward slippage causes a deletion of repeats (Strachan and Read, 2011). There are approximately one in five-hundred STR slippage mutations per locus per generation, which is 200 times higher than the rate of mutation for spontaneous CNVs and 200,000 times higher than de novo SNVs (Gymrek et al., 2012).

Unlike SNV, indel and CNV mutations, STR slippage mutations are readily reversible, where the repeat number may revert to a previous state.

Tandem repeats where the sequence is repeated perfectly are termed pure, where as if the tandem repeat includes base substitutions or indels within the repeat locus we term them impure. There is no strict definition of what constitutes an STR, as short motifs
may occur a small number of times in tandem by chance rather than polymerase stutter. Repeats with shorter repeat units, or a large number of repeat units, are more likely to undergo polymerase stutter, making it more likely these loci are polymorphic. There are multiple algorithms that can detect tandem repeats within a reference sequence. We may use these algorithms to give a definition of what constitutes an STR (Schaper et al., 2012). One such algorithm is the Tandem Repeat Finder (TRF) (Benson, 1999), which finds tandem repeats of all repeat unit sizes. The TRF software was used to create the “Simple repeats” track in the UCSC Genome Browser (Kent et al., 2002). In the Simple Repeats track, the minimum STR length is 24 bp, with shorter repeats filtered by TRF as likely false-positive hits. The settings of TRF may be adjusted so that the minimum STR length is shorter.

When restricting the STR definition to loci with repeat units of 2–6 nucleotides there are approximately a quarter of a million loci (Gymrek et al., 2012). In Table 1.1 and Figure 1.10 we annotated the region of each of the Simple Repeats track STRs with ANNOVAR (Wang et al., 2010) to the UCSC Known Genes set (Hsu et al., 2006). ANNOVAR is a tool for annotating SNVs and indels to gene location (such as coding or intronic), protein changes, and population frequency. However, ANNOVAR does not annotate STRs; we therefore retrieved gene location by converting STR loci to deletions that cover the same reference bases. We see that most coding STRs are tri- or hexanucleotide repeats, where polymerase stutter does not cause frameshifts and is therefore more likely to be tolerated in the protein function. We also observe that the majority of STRs are intergenic, but this is driven by the fact that most regions of the genome are intergenic. We show the relative ratio of STRs in each genomic region in Figure 1.11. STRs are most underrepresented in coding regions, suggesting negative selection pressure for STRs in coding sequences. STRs are also underrepresented in UTRs and ncRNA exons.

STRs are overrepresented upstream of genes, which contain the gene promoters, as previously shown by Sawaya et al. (2013). STRs can change the distance between promoter elements (Yoge et al., 1991), affecting gene expression and therefore can contribute to variability in phenotypes. Genes with STRs in their promoters, 3'UTR, introns and exons have greater divergence of expression within the great apes than genes without STRs in these regions (Bilgin Sonay et al., 2015), suggesting that STRs facilitate rapid changes in evolution, and that the overrepresentation of STRs within promoter regions allow the expression of these genes to adapt faster. It has also been shown that STRs that affect
Table 1.1. Human UCSC Simple Repeats track 2–6 bp annotated by ANNOVAR to UCSC Known Genes

Splicing is within 2 bp of a splice site, but not marked as coding (exonic in ANNOVAR). Upstream is within 1 kb upstream of a 5’UTR (UTR5). Downstream is within 1 kb downstream of a 3’UTR (UTR3). If multiple region types apply for an STR, then the first member of this list takes precedence (ties will have both): coding, splicing, ncRNA_exonic, ncRNA_splicing, ncRNA_intronic, UTR5/UTR3, intronic, upstream/downstream, intergenic.

<table>
<thead>
<tr>
<th>ANNOVAR region</th>
<th>Total</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding (ORF)</td>
<td>601</td>
<td>9</td>
<td>416</td>
<td>17</td>
<td>12</td>
<td>147</td>
</tr>
<tr>
<td>UTR5</td>
<td>510</td>
<td>109</td>
<td>231</td>
<td>41</td>
<td>57</td>
<td>72</td>
</tr>
<tr>
<td>UTR3</td>
<td>1,226</td>
<td>698</td>
<td>85</td>
<td>212</td>
<td>127</td>
<td>104</td>
</tr>
<tr>
<td>UTR5;UTR3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Splicing</td>
<td>16</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Intrinsic</td>
<td>86,175</td>
<td>37,593</td>
<td>6,080</td>
<td>25,003</td>
<td>10,946</td>
<td>6,553</td>
</tr>
<tr>
<td>ncRNA_exonic</td>
<td>457</td>
<td>199</td>
<td>63</td>
<td>92</td>
<td>57</td>
<td>46</td>
</tr>
<tr>
<td>ncRNA_splicing</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ncRNA_intronic</td>
<td>10,080</td>
<td>4,610</td>
<td>635</td>
<td>2,972</td>
<td>1,085</td>
<td>778</td>
</tr>
<tr>
<td>Upstream</td>
<td>1,869</td>
<td>700</td>
<td>251</td>
<td>454</td>
<td>245</td>
<td>219</td>
</tr>
<tr>
<td>Downstream</td>
<td>1,464</td>
<td>640</td>
<td>92</td>
<td>441</td>
<td>172</td>
<td>119</td>
</tr>
<tr>
<td>Upstream;Downstream</td>
<td>61</td>
<td>23</td>
<td>7</td>
<td>13</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>Intergenic</td>
<td>136,520</td>
<td>58,347</td>
<td>8,901</td>
<td>43,450</td>
<td>15,915</td>
<td>9,907</td>
</tr>
<tr>
<td>Total</td>
<td>238,984</td>
<td>102,936</td>
<td>16,765</td>
<td>72,700</td>
<td>28,632</td>
<td>17,951</td>
</tr>
</tbody>
</table>

expression (expression STRs (eSTRs)) are most strongly enriched in weakly regulated promoters compared to other regions (Gymrek et al., 2016).

1.5 Repeat expansion disorders and spinocerebellar ataxia (SCA)

The number of repeats in an STR can affect molecular phenotypes by varying transcription rates, transcription stability, rates of protein folding and turnover, and protein-protein interactions, aggregation and subcellular location. Repeat number mutations are able to give incremental quantitative effects, being described metaphorically as ‘biological tuning-knobs’ that are postulated to be important in evolution (Fondon et al., 2008). STRs that are non-trinucleotide within coding regions are able to reversibly disrupt gene function due to frameshift changes leading to premature stop codons or a drastically different amino acid sequence. One example can be found in pigs where in the melanocortin receptor 1 (MC1R) gene, a frameshift dinucleotide somatic repeat mutation causes a premature stop codon, leading to the spotting of coat colour (Kijas et al., 2001).
There are at least twenty-two disorders caused by expansions of STRs. At present all known human repeat expansions disorders manifest with neurological symptoms, with only associations for other disorders (Usdin, 2008; Kozlowski et al., 2010). A summary of known repeat disorders is given in Figure 1.12 and Table 1.2. These repeat expansion disorders are typically characterised by genetic anticipation, where individuals in later generations have an earlier age of onset than previous generations. It should be noted that anticipation only describes the change in age of onset and not the underlying cause. Before the discovery that repeat expansions caused disease, it was debated whether anticipation was an actual phenomenon or ascertainment bias, resulting from increased family awareness, leading to earlier diagnosis (Friedman, 2011). Diseases with anticipation will usually have an increase in the severity of symptoms in younger generations. Repeat expansion disorders show anticipation due to the STR increasing in size through polymerase stutter. An example of anticipation in a Huntington disease (HD) (also known as Huntington’s disease) family is seen in Figure 1 of Ranen et al. (1995). It is suspected that STRs affect other non-pathogenic neurological traits, but evidence is limited (Fondon et al., 2008). It is highly likely that further repeat expansion disorders await discovery.
Chapter 1. Introduction

Number of STRs in each class of genomic region over the expected frequency if STRs were uniformly distributed across the genome. Red indicates regions with fewer than expected number of STRs and blue an increased number of STRs. Region classes with less than 100 members have been excluded due to high variability. See Table 1.1 on page 17 for region definitions.

The mode of inheritance varies between repeat expansion disorders, with most showing autosomal dominant inheritance, some X-linked and others autosomal recessive such as in Friedreich ataxia. As Friedreich ataxia is an autosomal recessive disorder, anticipation is not usually observed as affected parent-offspring pairs are uncommon, but longer repeats lead to an earlier age of onset (Monros et al., 1997). Most repeat expansion disorders are trinucleotide repeats. Polymerase stutter does not cause frameshift mutations in coding regions, therefore under less selection pressure, unlike other size repeat units that are not multiples of 3 bp. Additionally, the smaller repeat unit size of 3 bp increases the chance of expansions. Most expansions within coding regions affect in-frame CAG repeats, coding for the amino acid glutamine (Q) and are known as polyglutamine (polyQ) tracts. Each disorder has a stable repeat size range that rarely undergoes expansion, and unstable, larger repeat sizes that are pathogenic. There is a gap between the stable and unstable repeat sizes (an example is seen in Figure 1 of Ranum et al. (1994)), as the few individuals with intermediate repeat number are typically healthy and not assessed clinically, but these alleles will continue to undergo further expansion becoming pathogenic in successive generations. Repeat expansion disorders where the STR is inside a coding region generally become pathogenic at smaller expansion sizes, compared to other repeat expansion disorders. This reflects the lower tolerance of protein coding sequences to variation that influences the
structure or function of the encoded protein.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Disease symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTT</td>
<td>HD (3)</td>
</tr>
<tr>
<td>AR</td>
<td>SBMA (3)</td>
</tr>
<tr>
<td>ATXN1</td>
<td>SCA1 (3)</td>
</tr>
<tr>
<td>ATXN2</td>
<td>SCA2 (3)</td>
</tr>
<tr>
<td>ATXN3</td>
<td>SCA3 (3)</td>
</tr>
<tr>
<td>CACNA1A</td>
<td>SCA6 (3)</td>
</tr>
<tr>
<td>ATXN7</td>
<td>SCA7 (3)</td>
</tr>
<tr>
<td>TBP</td>
<td>SCA17 (3)</td>
</tr>
<tr>
<td>ATN1</td>
<td>DPRPLA (3)</td>
</tr>
<tr>
<td>JPH3</td>
<td>HDL2 (3)</td>
</tr>
<tr>
<td>FMR1</td>
<td>FRAXA (3)</td>
</tr>
<tr>
<td>FMR2</td>
<td>FRAXE (3)</td>
</tr>
<tr>
<td>DMPK</td>
<td>DM1 (3)</td>
</tr>
<tr>
<td>FTXN</td>
<td>FRDA (3)</td>
</tr>
<tr>
<td>ZNF9</td>
<td>DM2 (4)</td>
</tr>
<tr>
<td>C9orf72</td>
<td>FTDALS (6)</td>
</tr>
<tr>
<td>NOP56</td>
<td>SCA36 (6)</td>
</tr>
<tr>
<td>ATXN10</td>
<td>SCA10 (5)</td>
</tr>
<tr>
<td>CSTB</td>
<td>EPM1 (12)</td>
</tr>
<tr>
<td>PPP2R2B</td>
<td>SCA12 (3)</td>
</tr>
<tr>
<td>ATXN8/ATXN8OS</td>
<td>SCA8 (3)</td>
</tr>
</tbody>
</table>

See Table 1.2 for full disease names. Number in parentheses indicates the length of the repeat motif. In the whole-genome sequencing cohort 1 we discuss in Chapter 2, the median insert size was 397 bp and we have indicated this as a vertical bar.
### Table 1.2. Human repeat expansion disorders

<table>
<thead>
<tr>
<th>Disease</th>
<th>Symbol</th>
<th>OMIM</th>
<th>MOI</th>
<th>Gene</th>
<th>Gene location</th>
<th>Repeat motif</th>
<th>Stable repeat number</th>
<th>Unstable repeat number</th>
<th>Strand</th>
<th>Start hg19</th>
<th>Reference repeat number</th>
<th>TRF match</th>
<th>TRF indel</th>
<th>Reference STR size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huntington disease</td>
<td>HD</td>
<td>143100</td>
<td>AD</td>
<td>HTT</td>
<td>4p16.3 coding</td>
<td>CAG</td>
<td>6–34</td>
<td>36–100+</td>
<td>+</td>
<td>3,076,604</td>
<td>21.3</td>
<td>96%</td>
<td>0%</td>
<td>64 bp</td>
</tr>
<tr>
<td>X-linked spinal and bulbary muscular atrophy / Kennedy disease</td>
<td>SBMA</td>
<td>313200</td>
<td>X</td>
<td>AR</td>
<td>Xq12 coding</td>
<td>CAG</td>
<td>0–35</td>
<td>38–62</td>
<td>+</td>
<td>66,765,159</td>
<td>33.3</td>
<td>86%</td>
<td>9%</td>
<td>103 bp</td>
</tr>
<tr>
<td>Spinocerebellar ataxia 1</td>
<td>SCA1</td>
<td>183090</td>
<td>AD</td>
<td>ATXN1</td>
<td>12p14 coding</td>
<td>CAG</td>
<td>15–24</td>
<td>32–200</td>
<td>−</td>
<td>112,036,754</td>
<td>23.3</td>
<td>97%</td>
<td>0%</td>
<td>70 bp</td>
</tr>
<tr>
<td>Machado-Joseph disease</td>
<td>SCA3</td>
<td>109150</td>
<td>AD</td>
<td>ATXN3</td>
<td>14q32.1 coding</td>
<td>CAG</td>
<td>13–36</td>
<td>61–84</td>
<td>−</td>
<td>92,537,355</td>
<td>14</td>
<td>84%</td>
<td>0%</td>
<td>42 bp</td>
</tr>
<tr>
<td>Spinocerebellar ataxia 6</td>
<td>SCA6</td>
<td>183086</td>
<td>AD</td>
<td>ATXN2</td>
<td>12q24 coding</td>
<td>CAG</td>
<td>4–7</td>
<td>21–33</td>
<td>−</td>
<td>32,440,014</td>
<td>13.3</td>
<td>100%</td>
<td>0%</td>
<td>68 bp</td>
</tr>
<tr>
<td>Spinocerebellar ataxia 7</td>
<td>SCA7</td>
<td>184500</td>
<td>AD</td>
<td>ATXN7</td>
<td>3p14.1 coding</td>
<td>CAG</td>
<td>4–35</td>
<td>37–106</td>
<td>+</td>
<td>63,898,361</td>
<td>10.7</td>
<td>100%</td>
<td>0%</td>
<td>83 bp</td>
</tr>
<tr>
<td>Spinocerebellar ataxia 10</td>
<td>SCA10</td>
<td>603516</td>
<td>AD</td>
<td>ATXN10</td>
<td>22q13.31 intron 9</td>
<td>ATTCT</td>
<td>10–20</td>
<td>50–4500</td>
<td>+</td>
<td>46,191,235</td>
<td>14</td>
<td>100%</td>
<td>0%</td>
<td>70 bp</td>
</tr>
<tr>
<td>Dentine-pallidoluysian atrophy</td>
<td>DRPLA</td>
<td>125370</td>
<td>AD</td>
<td>ATNR1</td>
<td>12p13.31 coding</td>
<td>CAG</td>
<td>7–34</td>
<td>49–88</td>
<td>+</td>
<td>7,045,880</td>
<td>19.7</td>
<td>92%</td>
<td>0%</td>
<td>59 bp</td>
</tr>
<tr>
<td>Huntington disease-like 2</td>
<td>HDL2</td>
<td>606438</td>
<td>AD</td>
<td>JPH3</td>
<td>3p14.1 coding</td>
<td>CAG</td>
<td>1–20</td>
<td>32–200</td>
<td>−</td>
<td>32,440,014</td>
<td>13.3</td>
<td>100%</td>
<td>0%</td>
<td>68 bp</td>
</tr>
<tr>
<td>Fragile-X site A</td>
<td>FRAXA</td>
<td>109150</td>
<td>AD</td>
<td>ATXN3</td>
<td>14q32.1 coding</td>
<td>CAG</td>
<td>13–36</td>
<td>61–84</td>
<td>−</td>
<td>92,537,355</td>
<td>14</td>
<td>84%</td>
<td>0%</td>
<td>42 bp</td>
</tr>
<tr>
<td>Myoclonus epilepsy of Unverricht and Lundborg</td>
<td>EPM1</td>
<td>254800</td>
<td>AR</td>
<td>CSTB</td>
<td>19p13.33 promoter</td>
<td>CCGGCCCGCCC</td>
<td>2–3</td>
<td>50–400</td>
<td>−</td>
<td>45,092,324</td>
<td>3.1</td>
<td>100%</td>
<td>0%</td>
<td>55 bp</td>
</tr>
<tr>
<td>Spinocerebellar ataxia 12</td>
<td>SCA12</td>
<td>604328</td>
<td>AD</td>
<td>ATXN12</td>
<td>12q22</td>
<td>AtRNA</td>
<td>7–45</td>
<td>55–75</td>
<td>−</td>
<td>146,258,291</td>
<td>10.7</td>
<td>100%</td>
<td>0%</td>
<td>70 bp</td>
</tr>
<tr>
<td>Spinocerebellar ataxia 17</td>
<td>SCA17</td>
<td>117210</td>
<td>AD</td>
<td>BEAN</td>
<td>16q21 insertion</td>
<td>TGGAA**</td>
<td>16–34</td>
<td>74–+</td>
<td>+</td>
<td>70,713,519</td>
<td>15.3</td>
<td>100%</td>
<td>0%</td>
<td>46 bp</td>
</tr>
</tbody>
</table>

Grouped by location of repeat within gene. OMIM=Online Mendelian Inheritance in Man accession number. MOI=Mode of inheritance (AD=Autosomal dominant, AR=autosomal recessive, X=X-linked.) Unstable repeat numbers are as usually observed in affected individuals, but may be longer. TRF=Tandem Repeats Finder, 100% TRF match is a pure repeat in the reference. *The repeat expansion in HDL2 is either coding, 3'UTR or intronic depending on the transcript. **SCA31 is an insertion of an STR and therefore not located on the reference or in TRF. (Strachan and Read, 2011; Fondon et al., 2008; Taroni et al., 2013; Renton et al., 2011; Kobayashi et al., 2011).
1.5.1 Spinocerebellar ataxias

Spinocerebellar ataxia (SCA) is a family of rare genetic autosomal-dominant disorders. In Europe it is estimated to have a prevalence of 0.9 to 4 in 100,000 (Leone et al., 1995; Silva et al., 1997; van de Warrenburg et al., 2002). It typically involves a slow progressive incoordination of gait and often poor coordination of hands, speech and eye movements. Autosomal-recessive forms are referred to as spinocerebellar ataxia autosomal-recessive (SCAR). The genetic causes of the known types of spinocerebellar ataxia are given in Figure 1.13. The majority (>60%) of SCA cases are caused by polyglutamine repeat expansions, while approximately 30% of cases currently have an unknown cause (including those with and without known loci) (Bird, 2016).

![Figure 1.13. Summary of the known genetic causes of spinocerebellar ataxia](image)

STRs, the main focus of this thesis, are coloured in maroon. Includes disorders up to SCA43 and SCAR24.

Anticipation, with varying levels of evidence, has been observed in the 12 spinocerebellar ataxias caused by a repeat expansion or insertion. Most SCAs that are caused by polyglutamine (CAG) expansions expand paternally (SCA1, 2, 3, 6, 7, DRPLA), except SCA17 that shows expansions from both parents, but still more often paternally. SCA8, caused by a CTG expansion in the utRNA gene ataxin 8 opposite strand ATXN8OS (and its complement CAG in the ataxin 8 utRNA gene ATXN8), is the only known spinocerebellar ataxia where expansions primarily occur maternally, but similar behaviour is observed for other repeat expansion disorders including DM1, FRAXA, FRAXE and HDL2. In general, paternal expansions are more likely to occur for smaller repeat sizes with expansions of...
a few repeat units, while maternal expansions are more likely to occur for larger repeats, adding 10 to 100s of repeat units (McMurray, 2010). Although SCA31 is caused by an insertion containing pentanucleotide repeats of TGGAA rather than an expansion of an existing STR, the age of onset of disease is also inversely correlated with the size of the repeat. Anticipation of SCA31 has only been reported in one family (Sato et al., 2009). SCA36 has shown limited anticipation, but without a statistically significant correlation of repeat size and age of onset as of yet (García-Murias et al., 2012).

Repeat instability is less likely to occur when the underlying STR is non-pure. For example, SCA17 does not usually undergo expansion due to an interrupted repeat. When the SCA17 repeat is more pure due to variation, expansions of the repeat are more likely (Gao et al., 2008).

1.5.2 Discovery of repeat expansion disorders

The first repeat expansion disorder to be identified in humans was fragile X syndrome (FRAXA, OMIM #300624) in 1991. Previous studies had mapped the locus to Xq27 (Camerino et al., 1983; Szabo et al., 1984). The repeat expansion was discovered with polymerase-chain reaction (PCR) to find unstable sequences that would break. This mapped to the repeat sequence in the fragile X mental retardation (FMR1) gene (Kremer et al., 1991).

Evidence that repeat expansions were the causative mutation in disease was solidified with the discovery that a coding CAG repeat expansion in the androgen receptor gene (AR) caused X-linked spinal and bulbar muscular atrophy (SBMA, OMIM #313200) (La Spada et al., 1991), a CTG repeat expansion in the 3′UTR of dystrophia myotonica protein kinase gene (DMPK) caused myotonic dystrophy-1 (DM1, OMIM #160900) (Brook et al., 1992), and a CAG repeat in the coding region of exon 1 of the Huntington gene (HTT) caused Huntington Disease (HD, OMIM #143100) (The Huntington’s Disease Collaborative Research Group, 1993; Budworth and McMurray, 2013). Each of SBMA, DM1 and HD was first mapped through linkage analysis (described in detail in Section 4.2.2 and (Ott, 1999)), with the genetic disruption discovered by PCR, which found that affected individuals had longer STR alleles compared to controls.
Chapter 1. Introduction

The first spinocerebellar ataxia discovered to be caused by a repeat expansion was SCA1 (OMIM #164400), in the ataxin 1 gene (ATXN1) (Orr et al., 1993). This was followed by Machado-Joseph disease (SCA3, OMIM #109150) (Kawaguchi et al., 1994), then SCA2 (OMIM #183090) (Pulst et al., 1996). Repeat expansions causing a spinocerebellar ataxia have so far only been discovered through linkage mapping and PCR. No repeat expansions have been discovered with MPS thus far (Didonna and Opal, 2016). Friedreich ataxia (FRDA, OMIM #229300), an autosomal dominant ataxia with similar symptoms to spinocerebellar ataxia, was also discovered through linkage mapping and PCR. For approximately 98% of cases (Delatycki et al., 1999), it was found to be caused by a GAA repeat expansion in the first intron of the frataxin gene (FXN) (Campuzano et al., 1996).

Frontotemporal dementia and/or amyotrophic lateral sclerosis (FTDALS1\(^5\), OMIM #105550) is the first disorder discovered to be caused by a repeat expansion with MPS. This was found by detecting six variants within a 30 bp locus, something often reflective of variant calling errors. This lead to manual realignment of reads to discover a CCGGGG repeat in the first intron of the chromosome 9 open reading frame 72 gene (C9orf72). A novel repeat expansion has not been discovered by either: methods designed for STRs and their expansions, or by calling an insertion in a vanilla pipeline at an STR locus that is indicative of an expansion.

1.6 Current methods for analysis of STRs and expansions in MPS data

Several existing software packages can size STR alleles, that is, genotype the STR loci. We have divided these into three categories: (i) STR specific methods that use the repeat structure in some way to make repeat size calls such as lobSTR (Gymrek et al., 2012), RepeatSeq (Highnam et al., 2012), STRViper (Cao et al., 2013), HipSTR (Willems et al., 2017), STRait Razor (Warshauer et al., 2013) and GenoTan (Tae et al., 2014) (and an unreleased package by Yan et al. (2016)); (ii) methods to detect expansions, such as ExpansionHunter (Dolzhenko et al., 2016), exSTRA (the software developed in this thesis) (Tankard et al., 2017) and STRetch (Dashnow et al., 2017); and (iii) general MPS variant calling methods, where indels can indicate a change in repeat size from the reference, such as

\(^{5}\)Note that some plots in this thesis refer to FTDALS1 as FTDALS.


GATK HaplotypeCaller | Non-STR specific method. Local de-novo assembly of haplotypes from aligned data to call indels. | General method with few assumptions about the repeat; takes account of any other structure in the data such as non-pure repeats. | Requires STR to be fully contained within single-reads in pure-repeats. Requires interpreting indels.

Dindel | Non-STR specific method. With aligned data, uses Bayesian methods that realigns reads (and their unmapped mates when applicable) to candidate haplotypes. | Fastest from data to results. Aligns reads by STR without aligning non-STR reads. Takes into account polymerase-stutter in sequencing. | Requires STR to be fully contained within single-reads. Often reads aren't aligned on STR loci where general aligners have. Without the full reference genome to align to, there may be some reads that are mismapped but belong to the masked part of the genome.

lobeSTR | Aligns reads around known STRs by flanking sequence, with repeat number derived from the repeat number within aligned reads. | Takes into account polymerase-stutter in sequencing. | Requires STR to be fully contained within single-reads.

HipSTR | Haplotype-based method for genotyping. | Performas phasing. Associated publication reports high accuracy compared to other tools. | Limited to read length.

RepeatSeq | Takes as input aligned reads after duplicate marking and local-realignment with the GATK IndelRealigner tool. Non-spanning reads are filtered and a genotype is called with a Bayesian approach, assuming a diploid organism by default. | Fastest post-alignment method. Takes into account polymerase-stutter in sequencing. | Requires STR to be fully contained within single-reads.

STRViper | With aligned reads, statistically estimates repeat number from insert sizes of paired-end reads. | Should be able to size the largest repeats of all these algorithms. Allows STR to be longer than a single read. | Requires STR to be fully contained within paired-end reads. Paper only describes STR calling within a haploid bacteria sample so may have more difficulty in human data.

STRait Razor | Tool to interrogate STRs used in forensic testing by finding distance between flanking sequences in reads. | Focus on forensic loci. | Manual inspection to call genotypes, limited loci, limited to read lengths.

Genotan | Uses a discretized Gaussian mixture model to distinguish variants from noise, to call genotypes. | Incorporates base and mapping quality scores. Can classify expanded STRs at the FTDALS1 locus. | Limited by read length.

ExpansionHunter | With PCR-free WGS, assesses repeat-spanning and in-repeat reads with mates mapped with high quality within 1 kb of the STR. | Incorporates base and mapping quality scores. Can classify expanded STRs at the FTDALS1 locus. | Paper assesses on the longer repeat motif (GGCCCC) of FTDALS1 and were able to include paired-end reads entirely within the repeat.

exSTRa | Recovers unmapped and mismapped reads by mates which are close to the STR. Counts the repeat motif content to compare cases to controls. | Can recover unmapped reads by making use of paired-end information. Robust to impure repeats and sequencing errors. | Requires control data. Requires the aligner mapping an anchor read close to the STR, and not undergoing interference by the STR itself.

STRetch | Recovers reads with STRs to a decay reference of repeats. Cases are compared to controls for the number of reads mapping to decay reference with a robust z-score. | Can recover unmapped reads by making use of paired-end information, explicitly investigating reads that map to STR decay reference sequences. | Requires control data: Repeats shorter than the read length may not be detected. Relies on the aligner mapping anchor reads correctly.

TRhist | Generates histograms of STR for long-read technologies. | For the correct data. | Requires sequencing over the known STR site.

ReviSTER | Pipeline that maps with BWA and BLAT, then revises mismapped or partially misaligned reads at STR loci. Doesn’t directly give allele calls that need to be derived from alignments. | Recovers unmapped and misaligned reads. | Requires STR to be fully contained within single-reads. Doesn’t give STR calls directly.
Chapter 1. Introduction

STR specific detection methods benefit from being able to use the repeat structure, and may explicitly model errors associated with repeats such as polymerase stutter. Often these perform faster than general indel variant detection software, as they only interrogate STRs, usually from a precompiled database, rather than the entire genome.

Most current software that interrogate STRs with MPS is focused on genotyping repeats exactly (i.e. determining the precise number of repeats for both alleles). This is always limited by the read insert size and usually by the size of a single read. As repeat size pathogenic coding SCAs can be in excess of 300 repeat units (900 bp), this is unsatisfactory for identifying the disease-causing repeat in diagnosis. The major goal in this thesis is to be able to discriminate healthy repeat sizes from pathogenic repeat sizes; we aim to treat this problem as an outlier detection exercise.

Published methods to detect expansions have been limited, and none have successfully detected a novel repeat expansion locus that was not already implicated in a repeat disorder. ExpansionHunter (Dolzhenko et al., 2016) has been tested on the C9orf72 repeat expansion FTDALS1 locus, and classified the largest allele as either pathogenic or normal, with 99.5% (212/213) sensitivity and 100% (2,788/2,788) specificity. In this case the pathogenic size was known a priori, but this is something we do not know for novel loci. We may however, be able to guess what is in the pathogenic range by whether the allele is unusual in comparison to what is observed in healthy individuals. Additionally, the location in the genome will be important, because coding regions are less tolerant of large expansions compared to intronic STRs that can generally have more repeats before a pathogenic effect is observed.

The STRetch (Dashnow et al., 2017) algorithm adds a decoy reference of repetitive sequences, encompassing all possible STR motifs, to the genome reference for alignment. Reads that map to STR decoys are extracted, and are allocated to STR loci by the alignment of their mates. At each STR locus, expansions are called with a robust z-score comparing the number of allocated reads for cases and controls. STRetch was tested on eight different repeat expansion disorders in nine affected individuals. Three out of five known affected individual samples were correctly identified (being successful in one SCA1, one SCA3 and one SBMA patient, while failing in one SCA6 patient and one SBMA patient), while the four blind samples were correctly identified by the disease locus with the smallest p-value (FTDALS1, DM1, DM2 and FRDA). Note that FRDA was not initially
detected as the underlying STR is not present in Tandem Repeat Finder (TRF) due to its short length of 6 repeats, but was detected after being added to the input data for STRetch. STRetch was only demonstrated on WGS with PCR-free library preparation.

General methods benefit from their lack of assumptions of STRs being tested but may fail to take into account the error profiles of these locations caused by polymerase stutter. Methods such as GATK HaplotypeCaller and Dindel that reassemble or realign the MPS data can take advantage of any local genomic structure in the sample genome and can perform better than STR specific algorithms at some locations. Methods such as SAMtools mpileup and GATK UnifiedGenotyper, which consider reads as multiple lines of evidence in a Bayesian genotype likelihood model, are likely to perform poorly, as they are unable to use additional local genomic structure information.

It is expected that future technologies will have longer reads that will span even very large STRs, such as PacBio’s long-read technologies that can have reads in excess of 20 kb or Oxford Nanopore with reads of 200 kb. These will make the STR size problem that repeat expansions cause to MPS methods less relevant, as it is expected that STR sizes could be read more directly off reads than we can with Illumina HiSeq data at present. It is still to be seen how well these technologies will perform in practice, and how they will compete in price. Furthermore, we have many samples with various diseases sequenced with Illumina sequencing that are unsolved, and any caused by repeat expansions may be solved with our current data by methods developed specifically for short read sequencing. Additionally, data repositories containing MPS data from previous studies, such as NCBI’s Database of Genotypes and Phenotypes (dbGaP) (Mailman et al., 2007; Tryka et al., 2014), contain thousands of WES and WGS data worth millions of dollars that could be reanalysed with these methods.

1.7 Thesis outline

Chapter 2 presents the methods developed for analysing STR expansions, including details on samples with known repeat expansions that underwent sequencing.

Chapter 3 presents the results of Chapter 2, including basic metrics of the sequencing data and the ability to identify expanded samples from our data.
Chapter 4 presents a study of a family with members affected by an autosomal dominant spinocerebellar ataxia with variable age of onset that was suggestive of anticipation, a feature observed in most repeat expansion disorders. Linkage analysis ruled out all known SCA loci except for SCA25, but did not rule out the possibility of a novel locus with a total of four linkage regions. Basic repeat expansion detection was performed heuristically and by visual assessment. We did not attempt detection with methods presented in Chapter 2 as these have not yet been sufficiently developed to overcome multiple testing penalties associated with testing thousands of STR loci.

Finally, Chapter 5 gives a discussion of repeat expansion detection in MPS, including challenges and further directions.
Chapter 2

Methods to detect repeat expansions in MPS data

In this chapter, we explore methods to detect repeat expansions in MPS data. Many expansions are larger than one end of an Illumina HiSeq X Ten read (150 bp as standard). As such, we cannot expect to be able to accurately determine the STR length. One approach is to estimate the size of the expansion. This is the focus of currently published methods. In the first section of this chapter, Section 2.1, we have taken a different approach by treating this as an outlier detection problem of the distribution of the number of repeats in reads surrounding a locus of interest, the goal being to classify a sample at a locus as expanded or normal, for many loci.

The subsequent section describes generation of our data and preparation for further analysis steps. The next sections outline our comparison to existing methods, and the performance of these algorithms to detect repeat expansions.

2.1 exSTRa: detecting outlier read length data

Current repeat expansion detection methods have relied on inference of allele sizes, that may be difficult to accurately determine from MPS data. Furthermore, allele calling for repeat expansions is inherently fraught with difficulty. At best, one can detect a bound for the repeat expanded allele, but not its true size. Given the difficulties in determining alleles at repeat expansion loci, we decided to look at the underlying data instead: the
repeat content within each read. This section details our methods for detecting repeat
expansions with this type of data.

Our algorithm has two parts. The first part (1) (Section 2.1.1) involves extracting repeat
information from reads. The second part (2) (Section 2.1.2) involves the visual and statistic-
al analysis of part (1) data to call outlier alleles.

2.1.1 exSTRa: Detecting expansions by repeat motif score

As alignment can be inaccurate or soft-clipped around STRs, and determination of STR
length from the alignment still requires confirmation that insertions form part of the repeat,
that is instead easier to determine the repeat length for each read directly from the read
sequence. As determining the start and end of an STR in a read is computationally
intensive and can be inaccurate due to sequencing errors and impure repeats, particularly
at the ends of reads, we simply summarise the proportion of the read showing the repeat
motif (at any starting base). This approximates the actual STR length, with mutations
and impure repeats causing underestimates, and matches to the motif outside STR leading
to overestimates. As we are comparing cases to controls, systematic bias at a locus should
be accounted for, whilst still encapsulating the signal from expanded STRs.

Read information was extracted from a database of STR locations, such as 2–6 bp repeat
unit features from the Simple Repeats track of UCSC Genome Browser. We extract infor-
mation for each STR at a time, repeating the following algorithm for each STR:

1. We first identify ‘anchor’ reads, that assist in identifying reads within or overlapping
the STR. To qualify as an anchor, we require that reads are mapped within 800 bp\(^1\)
of the STR, with the anchor orientated towards the STR. An anchor may overlap
the STR (Figure 2.1). It is possible that both ends of a read pair are identified as
anchors, in this case, the following procedure is repeated for both ends with both
results included that can be resolved during the statistical analysis of the data (the
implementation in this thesis did not investigate resolving these further, with both
ends left in the analysis if any).

\(^{1}\)800 bp is chosen to avoid discarding reads overlapping the STR, with the insert size of read pairs
having median approximately 360 bp. Some protocols may need to analyse reads further than 800 bp and
can be adjusted when calling the Perl module.
2. The anchor-mate mapping is checked. If the anchor-mate is mapped near the STR and not overlapping or adjacent, then the read is discarded, while those reads overlapping the STR are taken forward to the next analysis step (Figure 2.1). Sometimes the read is unmapped, or mapped to another locus, which is then recovered for further interrogation in the next step.

3. Remaining anchor-mates have their sequence content matched for the presence of the repeat unit in the correct direction, allowing for the repeat to start at any base, or phase, of the repeat unit. For example, if the repeat unit is CAG, we also match AGC and GCA. The number of bases designated as part of the repeat unit is counted to derive a repeat score for that read, that we denote at a given locus as $x_{ij}$ for sample $i$ and read $j$ (note that the maximum defined $j$ depends on the sample). As an example, matching a CAG repeat on the opposite strand, thus CTG, and equivalently TGC and GCT, is:

\[
\text{CGTTCA}cctgGATGUTGAAACTctgTCctgA\text{TAGTCCOCCCTgctgctgctgctgctgctgTbgtgctgTTbctgctgTctgAAA}
\]

where matches to the CTG repeat are red and lower case. This 87 bp sequence has 51 bp marked as part of the repeat.

4. We filter out reads where the score was not greater than expected in random base sequences. While not precisely true, we assume the four bases are uniformly distributed and independent with respect to other positions. Short motifs are more likely to appear by chance. We filter out scores where $x_{ij} \leq lk/4^k$, where $l$ is the read length and $k$ is the motif length.

Note that we do not attempt to make use of read-pairs that are entirely within the repeat, as these are usually difficult to map correctly due to their repetitive nature. ExpansionHunter, that we discuss in Section 2.3, attempts to use these reads for the FTDALS1 locus where there are only a few STRs with the same repeat motif.

### 2.1.2 Visualisation of repeat scores

Empirical cumulative distribution function (ECDF) plots of repeat scores were generated to assist in developing appropriate statistical classifiers. The use of ECDFs to visually detect expansions meant we did not need to normalise our data for coverage, but coverage
The top panel shows the expanded haplotype of a hypothetical sample, with the origin of sequenced paired-end reads, labelled i to vi. The normal size allele is not shown for the sample. The bottom panel is the reference sequence with the same reads aligned. Reads inside the STR, such as vi, will usually not be mappable due to their repetitive nature. Some reads (ii, iv) are soft-clipped (orange) due to the repeat within reads being longer than the STR on the reference; these bases are not aligned but may be indicative of a repeat expansion. Anchor reads (*) are identified as those aligned within 800 bp of the STR, where the read faces towards any part of the STR from the end of the read. Reads such as i have no qualifying anchor reads (as the forward read is too far from the STR) and are not pursued. Note that for some reads such as iii, both ends are identified as anchors, and the following procedure is repeated for both ends. Detection reads are the mates of anchor reads. In this way, detection reads may be recovered as unmapped or mapped to another locus. If the detection read is aligned within this region, without overlapping or bordering the STR, it is unlikely to have originated from the STR and therefore filtered out (i.e. iv for the rightmost read as a detection read, and v for the leftmost read as a detection read). The remaining detection reads (blue) then undergo counting of the repeat motif.

still influenced the ECDF accuracy. With increased coverage, the ECDF would converge towards the cumulative distribution function and have less variability.

We first visually inspected ECDF plots to identify expanded samples. Here, expanded samples were those with more reads (to the right) with higher repeat scores on the upper part (dominant) or the whole graph (recessive and males in X-linked disorders) (Figure 2.2). As we were blind to the underlying expansion in cohorts WES_PCR and WGS_PCR_2, we were able to assess the reliability of our visual outlier detection.

For simplicity, the following description of the data and analysis methods is only for a single locus. Our classification problem consisted of detecting, for a group of N samples, when a particular sample was an outlier compared to the others (the background distribution). As the background, we could either specify a precise group of controls, or, as presented in this thesis, use all other samples with trimming of extremes to form a null hypothesis distribution.
Chapter 2. Methods to detect repeat expansions in MPS data

Methods to detect repeat expansions in MPS data

33

Figure 2.2. ECDF visual classification

(a) HD As Huntington disease (HD) is a dominant disorder, we expected higher on-average repeat scores in approximately half of the reads. This was observed for sample rptWEHI3 (dark blue) with the ECDF curve more to the right than other samples. Hence we visually classified rptWEHI3 as having a HD expansion. (b) As Friedreich ataxia (FRDA) is a recessive disorder, we expected higher on-average repeat scores in all reads. This was observed for samples WGSrpt_09 and WGSrpt_11, and hence visually classified as having expansions.

16 of 21 known expansion disorder loci are autosomal dominant, where approximately half of the reads come from the expanded allele. Therefore we reason that we should only keep the highest 50% of repeat scores for each sample at each locus. This is a crude assumption as the expanded haplotype is longer and therefore has more DNA to be sequenced from it, but this was also offset by greater difficulty in alignment due to the repetitive nature of the STRs. Additionally, longer repeats change GC content, thus leading to repeat expansion allele specific bias in sequencing. These examples may reduce the ability to sequence these repeats for the expanded allele: reads with only CAG repeats have 67% GC content, whereas those from AT-rich repeat motifs, such as the GAA intronic repeat of Friedreich ataxia (FRDA) will have 33% GC content in reads containing only the repeat motif.

For \( N \) samples, the repeat score data is in a vector:

\[
x = (x_1, \ldots, x_i, \ldots, x_N)
\] 

(2.1)
where the repeat score data for sample $i$ with $m_i$ reads is defined as:

$$
\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{im_i})
$$

(2.2)

where $x_{ij}$ is the repeat score for sample $i$ and read $j$. For simplicity, we order the repeat score data in $\mathbf{x}_i$ such that subsequent indices are monotonically increasing, that is:

$$
x_{i1} \leq x_{i2} \leq \ldots \leq x_{im_i}
$$

(2.3)

At a locus, we have a total of $\sum_{i=1}^N m_i$ observations (reads).

### 2.1.3 A simple one-sided, Mann-Whitney two-sample test

Classic two-sample tests, such as the Student’s t-test or the non-parametric Mann-Whitney, only test up to two groups. To perform these tests, we have one sample that we declare as the subject, and compare this to the combined scores of the other samples, that we declare the background. As we are looking for samples with expanded alleles, rather than both expanded and contracted, we perform tests only one-sided, with the alternative hypothesis that the subject sample has more repeat motifs contained within its reads, and thus an expanded allele.

As the data with background samples combined does not appear to be normally distributed, an assumption of the Student’s t-test, we perform one-sided two-sample Mann-Whitney tests that makes fewer assumptions about the underlying distribution. Each sample is declared the subject in turn, to see if the repeat scores are greater in the subject compared to the background of the other $N - 1$ samples (regardless of case or control grouping). The test statistic $U_i$ is derived using R’s `wilcox.test()` function, as:

$$
U_i = MW(\mathbf{x}_i, \mathbf{x}_{-i})
$$

(2.4)

Where $MW()$ is the two-sample Mann-Whitney test-statistic function, and $\mathbf{x}_{-i}$ indicates all data points except those from sample $i$. In most cases, the R `wilcox.test()` function will derive p-values using a normal approximation, as we have at least 50 data points (this is only calculated exactly for less than 50 observations and no ties).
Chapter 2. Methods to detect repeat expansions in MPS data

The Mann-Whitney and Student’s t-test assume observations are independent and identically distributed (i.i.d.), but our data is structured with $N - 1$ background samples. Ignoring that some reads observations may be counted twice due to pairing (affecting less than 5% of data), the observations from each sample come from the same underlying alleles, but from independently sequenced reads. Hence, observations are approximately i.i.d. from the same sample. As different samples can have different alleles, these were independent, but came from a different distribution. As such, we did not expect the Mann-Whitney test statistic ($U_i$) to follow the assumed distribution. Thus, we assessed the resulting p-value distribution. If the assumptions were not met for the test, then we expected that the p-values would not be uniformly distributed over the interval $[0,1]$.

2.1.4 exSTRa: an aggregated t-statistic

To incorporate the structure of multiple background samples to compare against, we propose a test statistic, $T$, that is the average of multiple t-statistics, with each arising from a quantile of each sample at the same probability. An example of the following steps is given in Figure 2.3.

As the number of observations varies between samples, we compare the subject and background samples at the same quantiles. Here, we set the number of quantiles, or the number of t-statistics that contribute to the average T test statistic, as the largest number of reads seen in any of the $N$ individuals in the cohort at this STR locus:

$$M := \max_{1 \leq i \leq N} \{m_i\} \quad (2.5)$$

but this can also be set to another number, such as the median of the number of reads from all the samples at the STR locus being investigated. For a given numeric vector, R’s `quantile()` function produces sample quantiles at the given probabilities, with nine different algorithms for calculating these quantiles. With $Q_i(p)$ as R’s `quantile()` function for data $x_i$, we define a quantile matrix:

$$Y_{N \times M} = (y_{ij}) := (Q_i(p_j)) \quad (2.6)$$

where the probabilities are:

$$p_j = \frac{j - 1}{M - 1}, 1 \leq j \leq M \quad (2.7)$$
Chapter 2. Methods to detect repeat expansions in MPS data

Figure 2.3. Sample quantiles across samples

A subset of cohort WGS_PCR_2 data of five samples with reduced read depth. (a) The ECDF of this example data (b)–(c): the ECDF of each sample in is shown as a line without points to mark the steps (unlike other ECDFs shown in this thesis). The sample quantiles at fixed probabilities are shown as solid circles. Note that for the sample with the largest number of reads, WGSrpt_12, the probabilities correspond to the middle of each jump. A quantile lower-limit of $h = 0.5$ is chosen as default (dashed horizontal line), such that only quantile probabilities above this level are used in the test statistic. (b) The sample quantiles derived for each sample as solid circles. (c) To derive the background distribution, a trimming function is applied at each probability level. Remaining observations are solid circles.

Samples with less than 3 observations at this locus are discarded and not included in $Y$.

To be robust to additional cases appearing in our data, we define a trimming function, $V_v(y)$, that, given a proportion $v < 0.5$, removes $\lceil v/M \rceil$ of the highest observations and
[\lceil v/M \rceil] of the lowest observations from a vector \( y \) (Where \( \lceil \cdot \rceil \) is the ceiling function). We set \( v = 0.15 \) for cohorts WES_PCR, WGS_PCR_1 and WGS_PCR_2. As there were 3 of 8 cases of FRAXA in WGS_PF_3, we set \( v = 0.37 \) in this cohort that trims the highest (and lowest) 3 observations at each quantile. Note that the observation from the subject sample will often remain in the background samples.

We set a quantile lower-limit \( h = 0.5 \), such that we only perform our test on quantile probabilities above this limit. We did not explore the most optimal value for \( h \), or if the quantiles should be weighted; this is future work. Our test statistic \( T_i \) for sample \( i \) as the subject is defined as:

\[
T_i := \sum_{j > hM}^M \frac{t(y_{ij}, V_v(y_j))}{M(1-h)}
\]

(2.8)

where \( \lfloor \cdot \rfloor \) is the floor function, and \( t(\cdot, \cdot) \) is the t-statistic, assuming equal variance between groups as we had a single subject observation, precisely:

\[
t(u_0, u) = \frac{u_0 - \bar{u}}{S}
\]

(2.9)

where \( \bar{u} \) is the sample mean of \( u \), and

\[
S = s_u \sqrt{1 + \frac{1}{n_u}}
\]

(2.10)

where \( s_u \) is the sample standard deviation of \( u \). Consequently, higher values of \( T_i \) reflect longer underlying alleles and lower \( T_i \) reflect shorter underlying alleles.

We generate a \( T_i \) statistic for each sample, for each locus.

### 2.1.5 exSTRa: p-values for the aggregated t-statistic

Due to the set up of testing 1 vs N, permutation testing of the T statistic could not be performed as the number of permutations would be limited to \( N \) with smallest possible p-value \( \frac{1}{N+1} \). Bootstrap techniques were attempted at each locus to increase the number of possible unique resamplings. Here, each of the \( N + 1 \) bootstrap observations were derived from a random patient at the locus of interest. This did not lead to small p-values (<0.01), even for expansions. This failed because the bootstrap was not generating a null distribution when the data included an expansion at the locus being tested.
We derive p-values for the $T_i$ statistic empirically by simulating $B$ additional datasets under the null hypothesis, by making use of the empirical distribution contained in the read data.

We set:

$$\hat{\mu}_j = \text{median}\{y_{ij}\} \quad (2.11)$$

and

$$\hat{\sigma}_j = \frac{1}{(\Phi^{-1}(3/4))} \text{MAD}\{y_{ij}\} \quad (2.12)$$

where $\text{MAD}$ is the median absolute deviation and $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard normal distribution. R’s $\text{mad()}$ function incorporates this scaling factor by default. Note that for a normal distribution with mean $\mu_j$ and variance $\sigma_j^2$, $(\hat{\mu}_j, \hat{\sigma}_j)$ will asymptotically converge towards $(\mu_j, \sigma_j)$.

We simulated matrices $X^k$ for $k \in 1, 2, \ldots, B$ by:

$$x_{ij}^k := \hat{\sigma}_j z_{ij}^k + \hat{\mu}_j \quad (2.13)$$

where each $z_{ij}^k$ was a randomly generated realisation from a standard normal distribution using R’s $\text{rnorm()}$ function. As this may result in a matrix where each row may no longer be monotonically increasing, as expected for quantiles, we sort each row to form $Y^k$:

$$y_{ik}^k := \text{sort}\{x_{ik}^k\} \quad (2.14)$$

where $\text{sort}\{\}$ reorders its given $m$ elements such that $\text{sort}\{x\} = (y_1, y_2, \ldots, y_m)$ satisfies $y_1 \leq y_2 \leq \ldots \leq y_m$.

For each $Y^k$, we calculate the statistic for a single sample $T_1^k$ (choice of sample is not important as all samples are simulated under the same distributions). The one-sided p-value for a test statistic $T_1$ is:

$$P(T_i) = \frac{1 + \sum_{k=1}^{B} \mathbb{1}[T_i > T_1^k]}{B + 1} \quad (2.15)$$

where $\mathbb{1}[]$ is the indicator function. Setting $B = 10^i - 1$ gives p-values that are decimal fractions. In this study, we set $B = 10^6 - 1$. 

This p-value derivation is able to take into account locus specific variability and bias that may not be reflected in the raw $T_i$ statistic.

### 2.1.6 Software implementations

The algorithms implemented here are available in the R package **exSTRa** (expanded STR algorithm), available from [https://github.com/bahlolab/exSTRa/](https://github.com/bahlolab/exSTRa/), and a Perl module called Bio::STR::exSTRa available at [https://github.com/bahlolab/Bio-STR-exSTRa/](https://github.com/bahlolab/Bio-STR-exSTRa/).

The Perl module Bio::STR::exSTRa implements the algorithms described in Section 2.1.1 to extract read information from BAM or CRAM files. This also reads in an STR database, that is either the Simple Repeats track download from the UCSC Genome Browser (intended for analysis of STRs not yet implicated in disease), or an Excel or tab-delimited file that is a list of known repeat expansion loci. The package then provides functions to generate the repeat score that is output to a tab-delimited text file to be read with the exSTRa R package.

The R package exSTRa implements the methods of Sections 2.1.4 and 2.1.5. This reads in the same STR database as Bio::STR::exSTRa, providing functions to visualise the data, perform statistical analysis and identify outlier alleles to classify expansions. The exSTRa package makes use of the R data.table package ([Dowle and Srinivasan, 2017](https://doi.org/10.1093/bioinformatics/btw560)) to store and manipulate data.

### 2.2 Data gathering and preparation

#### 2.2.1 Cohorts of MPS data for STR investigation

In order to develop our repeat expansion detection method, we firstly collected samples affected by known repeat expansion disorders. Study approval was provided by the Royal Children’s Hospital Research Ethics Committee (Human Research Ethics Committee #28097, #29077, #25043 and #22073), with informed consent provided by participants or their guardians. Clinical diagnosis was obtained by review of medical records and examination of participants. Here, samples with a known repeat expansion were our cases and those without were controls. As each case had only one expansion, we also included the case samples as controls for other loci during performance testing.
Chapter 2. Methods to detect repeat expansions in MPS data

Agilent SS = Agilent SureSelect Human All Exon; AGRF = Australian Genome Research Facility, Melbourne; KCCG = Kinghorn Centre for Clinical Genomics, Sydney; PE = paired end. Library preparation is TruSeq Nano = Illumina TruSeq Nano DNA Library Preparation Kit, Hyper PCR-Free = KAPA Hyper Prep Kit with PCR-free workflow. *Fourth WES sample obtained at later date.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Regions</th>
<th>Centre</th>
<th>Sequencer</th>
<th>Library prep</th>
<th>Read length</th>
<th>Cases</th>
<th>Controls</th>
<th>Date received</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>WES_PCR</td>
<td>Agilent SS</td>
<td>AGRF</td>
<td>Illumina HiSeq 2500</td>
<td>100 bp PE</td>
<td>4</td>
<td>58</td>
<td></td>
<td>5-Jun-2014, 7-Nov-2014*</td>
<td>Controls obtained at multiple dates</td>
</tr>
<tr>
<td>WGS_PCR_1</td>
<td>Genome</td>
<td>KCCG</td>
<td>Illumina HiSeq X Ten</td>
<td>TruSeq Nano</td>
<td>151 bp PE</td>
<td>3</td>
<td>14</td>
<td>7-Nov-2014</td>
<td></td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>Genome</td>
<td>KCCG</td>
<td>Illumina HiSeq X Ten</td>
<td>TruSeq Nano</td>
<td>150 bp PE</td>
<td>16</td>
<td>2</td>
<td>17-Jan-2017</td>
<td>14 cases at 60x, 2 cases at 30x</td>
</tr>
<tr>
<td>WGS_PF_3</td>
<td>Genome</td>
<td>KCCG</td>
<td>Illumina HiSeq X Ten</td>
<td>Hyper PCR-Free</td>
<td>150 bp PE</td>
<td>3</td>
<td>5</td>
<td>30-Mar-2017</td>
<td>3 cases are replicates of WGS_PCR_2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>MOI</th>
<th>Diagnosis</th>
<th>Allele Sizes</th>
<th>Sex</th>
<th>WES_PCR</th>
<th>WGS_PCR_1</th>
<th>WGS_PCR_2</th>
<th>WGS_PF_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PolyQ</td>
<td>AD</td>
<td>SCA1</td>
<td>36, 52</td>
<td>female</td>
<td>WGSrpt 14</td>
<td>WGSrpt 16</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SCA2</td>
<td>21, 42</td>
<td>female</td>
<td>WGSrpt 18</td>
<td>WGSrpt 20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SCA6</td>
<td>11, 22</td>
<td>female</td>
<td>WGSrpt 05</td>
<td>WGSrpt 07</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SCA7</td>
<td>13, 39</td>
<td>female</td>
<td>WGSrpt 08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S’UTR</td>
<td>X</td>
<td>FRAXA (pre)</td>
<td>~100</td>
<td>female</td>
<td>WGSrpt 21</td>
<td>WGSrpt 21_F</td>
<td>WGSrpt 19_F</td>
<td>WGSrpt 19_F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FRAXA</td>
<td>613–1680</td>
<td>male</td>
<td>WGSrpt 17_F</td>
<td>WGSrpt 19_F</td>
<td>WGSrpt 19_F</td>
<td>WGSrpt 19_F</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DM1</td>
<td>8, 173</td>
<td>female</td>
<td>WGSrpt 13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FRDA</td>
<td>320, 320</td>
<td>male</td>
<td>WGSrpt 15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>788, 788</td>
<td>male</td>
<td>WGSrpt 11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(controls)</td>
<td></td>
<td></td>
<td></td>
<td>58</td>
<td>14</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

Our data were gathered into four cohorts (Table 2.1). Table 2.2 lists the disorders being investigated and which individuals were replicated across cohorts. Replicated individuals were given different sample names between cohorts. In samples with a known repeat expansion, we (Melanie Bahlo and I, the bioinformaticians) were blinded to the specific repeat expansion in cohorts WES_PCR and WGS_PCR_2.

The WES_PCR cohort cases consisted of four WES samples with different expansion disorders, HD, SCA1, SCA2 and SCA6. Each patient had blood collected and DNA extracted. The DNA was enriched for the exome with the Agilent SureSelect Human All Exon v5+ UTR kit, and sequenced 100 bp paired-end to 50x coverage on the Illumina HiSeq 2500 sequencer, at the Australian Genome Research Facility (AGRF) Melbourne. We made use of 58 samples from other internal projects which were affected by other...
diseases. These WES samples underwent the same capture and sequencing at the same facility, but at varying times. Some of these controls were related individuals, with the largest family having 9 members. As our focus was on learning as much as possible about the background distribution of STR alleles, including technical artefacts, we did not filter out or account for related samples.

Additionally three cohorts (WGS_PCR_1, WGS_PCR_2 and WGS_PF_3) of whole-genome sequencing (WGS) were performed on the Illumina HiSeq X Ten sequencer at the Kinghorn Centre for Clinical Genomics (KCCG), Garvan Institute of Medical Research, Sydney.

The WGS_PCR_1 cohort included three cases, one each affected with HD, SCA2 and SCA6, sequenced to 60x coverage. As controls, fourteen samples from other projects were used to understand the background distribution of population alleles, although nine of these were performed at only 30x coverage (one sequencing unit). We were not blind at any point to the underlying disorders in this cohort. Library preparation was with the Illumina TruSeq Nano DNA Library Preparation Kit. Control samples not affected by ataxia were not tested with gold-standard methods for repeat expansions at the known loci, but as they did not show symptoms for these disorders they were unlikely to harbour repeat expansions at the 21 loci. The largest number of samples in the controls from any one pedigree was three. Included in these fourteen controls were two patients affected with spinocerebellar ataxia from the A1 family discussed in Chapter 4 (samples A1-11 [II-3], A1-2 [III-2] (Chapter 4 name in square brackets)); we were able to use them as controls here as the SCA loci we were investigating had been tested and known not to be causing SCA in this family. WGS_PCR_1 data were provided as 151 bp paired-end instead of the usual 150 bp. In Illumina sequencing, each base is calibrated by its following base, such that 151 bp must be sequenced in order to provide 150 well calibrated bases, and usually the 151st base is not provided. As it was provided in this case, the 151st base is likely to be of much lower quality than preceding bases. However, we did not remove this last base reasoning that the information was still useful for our work.

The WGS_PCR_2 cohort consisted of 16 cases with a known repeat expansion and 2 controls. There were eight different disorders represented: HD (2), SCA1 (2), SCA2 (2), SCA6 (2), SCA7 (1), DM1 (2), FRDA (2) and FRAXA (3), with the number of affected/carrier individuals in brackets. We were initially blinded to the underlying
disorders. Library preparation was with the Illumina TruSeq Nano DNA Library Preparation Kit. It was intended that all the cases would be performed with two units of sequencing giving 60x coverage and controls with one unit of sequencing (30x), but due to a labelling error 14 cases and the 2 controls were performed at 60x coverage and 2 cases at 30x coverage. Fragile X syndrome (FRAXA), the most common repeat disorder, occurs in approximately 1 in 4,000 males (Turner et al., 1996). The FRAXA individuals included two affected individuals (WGSrpt_17, WGSrpt_19) and one carrier (WGSrpt_21), the mother of WGSrpt_19.

For methods that showed some success at 60x coverage, we have further explored the performance of 30x coverage. For 60x coverage samples of cohort WGS_PCR_2, the first lane (lowest numbered lane) of data were assigned to WGS_PCR_2_30x_1, while the second to WGS_PCR_2_30x_2. The two 30x coverage samples were included in both WGS_PCR_2_30x_1 and WGS_PCR_2_30x_2.

The WGS_PF_3 cohort consisted of 3 cases of FRAXA, to 60x coverage, with the library prepared with a PCR-free workflow of the KAPA Hyper Prep Kit. We were blind to the underlying disorder of these patients, but knew that they were derived from the same DNA samples as three from cohort WGS_PCR_2. At first we did not have control samples for this cohort, as such we were not able to make disorder calls while blinded. At a later date, after we were unblinded, we added 5 control samples to this cohort that were sequenced using the same protocol at the same centre (KCCG), but at a different time. These controls were shared by Katie Ayers (Murdoch Children’s Research Institute, Melbourne), who had sequencing performed on these samples for a non-repeat expansion project.

2.2.2 MPS data preparation

This subsection outlines the initial analysis steps that were applied for our repeat score methods, exSTRa and a Mann-Whitney test, as well as ExpansionHunter. We outline our exSTRa analysis pipeline for repeat expansion in Figure 2.4. This includes wet-laboratory analysis steps from a DNA sample, bioinformatic analysis to prepare the sequencing data, and analysis using our exSTRa package. The following contains methods that may be considered standard for many analysis pipelines.

---

Figure 2.4. Sample preparation and data analysis pipeline summary

This graph represents three DNA samples undergoing analysis in our pipeline. Libraries may be sequenced across multiple lanes or runs, and we analysed the FastQ data for each of these separately, then merge after alignment. Data and samples are represented by ovals, and actions represented by rectangles. Blue represents wet-laboratory, orange dry-laboratory, purple input data, and red steps in the exSTRa analysis package. Some implied analysis steps, such as BAM indexing, are not included for simplicity.
FastQC (Andrews, 2016) is a quality-control program that quickly generates summaries of quality metrics, such as base score qualities by position, base frequencies by position, GC content and k-mer biases. The program does generate PCR duplicate rate estimation, but this does not take base calling errors or paired-end information into account. We ran FastQC on each FastQ file to check that the quality was adequate and that there were no outlier samples where sequencing performed poorly.

We did not attempt to perform optimal alignments with the Needleman-Wunsch or Smith-Waterman algorithms as this was unfeasible for our MPS data sets that had millions of reads requiring alignment to a 3 billion bp reference sequence in humans. In choosing an aligner appropriate for our work, we required one that would give the correct mapping as frequently as possible. As we inspected the sequence directly, examining the reads for repetitive DNA, the local accuracy of the alignment was less important. In an expansion, STRs would be much longer in the reads and could have large insertions that in full-alignment algorithms lead to highly penalized alignment scores. Instead, we chose an aligner that performed local alignment in order to not mismap these reads, that is, bases at either end of the read may not have been aligned (soft-clipping) if this increased the alignment score. Soft-clipping usually gives a penalty to the alignment score, but less than large gap penalties (indels) and multiple mismatch penalties. It was computationally unfeasible to extensively test aligners for optimal performance for repeat expansion detection. During Daniel Cameron’s PhD work, supervised by Tony Papenfuss in the Bioinformatics Division at the Walter and Eliza Hall Institute of Medical Research, he investigated the ability of many aligners to handle large structural variation as part of his work, finding that Bowtie 2 (Langmead and Salzberg, 2012) in local mode performed well (compared to other tested software) at mapping reads where there were large insertions or deletions in the sequence. Bowtie 2 was ultimately chosen due to its ability to allow more bases to be soft-clipped when only part of the read could be aligned. The BWA (Li and Durbin, 2009) aligner’s BWA-MEM algorithm is a possible alternative that also performs local alignment, but this was not tested in this thesis and remains as future work.

All samples were aligned with Bowtie 2 to the hg19 human reference genome (as provided in the GATK bundle3 2.3). Usage of Bowtie 2 requires that the reference genome was first converted to an FM-index that is based on the Burrows-Wheeler Transform, allowing fast searching of the reference sequence. Bowtie 2 attempted to map reads by first finding

matches to the index with the high-quality end of reads, with backtracking to allow for errors and genuine variations. These matches underwent alignment at the initial mappings, with each alignment given a score. When there was a uniquely-best alignment score, then this was chosen as the alignment and a mapping score was calculated from this alignment score and the second-best alignment score. When there was more than one best alignment, usually when there were multiple positions that matched exactly, then the read was considered to be multi-mapping, and given a mapping quality less than 2. Some reads may not have been able to be mapped, and these reads were reported as unmapped. We performed alignments with Bowtie 2 in very sensitive local mode (option --very-sensitive-local), that makes the software search more aggressively for the best alignment, and allowed the alignment algorithm to perform local alignment, soft-clipping bases if this increased the alignment score. The other important option was --maxins 1000 that allowed insert sizes of up to 1000 bp to be considered a valid paired-end alignment.

Read group information was provided. The number of threads, allowing faster alignment through parallelisation, varied between analyses, but this would be unlikely to change the results. We did not investigate modifying the Bowtie 2 alignment parameters further (e.g. reducing the reference gap penalties), due to the computation and disk space required to run alignments, which can take many months and terabytes of storage, and would be for limited gain while our downstream analysis is unrefined. The alignment outputs Sequence Alignment/Map (SAM) format that we directly converted to a compressed Binary Alignment/Map (BAM) format file for writing to disk.

After alignment, reads were sorted and merged for each sample with the Novosort tool (Novocraft Technologies Sdn Bhd, 2017), creating a single BAM file for each sample that contained multiple lanes and runs if applicable. Novosort was run in multi-threaded mode, significantly speeding up sorting, but this feature required a licence.

Next, we performed local realignment. In the initial alignment with Bowtie 2, each read pair was mapped and aligned independently from the other read pairs. This could result in the same insertions or deletions (indels) being represented differently across the reads. Local realignment aimed to represent indels uniformly across reads with respect to the reference, giving a consensus indel. We performed local realignment with the GATK (DePristo et al., 2011) IndelRealigner tool.

In MPS, each base in a read was given a quality score in the FastQ file that was estimated
by the Illumina base calling software. Quality score recalibration adjusted quality scores such that they more accurately reflected the actual error rates of the sequencing. This was calculated empirically by comparing mismatches to a dbSNP-masked (Sherry et al., 2001) reference and adjusting the quality scores to reflect these rates, increasing the accuracy of variant calling. The recalibration model used several covariates including the original base quality, the preceding and current base, position within the read and sequencing run. Although we applied the GATK Base Quality Score Recalibration tool for general analysis, at present our methods currently did not make use of base and mapping quality score information, but could be incorporated into future iterations.

Some BAM files were compressed to CRAM format to save disk space. As BAM and CRAM both hold the same information, this should be inconsequential to the results.

2.3 Repeat expansion detection with existing software

We firstly examined all samples in our cohorts with one of the best known STR MPS genotype callers for Illumina data, lobSTR (Gymrek et al., 2012), as well as a very recent method for detecting expansions, ExpansionHunter (Dolzhenko et al., 2016). We later go on to compare these results to our new method, exSTRa (Tankard et al., 2017). As STRetch (Dashnow et al., 2017), another tool for calling repeat expansions, was released close to the end of this thesis work, we have not assessed its performance; comparing STRetch to ExpansionHunter and exSTRa with our samples is future work.

A newer method for genotyping, haplotyping, and phasing STRs, HipSTR (Willems et al., 2017), may produce better results than lobSTR for expanded alleles. We did not have sufficient time to analyse the performance use of HipSTR, but Dashnow et al. (2017) showed that it performed poorly on expanded STRs as in attempting to make accurate size calls, it is limited by read length.

2.3.1 lobSTR

We performed alignment and genotyping with the STR dedicated software lobSTR (version 4.0.6). The lobSTR package contains two main programs: lobSTR, that aligns reads from either FastQ or BAM, and allelotype, that derives STR genotypes. Although lobSTR
does not call expansions, we have included it as it is a well-established method for calling alleles with STR data, acting as a baseline comparison and motivating the need for more dedicated expansion detection software.

For each read pair, the lobSTR program first performed a sensing step that detected STRs within reads, excluding reads without an STR from further analysis. The flanking regions were then used to align the read to a Burrows Wheeler Transform (BWT) (Burrows and Wheeler, 1994) index specific to the inferred repeat motif, (lobSTR uses BWT indexes). Current versions of lobSTR output a BAM file, that in our workflow was sorted with Novosort.

The allelotype program gave diploid allele calls (haploid calls were also possible with options, and would be applicable for analysis of X chromosome STRs in males). Reads without unique flanking sequence of the STR within the read were not able to be used in genotyping. The length of the STR in a read is determined with:

\[ L = s - (d - u) + L_{\text{ref}} \]  

(2.16)

where \( L \) is the observed STR length, \( s \) is the length of the read, \( d \) the genomic coordinate (1-based) at the 3’ end of the alignment, \( u \) is the genomic coordinate (0-based) at the 5’ end of the alignment, and \( L_{\text{ref}} \) is the length of the STR in the reference sequence. Alleles were sized using a maximum likelihood model that takes into account polymerase stutter introduced by sequencing. The modelling of polymerase stutter was trained with male X chromosome data, as there was only one allele to be called, thus simplifying the problem with no need to decouple two allele signals. Current versions of allelotype output a VCF file.

We identified the largest allele at each locus for each sample, and called it as an expansion when within the expanded range, and as an intermediate expansion if it is larger than the expected normal range but not an expansion as defined by Table 1.2. The ability and performance of lobSTR in expansion disorders was unknown. We do however, know lobSTR only makes use of reads that span the repeat.
2.3.2 ExpansionHunter

ExpansionHunter (Dolzhenko et al., 2016) is a dedicated repeat expansion method designed for Illumina WGS data prepared with the Illumina TruSeq PCR-free Library Preparation Kit. In our case, although all our cohort data were from Illumina sequencing, none of our samples were prepared in the way as those analysed in the ExpansionHunter preprint. Our samples were either WES (WES_PCR), prepared with the Illumina TruSeq Nano DNA Library Preparation Kit (WGS_PCR_1 and WGS_PCR_2), or for PCR-free data we used a competitor library preparation protocol, the KAPA Hyper Prep Kit (WGS_PF_3). The performance of ExpansionHunter with these library preparation methods was unknown. All our WGS cohorts were sequenced on the Illumina HiSeq X. The ExpansionHunter preprint involved the study of 1,770 samples on Illumina HiSeq X machines, and another 1,231 samples sequenced on Illumina HiSeq 2000 machines. Additionally, the preprint data were aligned with the Isaac aligner (Raczy et al., 2013) instead of Bowtie 2.

The 2016 preprint of ExpansionHunter only interrogated patients with two specific repeat expansion disorders, these were (1) frontotemporal dementia and/or amyotrophic lateral sclerosis 1 (FTDALS1) in C9orf72, and (2) Friedreich ataxia (FRDA) in FXN. For FTDALS1 patients, ExpansionHunter showed an impressive 99.5% (212/213) sensitivity and 100% (2,788/2,788) specificity, but this was with a 30 repeat cutoff, which is higher than those usually observed in normal samples of 19 repeats, but much lower than the 250 repeats or more, often seen in those affected with FTDALS1. The ability to detect expansions in Friedreich ataxia was also demonstrated in eight affected individuals, but no estimate of specificity was given. The preprint was updated in July 2017, close to the submission of this thesis, which showed ExpansionHunter could successfully identify repeat expansion in PCR-free WGS for the FTDALS1, FRDA, HD, SCA1, SCA3, FRAXA, DRPLA, SBMA and DM1 loci (Dolzhenko et al., 2017).

ExpansionHunter identifies reads that either: (a) span the repeat (SPANNING reads), (b) flank one end of the STR (FLANKING reads), or (c) are fully within the repeat, called “in-repeat” reads (IRRs). The IRR strategy relies on the repeat motif being long enough such that it only occurs in a few places on the genome, as such this is only used for the FTDALS1 expansion in the Dolzhenko preprint and this thesis, with the 6-mer repeat motif GGGGCC in the first intron of C9orf72. ExpansionHunter extracts exact allele lengths for SPANNING reads (based on what is observed), and produces confidence intervals with
point estimates for FLANKING and IRR reads. The only quality metric given for allele
calls is the number of supporting reads. Even though human data were diploid, this strategy
sometimes gave more than two alleles for SPANNING reads in WGS with a PCR step in
library preparation, as ExpansionHunter did not attempt to resolve alleles that were close
(such as one repeat differences). At the time of analysis, ExpansionHunter did not provide
a method to combine allele size calls from these three strategies. ExpansionHunter also did
not provide calls of what was an expansion; the user was required to decide a threshold to
filter for expansions.

We performed analysis with the ExpansionHunter software (version 2.0.9) on the cohorts
at the 21 repeat expansion loci in Table 1.2. The input data were BAM files that were
prepared as given in Section 2.2.2. Only specification files (in JSON format) for the HD,
FTDALS1 and FRAXA loci were provided with the software. The JSON files for the other
18 loci were obtained by personal communication with Egor Dolzhenko (Illumina, Inc.
San Diego, CA, USA), the developer of the ExpansionHunter software and lead author
of the associated preprint, after providing him with the location of these loci. We ran
ExpansionHunter on all four cohorts, and the 30x coverage split of WGS_PCR_2, for
comparison. As ExpansionHunter may call more than two alleles, we only used the largest
allele call for classification of expansions. A locus was classified as expanded for a sample
when the longest allele call was within the expanded range. Additionally, we repeated the
classification with a more lenient threshold, classifying a sample at a locus as expanded
when the allele was above the normal range (intermediate or larger). Furthermore, we
inspected calls manually to ensure our thresholds were not misspecified.

We reported the sensitivity and specificity of this method, but did not construct a receiver
operating characteristic (ROC) curve as there was not an obvious predictor metric given
that the outcomes are allele calls. The number of supporting reads would likely be inappro-
priate, as the normal allele size alleles usually had many more supporting reads even when
an expansion was known. 95% confidence intervals for sensitivity and specificity were calcu-
lated with a stratified bootstrap (10,000 replicates) as implemented in the pROC package
in R (Robin et al., 2011).
2.4 Comparing repeat expansion detection algorithms

The repeat score methods we developed (with a Mann-Whitney test or exSTRa (aggregated T-statistic with p-values derived by simulation)), gave p-values, and we assessed these two methods with similar diagnostic tools.

To assess the p-value distributions of the Mann-Whitney test, we produced histograms of p-values at each locus for each cohort with the ggplot2 package (Wickham, 2016), grouped by cases and controls for the locus. This shows visually the p-value locations for both groups, and at a glance which loci have small p-values for cases, and which expansions would not be detected even without multiple testing correction (p-value > 0.05). With the exSTRa method, we produced histograms of the aggregated T-statistic at each locus for each cohort, and grouped by case-control status.

Histograms of p-values across all loci in each cohort, grouped by case-control status, were produced. The bins were 0.01 in width for p-values less than 0.05, and 0.05 in width otherwise, with the height of bars scaled to the density (each sample contributes the same area to the graph). For each method, we produced quantile-quantile plots (Q-Q plots) of p-values per cohort.

We assessed the performance of the p-values with receiver operating characteristic (ROC) curves and the area under curve (AUC), with the ROCR R package (Sing et al., 2005). 95% confidence intervals of AUC were produced with a stratified bootstrap (10,000 replicates) as implemented in the pROC package in R (Robin et al., 2011).

To classify a sample as expanded at a locus, we chose a significance level of 0.05 and applied Bonferroni correction with the total number of tests in the respective cohort. Sensitivity and specificity were calculated from this classification; confidence intervals (95%) were calculated with a stratified bootstrap (10,000 replicates) as implemented in the pROC package in R (Robin et al., 2011). The classification results were pooled across the main four cohorts (WES_PCR, WGS_PCR_1, WGS_PCR_2, WGS_PF_3), in order to estimate overall sensitivity and specificity.
Chapter 3

Results of detecting repeat expansions in MPS data

This chapter presents the assessment of the performance of repeat expansion detection methods in patients with known disease-causing repeat expansions. The methods and cohorts for this chapter were introduced in Chapter 2. Here, we seeked to determine which repeat expansion detect methods performed the best and if these could be used in screening. Throughout the chapter, the cohorts have been coloured consistently across each figure.

3.1 Quality and alignment summary

FastQC results are summarised in Figure 3.1. FastQC indicated that the base score qualities were generally not as high in the WGS_PCR_1 cohort compared to others. A plot from FastQC of per base score qualities, for one FastQ file with high quality scores, is given in Figure 3.2. In the WGS cohorts, per base sequence content usually failed and per base GC content gave a warning; this is likely due to a low call rate of A bases toward to the end of reads that can be seen for one sample in Figure 3.3. The WGS_PCR_1 cohort had more sequences with uncalled bases (N) than seen in other cohorts. The WGS_PCR_2 and WGS_PF_3 cohorts had many reads hard clipped, hence giving a warning in the sequence length distribution. FastQC indicated the WGS cohorts had some overrepresented sequences; if these are due to duplicates than they can be filtered after alignment.
Chapter 3. Results of detecting repeat expansions in MPS data

The \textit{kmer content}, that is, short subsequences overrepresented, failed in most samples; we have found that this is often observed in MPS data.

The GC content of FASTQ files is given in Figure 3.4, showing a range of 45–48 in the WES\_PCR\_1 cohort and 40–43 in the WGS cohorts. This suggests, in our experiments, that the Agilent SureSelect Human All Exon V5+UTR enrichment was more efficient at a higher GC content than the WGS library preparation protocols.

All FASTQ files come in pairs as all our samples are paired-end sequencing. Some samples will have more FASTQ files either due to higher coverage or subsequent runs to meet required coverage. FastQC uses heuristic methods to determine module results. The FastQC modules are: BS = “basic statistics”, BSQ = “per base sequence quality”, TSQ = “per tile sequence quality”, SQS = “per sequence quality scores”, BSC = “per base sequence content”, SGCC = “per sequence GC content”, BNC = “per base N content”, SLD = “sequence length distribution”, SDL = “sequence duplication levels”, OS = “overrepresented sequences”, AC = “adapter content”, and KMER = “kmer content”.

\textbf{Figure 3.1. FastQC module results}

All FASTQ files come in pairs as all our samples are paired-end sequencing. Some samples will have more FASTQ files either due to higher coverage or subsequent runs to meet required coverage. FastQC uses heuristic methods to determine module results. The FastQC modules are: BS = “basic statistics”, BSQ = “per base sequence quality”, TSQ = “per tile sequence quality”, SQS = “per sequence quality scores”, BSC = “per base sequence content”, SGCC = “per sequence GC content”, BNC = “per base N content”, SLD = “sequence length distribution”, SDL = “sequence duplication levels”, OS = “overrepresented sequences”, AC = “adapter content”, and KMER = “kmer content”.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fastqc_modules.png}
\caption{FastQC module results}
\end{figure}
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.2. FastQC per base sequence quality module for WGSrpt_21 read 1
The mean quality is given by a blue line. The median and interquartile range (IQR) are given by boxes, with whiskers spanning 10–90% of quality scores. Higher quality is better, as indicated by the background colours; quality scores correspond to an error rate of less than 1 in 600 in the green region, and less than 1 in 100 in the orange region.

Figure 3.3. FastQC per base sequence content module for WGSrpt_21 read 1
Note that there were almost no bases called as A for the 151st base.

Figure 3.4. Box plots of GC percent in each FASTQ file
Figure 3.5 shows a summary of the median coverage of each cohort over the WES target regions (for direct comparison), showing that the WGS_PCR_2 and WGS_PF_3 cohorts had the highest coverage, and WES_PCR data had samples with the lowest coverage. Note that the target coverage was heterogeneous within each cohort, but cases generally were targeted to higher depth; WES cases had target 50x coverage, while WGS cases had target depth 60x with the exception of two WGS cases in WGS_PCR_2 (due to a sample mixup, but was corrected at the analysis stage). One sample in the WES_PCR cohort was sampled to very high depth; we have not removed this sample as methods are not using
total coverage to detect outliers.

![Box plots of median coverage over capture regions](image)

**Figure 3.5. Box plots of median coverage over capture regions**

For each sample, we show box plots of coverage of WES over target regions (Figure 3.6(a)) and over the same regions in WGS cohorts (Figures 3.7(a), 3.8(a) and 3.9(a)), and the total reads in each sample for all the cohorts in Figures 3.6(b), 3.7(b), 3.8(b) and 3.9(b). In the WES_PCR, the four cases did not quite reach the 50x coverage requested over capture regions after filtering of reads. Case samples in the WGS were requested to 60x coverage, but this failed for two case samples (WGSrpt_20 and WGSrpt_21) in WGS_PCR_2 due to a sample swap during sequencing; instead they only satisfied the expected coverage for a single unit of sequencing of 30x. This sample swap was corrected in the analysis, but the lower 30x coverage remained.
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.6. WES median and interquartile range (IQR) of coverage, and read count over capture regions
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.7. WGS_PCR_1 median and interquartile range (IQR) of coverage, and read count over WES capture regions for comparison to WES

Figure 3.8. WGS_PCR_2 median and interquartile range (IQR) of coverage, and read count over WES capture regions for comparison to WES

Figure 3.9. WGS_PF_3 (PCR-free) median and interquartile range (IQR) of coverage, and read count over WES capture regions for comparison to WES
In order to make a fair comparison between WES and WGS, the coverage over target regions had to be normalised to the same value. It was preferable to do this by the mean coverage as this gave greater numerical precision (to two decimal places) than the median (integer), and these statistics were highly correlated (Pearson’s $R^2 > 0.998$). The read depth of known repeat expansion loci of WES compared to WGS is shown in Figure 3.10, normalised to have 30x coverage over the target regions for WES and 30x genome coverage for WGS. For most of the coding loci, WES appeared to have similar read depth to WGS with PCR library preparation, with the exception of the SCA3 locus which, had better read depth in WES. WES had lower read depth than WGS for all other loci except for the FRAXE locus where it had similar read depth to the PCR free WGS_PF_3. Read depth was low in the coding loci SCA2, SCA6 and SCA7 in both WES and WGS with PCR, but showed increased read depth with PCR-free library preparation. Overall, PCR-free WGS appeared to give the best coverage of pathogenic STR loci.

![Figure 3.10. Read depth of repeat expansion loci with coverage normalised to 30x](image)

A box plot summary of inferred PCR duplicates by percentage of total reads is shown in Figure 3.11, and by sample in Figure 3.12. Duplicates were lowest in the WES_PCR (median 4.0%), while WGS_PCR_1 had a median duplicate rate of 33%, substantially
higher than the other two WGS cohorts. Although the duplicate rate of WGS_PCR_1 was high, after filtering of duplicates the case samples still exceeded the 60x coverage requested of the sequencing provider.

![Box plots of duplicate percentages by cohort](image)

**Figure 3.11.** Box plots of duplicate percentages by cohort

The median coverage compared to duplication rates is shown in Figure 3.13. This shows that while the WGS_PCR_1 cohort had a high duplicate rate, this did not appear to be due to higher coverage causing duplicate hits by chance.

The median of the sample median insert size of each cohort was 189.5 bp (median absolute deviation as calculated by R’s `mad()` function (MAD) 5.2) for the WES_PCR cohort, and 397, 434.5 and 341 (MAD 3.0, 21, and 32) for the WGS_PCR_1, WGS_PCR_2 and WGS_PF_3 cohorts respectively (Figure 3.14). As the median median insert size of the WES_PCR was smaller than the sum of two 100 bp reads (200 bp), this meant that many mates overlapped at the ends of reads, therefore sequencing the same bases. This could cause double counting of some bases when calling variants when not properly accounted for. This was most likely to affect WES samples sequenced on 5\textsuperscript{th} of June 2014 with median median insert size of 182 bp, 18 bp lower than total sequenced bases per read-pair (Figure 3.15). While the median insert size in WES batches appeared to vary, this was not statistically significant (p-value 0.08, \( F = 1.8 \), \( df = (10, 61) \)).
Figure 3.12. Duplicate percentages by sample
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.13. Median coverage to inferred PCR duplicate rate

Figure 3.14. Insert size by cohort
Figure 3.15. WES insert size by batch

Figure 3.16 summarises the insert size distribution per sample. The PCR-free data appeared to have a greater variability of insert sizes than those from the PCR WGS data, but this may not have been a direct result of the PCR-step, as this variability is largely determined by the size-selection step in sequencing. It was possible that due to the higher DNA input requirement of PCR-free protocols that size-selection had to be performed with a greater bandwidth, allowing greater sequencing depth at the expense of higher insert size variability.
Chapter 3. Results of detecting repeat expansions in MPS data

<table>
<thead>
<tr>
<th>Sample</th>
<th>Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>WGSrpt_07</td>
<td>WES_PCR</td>
</tr>
<tr>
<td>WGSrpt_14</td>
<td>WGS_PCR_1</td>
</tr>
<tr>
<td>WGSrpt_04</td>
<td>WGS_PCR_2</td>
</tr>
<tr>
<td>WGSrpt_10</td>
<td>WGS_PF_3</td>
</tr>
<tr>
<td>WGSrpt_08</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_18</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_11</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_05</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_20</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_12</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_17</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_16</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_02</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_13</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_15</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_10</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_19</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_09</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_17</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_11</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_19</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_21</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_08</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_24</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_07</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_15</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_12</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_18</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_09</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_10</td>
<td></td>
</tr>
<tr>
<td>WGSrpt_06</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3.16. Insert sizes by sample**

Samples are displayed in decreasing order of the median insert size, indicated by a circle. Bars extend to cover 90% of insert sizes, at the 5th and 95th percentile. The interquartile range (IQR), covering 50% of the data, is indicated by small vertical bars. The dotted and dashed vertical lines indicates the threshold at which our WES and WGS samples respectively will usually have overlapping bases, between the two ends of the read.
3.2 Expansion detection in other software

3.2.1 lobSTR results

For most alleletype (genotype) calls, lobSTR gave allele marginal likelihood ratio scores, with those closer to 1 of higher quality. For some some genotypes, these scores were given as NaN (not a number), where we instead used the alleletype likelihood ratio score to infer the allele call qualities. Figure 3.17 shows the lobSTR likelihood ratio scores for the alleles called at the 21 pathogenic STR loci, across the four different sequencing cohorts. Due to the shorter reads of the WES_PCR cohort, the largest alleles called (28.3 repeats) were smaller than for WGS cohorts with longer reads (39 repeats).

![Figure 3.17. All allele sizes called by lobSTR](image)

Allele scores for lobSTR collapsed over loci. Allele category was determined separately for each locus by the repeat number to its population ranges. Each sample at each locus had at most two alleles called, each with a likelihood ratio score that was either: an allele marginal for each allele, or if the allele marginal was not defined, one score for the alleleotype.

We found that lobSTR was unable to call alleleotypes for FRAXE and EPM1. In the lobSTR database, the FRAXE STR was annotated with a 148 bp locus that makes interrogation with lobSTR difficult with 150 bp reads. The FRAXE repeat was also an example of varying definitions of an STR depending on the tolerance for mismatches and indels, with the lobSTR repeat longer than the 46 bp FRAXE repeat we list in Table 1.2. The EMP1
STR was not genotyped by lobSTR as it had a 12 bp repeat motif, longer than the lobSTR maximum of 6 bp.

Examination of the differences between the repeat number and expansion threshold showed that expansion alleles had only been detected in the WGS_PCR_2 cohort (Figure 3.18). We saw that alleles were not as close to the expansion threshold in WES compared to WGS, due to shorter read sizes in the WES data. Many allele calls were close to the expansion thresholds in the WGS data. The distribution of allele calls is given in Figure 3.19 showing that these loci often appeared to have multi-modal distributions as found in (Willems et al., 2014).

Figure 3.18. Allele size difference to minimum expansion size with lobSTR
Allele score differences to expanded for lobSTR collapsed over loci. The red line indicates the pathogenic expansion threshold (lowest unstable). Note some points are not displayed as we have restricted the horizontal axis for clarity. Allele category was determined separately for each locus by the repeat number to its population ranges. Each sample at each locus had at most two alleles called, each with a likelihood ratio score that was either: an allele marginal for each allele, or if the allele marginal was not defined, one score for the alleleotype.

Individual allele calls for each affected sample are shown in Figures 3.20 to 3.23. We saw that the only samples reaching the expansion threshold were in WGS_PCR_2, where WGSrpt_05 and WGSrpt_07 reached the SCA6 threshold, and WGSrpt_18 reached the threshold for SBMA. The SCA6 expansion was identified correctly, but WGSrpt_18 was not affected by SBMA, but instead an expansion at the SCA2 locus. The SBMA genotype
calls showed five case samples with an intermediate allele; these calls were very close to the expansion threshold. This was caused by the lobSTR annotation (and similarly the Tandem Repeat Finder annotation we used) covering more bases with impurities, compared to the pure repeat with whole CAGs normally used to size the SBMA locus that did not include an interruption of the repeat sequence. The following is the reference lobSTR repeat sequence in its entirety (hg19), compared to the stringent CAG repeat of La Spada et al. (1991) used in lower-case and red:

GcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagcagCAAGAGACTAGCCCCAGGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG

To check that intermediate calls were not reflective of repeat expansions, the largest intermediate allele calls were manually inspected in Figure 3.19. The largest intermediate allele calls were not concordant with the disorder these individuals had been previously diagnosed with.
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.19. WGS allele sizes from lobSTR

The vertical blue bar indicates the upper range of normal size alleles, and the red bar indicates the lower range of pathogenic repeat sizes (disease threshold). Some disease ranges are larger than shown, hence do not show vertical blue or red bars.
### Chapter 3. Results of detecting repeat expansions in MPS data

<table>
<thead>
<tr>
<th>HD</th>
<th>SCA2</th>
<th>SCA3</th>
<th>SCA6</th>
<th>SCA7</th>
<th>DRPLA</th>
<th>HDL2</th>
<th>DM1</th>
<th>FTDALS</th>
<th>SCA8</th>
<th>SCA10</th>
<th>SCA12</th>
<th>SCA36</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>rptWEHI1</strong></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><strong>rptWEHI2</strong></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><strong>rptWEHI3</strong></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><strong>rptWEHI4</strong></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

#### Figure 3.20. WES_PCR case sample lobSTR allele calls

The vertical blue bar indicates the upper range of normal size alleles, and the red bar indicates the lower range of pathogenic repeat sizes (disease threshold).

<table>
<thead>
<tr>
<th>HD</th>
<th>SBMA</th>
<th>SCA1</th>
<th>SCA2</th>
<th>SCA3</th>
<th>SCA6</th>
<th>SCA7</th>
<th>DRPLA</th>
<th>HDL2</th>
<th>DM1</th>
<th>FRDA</th>
<th>DM2</th>
<th>FTDALS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HD−1</strong></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><strong>SCA2−1</strong></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><strong>SCA6−1</strong></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

#### Figure 3.21. WGS_PCR_1 case sample lobSTR allele calls

The vertical blue bar indicates the upper range of normal size alleles, and the red bar indicates the lower range of pathogenic repeat sizes (disease threshold).
Figure 3.22. WGS_PCR_2 case sample lobSTR allele calls

The vertical blue bar indicates the upper range of normal size alleles, and the red bar indicates the lower range of pathogenic repeat sizes (disease threshold).
Figure 3.23. WGS _PF _3 PCR-free case sample lobSTR allele calls

The vertical blue bar indicates the upper range of normal size alleles, and the red bar indicates the lower range of pathogenic repeat sizes (disease threshold).
3.2.2 ExpansionHunter results

ExpansionHunter (Dolzhenko et al., 2016) did not give a statistical framework for interpreting the results, instead we relied on the identification of alleles that were longer than the normal range. In Figure 3.24, the raw allele sizes are plotted with the number of reads supporting them; generally longer alleles were supported by fewer reads. In the WES_PCR cohort, many alleles were called as 33 repeat units, but we note that ExpansionHunter was not designed for this kind of data.

![Figure 3.24. All allele sizes called by ExpansionHunter](image)

Allele scores for ExpansionHunter collapsed over loci. Allele category was determined separately for each locus by the repeat number to its population ranges. Each sample at each locus could have more than two alleles. Source denotes the algorithm used by ExpansionHunter to make the allele call.

In Figure 3.25, the difference of the repeat number and expansion threshold has been calculated. This showed that expansion alleles were detected in the WES_PCR, WGS_PCR_1 and WGS_PCR_2 cohorts. We did not identify any expansions in the PCR-free cohort (WGS_PF_3). It could be seen that many alleles reaching pathogenic length were only supported by a few reads. There were still many alleles that are close to the expansion threshold, including intermediate alleles in the WGS_PCR_1 and WGS_PCR_2 cohorts.
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.25. Allele size difference to minimum expansion size with ExpansionHunter

Allele score differences for ExpansionHunter collapsed over loci. The red line indicates the pathogenic expansion threshold (lowest unstable). Note some points are not displayed as we have restricted the horizontal axis for clarity. Allele category was determined separately for each locus by the repeat number to its population ranges. Each sample at each locus may have more than two alleles. Source denotes the algorithm used by ExpansionHunter to make the allele call.

In Figure 3.26, we sought to identify any alleles that were large outliers, as the expansion threshold may not be reached. The loci for SCA6, SCA7, DM1, FRDA and FTDALS1 appeared to show some alleles that were not large enough to be in the expanded range, but were identified as outliers compared to other samples at the respective locus. Additionally, the FRAXA locus appeared to show longer alleles in the WGS_PF_3 cohort, but it was difficult to tell if this would be representative of an expansion due to the small sample size in this cohort.

Individual allele calls for each affected sample are shown in Figures 3.27 to 3.30. ExpansionHunter found alleles within the expanded range for 7 WES_PCR samples (2 of 4 cases, 5 from controls), 3 of 3 cases from WGS_PCR_1, 9 of 16 cases from WGS_PCR_2, and none from WGS_PF_3.

Sensitivity and specificity results are presented, discussed and compared to other methods in Section 3.5, along with the expansions called for each case sample.
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.26. WGS allele sizes from ExpansionHunter

Note that samples are not evenly represented as some samples only had one allele called, while others had as many as nine calls that are all included. The vertical blue bar indicates the upper range of normal size alleles, and the red bar indicates the lower range of pathogenic repeat sizes (disease threshold). In some individual plots the red (or blue) bar is missing because the pathogenic repeat size (or highest normal range) is larger than 50 of which no alleles were called at this size.
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.27. WES_PCR case sample ExpansionHunter allele calls

The vertical blue bar indicates the upper range of normal size alleles, and the red bar indicates the lower range of pathogenic repeat sizes (disease threshold).

Figure 3.28. WGS_PCR_1 case sample ExpansionHunter allele calls

The vertical blue bar indicates the upper range of normal size alleles, and the red bar indicates the lower range of pathogenic repeat sizes (disease threshold).
Figure 3.29. WGS_PCR_2 case sample ExpansionHunter allele calls

The vertical blue bar indicates the upper range of normal size alleles, and the red bar indicates the lower range of pathogenic repeat sizes (disease threshold).
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.30. WGS_PF_3 PCR-free case sample ExpansionHunter allele calls

The vertical blue bar indicates the upper range of normal size alleles, and the red bar indicates the lower range of pathogenic repeat sizes (disease threshold).
Chapter 3. *Results of detecting repeat expansions in MPS data* 76

### 3.3 Assessing the repeat score procedure in exSTRa

This section gives the results of analysis of all of our data with the repeat score procedure that I developed, described in Section 2.1. The repeat score was defined as the number of bases covered by a repeat motif; this is described in detail in Section 2.1.1. We first compared the differences in the repeat score distributions between cohorts in Figures 3.31 and 3.32 with ECDF plots as outlined in Section 2.1.2. Due to the shorter read lengths in the WES_PCR cohort data, we found that our WES data generally had lower repeat scores than the WGS cohorts, and hence we could not use this as control data in the absence of a correction procedure, the development of which is a future research direction. We found 19 of the 21 loci showed similar distributions across the WGS cohorts, with exceptions in two STRs with 100% GC in FRAXA (CGG), FRAXE (CCG). The other 100% GC content FTDALS1 (GGGGCC) appeared to show similar behaviour across the WGS cohorts.

ECDF plots of the repeat scores in the WES_PCR cohort (Figure 3.33) suggested that the HD, SCA2 and SCA6 loci had expanded samples in the blinded cases, and that identification of expansions at these loci with the Agilent SureSelect Human All Exon V5+UTR platform would be feasible. We did not visualise a clear repeat expansion in the rptWEHI4 sample (known to be affected with SCA1), although it had weak evidence of a SCA1 expansion, but similar evidence was also apparent in one of the control samples. We did not visually identify any expansion at other loci in the WES_PCR cohort.

The repeat score at each locus is shown for WGS_PCR_1 in Figure 3.34. We saw that the HD-1 sample appeared to show an expansion at the HD locus, as well as SCA6-1 showing an expansion at the SCA6 locus. SCA2-1 had some reads that showed a repeat expansion at the SCA2 locus, but this was only for a small proportion (approximately 10%) of the reads. The SCA2-1 sample also appeared to have a larger repeat than other samples at the DM1 locus, but the repeat size of DM1 was not independently verified.

The repeat score at each of the loci are shown for WGS_PCR_2 in Figures 3.35 and 3.36. Here we identified expansions visually in 12 of the 16 cases, with the results detailed in Table 3.1. The Friedrich ataxia (FRDA) result appeared most prominent, and may have been due to one or both of the following reasons: the large size of the expanded allele compared to the other loci where the difference between normal and disease threshold alleles was smaller, and/or that FRDA is a recessive disease such that the affected individuals
Chapter 3. Results of detecting repeat expansions in MPS data

would have both alleles as expanded, compared to approximately 1 in 100 unaffected individuals carrying the disease allele. The FRAXA and FRAXE loci were difficult to interpret due to low coverage. Although some samples appeared to have larger alleles, these were only observed in one or two reads per sample. We also split the samples by sex for loci on the X chromosome (Figure 3.37), but due to low coverage of the FRAXA and FRAXE it was difficult to determine the presence of a repeat expansion. The sample WGSrpt_15 appeared to have outlier scores in FRAXA, but this was not reflective of an expansion. The X chromosome disorder SBMA had adequate coverage and there appeared to be no repeat expansions in the male samples as expected.

The repeat score at each locus is shown for WGS_PF_3 in Figure 3.38. As there were only three case samples in this cohort it was difficult to identify expansions from normal samples without comparing to the WGS cohorts with PCR library preparation. The FTDALS1 locus appeared to be larger in the WGSrpt_21_F sample, but the apparent increase was smaller than what would be expected for an expansion. As well as increased coverage at the FRAXA and FRAXE loci in PCR-free cohort WGS_PF_3, the repeat scores were also on average greater, exemplified by a shift to the right of the ECDF. This made use of WGS_PCR_1 and WGS_PCR_2 cohorts as control data for the WGS_PF_3 samples inappropriate. The WGSrpt_17_F sample at the FRAXA locus showed some evidence of a larger allele, but with low confidence due to the lack of appropriate control data.

The complete set of ECDF plots for all 21 expansion loci for all four cohorts is presented in Appendix A.

Overall, these ECDF plots suggested that the repeat score procedure would be appropriate for identifying expansions for some loci. Therefore, this would enable the development of computational methods to identify an expansion when many loci or samples are to be tested.
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.31. repeat score by cohort for select loci part 1

Plot titles indicate the disease symbol, the STR region, the reference length (norm), the minimum expanded pathogenic length (exp). The dashed vertical blue line is the reference repeat length in bp, and the vertical red dashed line is the smallest expansion size (only when within plot range).
Figure 3.32. repeat score by cohort for select loci part 2

Plot titles indicate the disease symbol, the STR region, the reference length (norm), the minimum expanded pathogenic length (exp). The dashed vertical blue line is the reference repeat length in bp, and the vertical red dashed line is the smallest expansion size (only when within plot range).
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.33. WES_PCR ECDF of select loci

Plot titles indicate the disease symbol, the STR region, the reference length (norm), the minimum expanded pathogenic length (exp). The dashed vertical blue line is the reference repeat length in bp, and the vertical red dashed line is the smallest expansion size (only when within plot range). See Appendix A for the other 15 disease loci.
Figure 3.34. WGS_PCR_1 ECDF of select loci

Plot titles indicate the disease symbol, the STR region, the reference length (norm), the minimum expanded pathogenic length (exp). The dashed vertical blue line is the reference repeat length in bp, and the vertical red dashed line is the smallest expansion size (only when within plot range). See Appendix A for all 21 disease loci with all case samples coloured.
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.35. WGS_PCR_2 ECDF of select loci with select samples coloured part 1

Most samples coloured as grey for clarity. Plot titles indicate the disease symbol, the STR region, the reference length (norm), the minimum expanded pathogenic length (exp). The dashed vertical blue line is the reference repeat length in bp, and the vertical red dashed line is the smallest expansion size (only when within plot range). See Appendix A for all 21 disease loci with all case samples coloured.
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.36. WGS_PCR_2 ECDF of select loci with select samples coloured part 2

Most samples coloured as grey for clarity. Plot titles indicate the disease symbol, the STR region, the reference length (norm), the minimum expanded pathogenic length (exp). The dashed vertical blue line is the reference repeat length in bp, and the vertical red dashed line is the smallest expansion size (only when within plot range). See Appendix A for all 21 disease loci with all case samples coloured.
Figure 3.37. WGS_PCR_2 ECDF of X loci for males

Plot titles indicate the disease symbol, the STR region, the reference length (norm), the minimum expanded pathogenic length (exp). The dashed vertical blue line is the reference repeat length in bp, and the vertical red dashed line is the smallest expansion size (only when within plot range). Note the both controls were female and so are excluded here.
Figure 3.38. WGS_PF_3 PCR-free ECDF of select loci

Plot titles indicate the disease symbol, the STR region, the reference length (norm), the minimum expanded pathogenic length (exp). The dashed vertical blue line is the reference repeat length in bp, and the vertical red dashed line is the smallest expansion size (only when within plot range). See Appendix A for the other 15 disease loci.
3.4 Classification of repeat expansions

3.4.1 Visual classification

The classifications of expansions for each sample is given in Table 3.1. In the WES_PCR cohort we were able to correctly call three of the four known repeat expansion samples, while we could not make a confident call for rptWEHI4, although the correct locus was identified with low confidence. For the WGS_PCR_2 cohort, we were able to correctly identify 12 out of the 16 repeat expansion samples, with WGSrpt_20 having low confidence calls in SCA2 (the correct locus) due to only a few reads supporting an expansion. FRAXA was a problematic locus for the WGS_PCR_2 cohort, representing 3 of the 4 loci without a confident call. This was due to low coverage at the FRAXA locus, likely because of GC bias introduced by the PCR step of library preparation.

Table 3.1. Classification of expansions by visual inspection of ECDFs

All visual calls were made by the author when blinded to the underlying repeat expansion, and thus only cases from cohorts WES_PCR and WGS_PCR_2 are included. At most one best locus was called, and further loci that appeared to show more repeat sequence in ECDFs were also noted. Loci where no visual call could be confidently made, but appeared suggestive, were given low confidence calls. Low confidence calls are ordered from what appeared to be the most likely repeat expansion locus, to the least. Correct calls are indicated by a checkmark (✓), while incorrect calls indicated by an x mark (✗).

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Sample</th>
<th>Actual locus</th>
<th>Best locus</th>
<th>Low confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>WES_PCR</td>
<td>rptWEHI1</td>
<td>SCA2</td>
<td>SCA2 ✓</td>
<td>-</td>
</tr>
<tr>
<td>WES_PCR</td>
<td>rptWEHI2</td>
<td>SCA6</td>
<td>SCA6 ✓</td>
<td>-</td>
</tr>
<tr>
<td>WES_PCR</td>
<td>rptWEHI3</td>
<td>HD</td>
<td>HD ✓</td>
<td>-</td>
</tr>
<tr>
<td>WES_PCR</td>
<td>rptWEHI4</td>
<td>SCA1</td>
<td>-</td>
<td>SCA1 ✓</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_05</td>
<td>SCA6</td>
<td>SCA6 ✓</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_07</td>
<td>SCA6</td>
<td>SCA6 ✓</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_08</td>
<td>SCA7</td>
<td>SCA7 ✓</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_09</td>
<td>FRDA</td>
<td>FRDA ✓</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_10</td>
<td>HD</td>
<td>HD ✓</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_11</td>
<td>FRDA</td>
<td>FRDA ✓</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_12</td>
<td>HD</td>
<td>HD ✓</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_13</td>
<td>DM1</td>
<td>DM1 ✓</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_14</td>
<td>SCA1</td>
<td>SCA1 ✓</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_15</td>
<td>DM1</td>
<td>DM1 ✓</td>
<td>FRAXE X</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_16</td>
<td>SCA1</td>
<td>SCA1 ✓</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_17</td>
<td>FRAXA</td>
<td>-</td>
<td>FRAXE X</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_18</td>
<td>SCA2</td>
<td>SCA2 ✓</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_19</td>
<td>FRAXA</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_20</td>
<td>SCA2</td>
<td>-</td>
<td>SCA2 ✓, FRAXA X</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>WGSrpt_21</td>
<td>FRAXA</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
3.4.2 Classification with Mann-Whitney test

The distribution of the Mann-Whitney test p-values by locus for each cohort is shown in Figures 3.39 to 3.42. These suggested that the p-values were not well calibrated, with many p-values of 1. This was further confirmed by quantile-quantile (Q-Q) plots (Figure 3.45), showing many p-values were smaller than expected.

The Mann-Whitney test was able to detect the expansion in most loci but was not as effective for identifying the expansions in SCA2 and FRAXA (Tables 3.3 and 3.4). There were also many false-positive calls, with specificities of 0.85 to 0.96. The distribution of p-values of the Mann-Whitney test for each cohort (collapsed over all loci) is shown in Figure 3.43. P-values distributions were heavy tailed in all cohorts.

ROC curves (Figure 3.44) showed good performance with AUC in most cohorts, except the WGS_PF_3 cohort which only had FRAXA cases. The WES_PCR cohort had the highest AUC value of 0.997 (95% CI = (0.989,1)), but this was likely due to the selection of loci rather than WES performing better than WGS in general. Similarly, the WGS_PCR_1 cohort only contained coding repeat expansions, explaining the high AUC of 0.990 (95% CI = (0.971,1)). The WGS_PCR_2 cohort included seven affected individuals with a non-coding repeat expansion, contributing to a lower AUC of 0.857 (95% CI = (0.737,0.977)). The FRAXA locus, with high GC-content, had poor coverage in the WGS_PCR_2 cohort, thus making FRAXA expansion detection difficult. Reducing coverage to 30x for WGS_PCR_2 cohort samples had negligible affect on the AUC [WGS_PCR_2_30x_1 0.863 (95% CI = (0.750,0.975)), WGS_PCR_2_30x_2 0.836 95% CI = (0.703,0.969)], but the p-values were not as small as for the full WGS_PCR_2 cohort data.

The WGS_PF_3 cohort consisted of only FRAXA affected and pre-mutation cases; despite using a PCR-free library preparation protocol, the AUC was only 0.772 (95% CI = (0.566,0.977)). The ECDF of FRAXA in WGS_PF_3 (Figure 3.38) showed that repeat scores were not much larger in the expanded samples, and thus consistent with a lower AUC.
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.39. Mann-Whitney p-values in WES_PCR by locus
Figure 3.40. Mann-Whitney p-values in WGS_PCR_1 by locus
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.41. Mann-Whitney p-values in WGS_PCR_2 by locus
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.42. Mann-Whitney p-values in WGS_PF_3 by locus
Figure 3.43. Mann-Whitney p-values per cohort

Histograms of the frequency density for Mann-Whitney test p-values for all STR loci. The bins on the far left, where p-value < 0.05, are plotted at smaller bin sizes of 0.01 to show greater detail, whilst other bins were plotted with bin size 0.05. Also included are the two 30x coverage splits of cohort WGS_PCR_2.
Chapter 3. Results of detecting repeat expansions in MPS data

Receiver operating characteristic (ROC) curves were summarised over each cohort. The predictor variable was the $-\log_{10}$ transform of the p-value of the one-sided Mann-Whitney test. The colour of the ROC curve indicates the p-value threshold at that point after a $-\log_{10}$ transform. The p-value threshold used after Bonferroni correction is indicated by an ‘x’. The area under curve (AUC) is also given for each cohort. Also included are the two 30x coverage splits of cohort WGS_PCR_2.
Figure 3.45. Q-Q plots of $-\log_{10}$ transform p-values from Mann-Whitney per cohort

The solid blue line is the desired fit, samples with known expansions (for the respective locus being tested) are red x’s, and the red dashed line is the Bonferroni threshold, where observations above this are called as an expansion. The threshold varies depending on the number of samples and loci being tested in each cohort. Also included are the two 30x coverage splits of cohort WGS_PCR_2.
3.4.3 Classification with p-values of the aggregated t-statistic

The distribution of the aggregated t-statistic $T$ by locus for each cohort is shown in Figures 3.46 to 3.49. For most loci, the cases had the highest $T$ statistic, with the exception of the FRAXA locus. While the distribution of $T$ appeared to vary per locus, we did not identify a pattern. The SCA7 affected individual (WGSrpt_08) had the highest $T$-value that appeared to be caused by low variability of the control samples at this locus. The size of the $T$ statistic for SCA2 in the sample rptWEHI2/SCA2-1/WGSrpt_05 was somewhat variable between the cohorts, being much larger in WES_PCR and WGS_PCR_1 while only a little larger in WGS_PCR_2.

The distribution of p-values, derived through simulation, for each cohort (collapsed over all loci) is shown in Figure 3.50. Most p-values of cases were $\leq 0.01$.

While the p-values still had heavy tails, these were less heavy than the p-values from the Mann-Whitney test statistic. ROC curves in Figure 3.51 showed good performance with AUC in most cohorts, except cohort WGS_PF_3 that only had FRAXA cases. Notably, the WES_PCR cohort achieved a very high AUC of 0.997 (95% CI 0.989–1.000), but this was likely due to all the cases being caused by coding CAG expansions. Q-Q plots in Figure 3.52 shows our simulation was deriving small p-values as small as the simulation allowed ($10^{-6}$) for most samples, but it did appear that for controls p-values were smaller than expected.

Sensitivity and specificity results are given, discussed and compared to other methods in Section 3.5. We also compare the expansions called for case samples in that section.
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.46. Aggregated t-statistic in WES_PCR by locus
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.47. Aggregated t-statistic in WGS_PCR_1 by locus
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.48. Aggregated t-statistic in WGS_PCR_2 by locus
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.49. Aggregated t-statistic in WGS_PF_3 by locus
Chapter 3. Results of detecting repeat expansions in MPS data

Figure 3.50. Aggregate t-statistic p-values per cohort

Histograms of the frequency density for the empirically derived p-values for all STR loci. The bins on the far left, where p-value < 0.05, are plotted at smaller bin sizes of 0.01 to show greater detail, whilst other bins were plotted with bin size 0.05. Also included are the two 30x coverage splits of cohort WGS_PCR_2.
Chapter 3. Results of detecting repeat expansions in MPS data

Receiver operating characteristic (ROC) curves are summarised over each cohort. The predictor variable was the $-\log_{10}$ transform of the p-value derived from the simulation of the aggregated t-statistic. The colour of the ROC curve indicates the p-value threshold at that point after a $-\log_{10}$ transform. The p-value threshold used after Bonferroni correction is indicated by an ‘x’. The area under curve (AUC) is also given for each cohort. Also included are the two 30x coverage splits of cohort WGS_PCR_2.
Figure 3.52. Q-Q plots of $-\log_{10}$ transform p-values from T per cohort

The solid blue line is the desired fit, samples with known expansions (for the respective locus being tested) are red x’s, and the red dashed line is the Bonferroni threshold, where observations above this are called as an expansion. The threshold varies depending on the number of samples and loci being tested in each cohort. Observations at 6 ($p-value = 10^{-6}$) means that the T statistic was larger than all the 999,999 simulations for that locus. Also included are the two 30x coverage splits of cohort WGS_PCR_2.
3.5 Results summary

As lobSTR was not designed to be able to detect repeat expansions, and its results showed it was limited in detecting expansions, we have not attempted to generate comprehensive performance metrics for lobSTR. A comparison of ExpansionHunter (with intermediate and expanded thresholds) to the repeat score procedure with Mann-Whitney and aggregated t-statistic is summarised in Table 3.3, with calls for all case samples in Table 3.4. Using the repeat score procedure, the T statistic p-values gave higher specificity than the Mann-Whitney test, with similar sensitivity. When using ExpansionHunter, an expanded allele size threshold missed 4 expansions in WGS_PCR_2 that would otherwise be detected with a more lenient intermediate allele size threshold; as we prioritised minimising false-negatives, we made further comparisons to exSTRa with an intermediate size threshold in ExpansionHunter. ExpansionHunter (intermediate threshold) performed similarly to exSTRa, where the same expansions were correctly identified except for 1 true-positive only identified by exSTRa and 2 true-positives only identified by ExpansionHunter (Table 3.2). Reducing the WGS_PCR_2 cohort to 30x coverage did not show an impact on sensitivity (due to small sample sizes) and may have increased specificity due to a reduction in power. The FRAXA locus was difficult to call in all cohorts, with all cases being false-negative except for one sample when using the Mann-Whitney test; however, since the Mann-Whitney test had a large number of false-positives, this may have been due to chance.
Chapter 3. \textit{Results of detecting repeat expansions in MPS data} 

### Table 3.2. Comparing known expansions detected

Comparing the detection of known cases only. Only the four main cohorts are included (excluded WGS\_PCR\_2\_30X\_1 and WGS\_PCR\_2\_30X\_2). See Table 3.3 for specificity. Only ExpansionHunter with an intermediate repeat size threshold is shown, as this leads to more true positives. TP = true-positives, FN = false-negatives.

<table>
<thead>
<tr>
<th>ExpansionHunter (intermediate)</th>
<th>exSTRa</th>
<th>MNW</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>FN</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>TP</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>FN</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>TP</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>FN</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

TP = true-positives, FN = false-negatives.
### Table 3.3. Performance summary of algorithmic expansion calls

For exSTRA, the performance of the p-values derived by simulation is given. ExpansionHunter did not give an appropriate metric to derive ROC curves and hence area under curve (AUC) could not be computed. For ExpansionHunter, two thresholds were used at each locus: either those where the largest allele was in the expanded (exp) range, and those in or above the intermediate (int) range. In the WES_PCR cohort, only loci with local baits on the Agilent SureSelect Human All Exon V5+UTR are tested. Calls were classified as significant at the 0.05 level after Bonferroni correction for the number of tests in the cohort. Also included are the two 30x coverage splits of cohort WGS_PCR_2. *The pooled results include the main four cohorts, but not WGS_PCR_2_30X_1 and WGS_PCR_2_30X_2. TP = true-positives, FN = false-negatives.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Cases</th>
<th>Controls</th>
<th>Method</th>
<th>AUC</th>
<th>TP</th>
<th>FN</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>WES_PCR</td>
<td>4</td>
<td>58</td>
<td>exSTRA</td>
<td>0.997 (0.989–1.000)</td>
<td>3</td>
<td>1</td>
<td>0.75 (0.25–1.00)</td>
<td>0.994 (0.988–0.998)</td>
<td>682</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Score Mann-Whitney</td>
<td>0.965 (0.912–1.000)</td>
<td>3</td>
<td>1</td>
<td>0.75 (0.25–1.00)</td>
<td>0.97 (0.939–0.970)</td>
<td>682</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (exp)</td>
<td>0.999 (0.973–1.000)</td>
<td>2</td>
<td>2</td>
<td>0.50 (0.00–1.00)</td>
<td>0.994 (0.988–0.999)</td>
<td>682</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (int)</td>
<td>0.912 (0.755–1.000)</td>
<td>2</td>
<td>2</td>
<td>0.50 (0.00–1.00)</td>
<td>0.945 (0.926–0.962)</td>
<td>682</td>
</tr>
<tr>
<td>WGS_PCR_1</td>
<td>3</td>
<td>14</td>
<td>exSTRA</td>
<td>0.890 (0.751–1.000)</td>
<td>2</td>
<td>2</td>
<td>0.67 (0.00–1.00)</td>
<td>0.989 (0.977–0.997)</td>
<td>357</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Score Mann-Whitney</td>
<td>0.890 (0.751–1.000)</td>
<td>2</td>
<td>2</td>
<td>0.67 (0.00–1.00)</td>
<td>0.989 (0.977–0.997)</td>
<td>357</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (exp)</td>
<td>0.912 (0.755–1.000)</td>
<td>2</td>
<td>2</td>
<td>0.67 (0.00–1.00)</td>
<td>0.915 (0.881–0.942)</td>
<td>357</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (int)</td>
<td>0.750 (0.50–1.00)</td>
<td>3</td>
<td>0</td>
<td>1.00 (1.00–1.00)</td>
<td>1.000 (1.000–1.000)</td>
<td>357</td>
</tr>
<tr>
<td>WGS_PCR_2</td>
<td>16</td>
<td>2</td>
<td>exSTRA</td>
<td>0.892 (0.748–1.000)</td>
<td>12</td>
<td>4</td>
<td>0.75 (0.00–1.00)</td>
<td>0.976 (0.954–0.992)</td>
<td>378</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Score Mann-Whitney</td>
<td>0.857 (0.737–0.977)</td>
<td>11</td>
<td>5</td>
<td>0.69 (0.47–0.93)</td>
<td>0.890 (0.856–0.920)</td>
<td>378</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (exp)</td>
<td>0.56 (0.31–0.81)</td>
<td>9</td>
<td>7</td>
<td>0.56 (0.31–0.81)</td>
<td>1.000 (1.000–1.000)</td>
<td>378</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (int)</td>
<td>0.56 (0.31–0.81)</td>
<td>13</td>
<td>3</td>
<td>0.81 (0.63–1.00)</td>
<td>0.994 (0.986–1.000)</td>
<td>378</td>
</tr>
<tr>
<td>WGS_PCR_2_30X_1*</td>
<td>16</td>
<td>2</td>
<td>exSTRA</td>
<td>0.899 (0.764–1.000)</td>
<td>13</td>
<td>3</td>
<td>0.81 (0.67–1.00)</td>
<td>0.989 (0.978–0.997)</td>
<td>378</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Score Mann-Whitney</td>
<td>0.862 (0.750–0.975)</td>
<td>9</td>
<td>7</td>
<td>0.56 (0.33–0.87)</td>
<td>0.931 (0.902–0.955)</td>
<td>378</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (exp)</td>
<td>0.56 (0.31–0.81)</td>
<td>9</td>
<td>7</td>
<td>0.56 (0.31–0.81)</td>
<td>1.000 (1.000–1.000)</td>
<td>378</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (int)</td>
<td>0.56 (0.31–0.81)</td>
<td>12</td>
<td>4</td>
<td>0.75 (0.56–0.94)</td>
<td>0.994 (0.986–1.000)</td>
<td>378</td>
</tr>
<tr>
<td>WGS_PCR_2_30X_2*</td>
<td>16</td>
<td>2</td>
<td>exSTRA</td>
<td>0.881 (0.727–1.000)</td>
<td>12</td>
<td>4</td>
<td>0.75 (0.60–1.00)</td>
<td>0.994 (0.986–1.000)</td>
<td>378</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Score Mann-Whitney</td>
<td>0.836 (0.703–0.969)</td>
<td>10</td>
<td>6</td>
<td>0.63 (0.40–0.87)</td>
<td>0.931 (0.903–0.956)</td>
<td>378</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (exp)</td>
<td>0.56 (0.31–0.81)</td>
<td>9</td>
<td>7</td>
<td>0.56 (0.31–0.81)</td>
<td>1.000 (1.000–1.000)</td>
<td>378</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (int)</td>
<td>0.56 (0.31–0.81)</td>
<td>12</td>
<td>4</td>
<td>0.75 (0.50–0.94)</td>
<td>0.994 (0.986–1.000)</td>
<td>378</td>
</tr>
<tr>
<td>WGS_PF_3</td>
<td>3</td>
<td>5</td>
<td>exSTRA</td>
<td>0.762 (0.552–0.971)</td>
<td>0</td>
<td>3</td>
<td>0.00 (0.00–0.00)</td>
<td>0.976 (0.952–0.994)</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Score Mann-Whitney</td>
<td>0.772 (0.566–0.977)</td>
<td>1</td>
<td>2</td>
<td>0.33 (0.00–1.00)</td>
<td>0.948 (0.794–0.903)</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (exp)</td>
<td>0.00 (0.00–0.00)</td>
<td>0</td>
<td>3</td>
<td>0.00 (0.00–0.00)</td>
<td>1.000 (1.000–1.000)</td>
<td>168</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (int)</td>
<td>0.00 (0.00–0.00)</td>
<td>0</td>
<td>3</td>
<td>0.00 (0.00–0.00)</td>
<td>0.970 (0.939–0.994)</td>
<td>168</td>
</tr>
<tr>
<td>Pooled</td>
<td>26</td>
<td>79</td>
<td>exSTRA</td>
<td>0.68 (0.46–0.81)</td>
<td>17</td>
<td>9</td>
<td>0.65 (0.46–0.81)</td>
<td>0.987 (0.962–0.992)</td>
<td>1585</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Score Mann-Whitney</td>
<td>0.966 (0.775–1.000)</td>
<td>16</td>
<td>8</td>
<td>0.65 (0.46–0.81)</td>
<td>0.920 (0.906–0.933)</td>
<td>1585</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (exp)</td>
<td>0.54 (0.35–0.73)</td>
<td>14</td>
<td>12</td>
<td>0.54 (0.35–0.73)</td>
<td>0.997 (0.994–0.999)</td>
<td>1585</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExpansionHunter (int)</td>
<td>0.69 (0.50–0.85)</td>
<td>18</td>
<td>8</td>
<td>0.69 (0.50–0.85)</td>
<td>0.970 (0.961–0.978)</td>
<td>1585</td>
</tr>
</tbody>
</table>
### Table 3.4. Summary of algorithmic expansion calls on case samples

Correct calls are indicated by a checkmark (✓), while incorrect calls are indicated by an × mark (X). When the correct call was made together with incorrect calls, this is indicated by a star (*). Calls were classified as significant at the 0.05 level after Bonferroni correction for the number of tests in the cohort. Also included are the two 30x coverage splits of cohort WGS_PCR_2.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Sample</th>
<th>Actual</th>
<th>Expanded</th>
<th>Match</th>
<th>Intermediate</th>
<th>Significant</th>
<th>Match</th>
<th>p&lt;0.01</th>
<th>Significant</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>WES_PCR</td>
<td>rpWES14</td>
<td>SCA2</td>
<td>SCA2</td>
<td>✓</td>
<td>SCA8</td>
<td>✓</td>
<td>SCA2, SCA12</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WES_PCR</td>
<td>rpWES12</td>
<td>SCA6</td>
<td>SCA6</td>
<td>✓</td>
<td>SCA8</td>
<td>✓</td>
<td>SCA8, SCA8</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WES_PCR</td>
<td>rpWES13</td>
<td>HD</td>
<td>✓</td>
<td>HD</td>
<td>✓</td>
<td>HD</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WES_PCR</td>
<td>rpWES14</td>
<td>SCA1</td>
<td>✓</td>
<td>SCA1</td>
<td>✓</td>
<td>SCA1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>SCA1</td>
</tr>
</tbody>
</table>

**Chapter 3. Results of detecting repeat expansions in MPS data**

106
3.6 Chapter summary

In this chapter we analysed four cohorts (Tables 2.1 and 2.2) with three different library preparation protocols: WES with Agilent SureSelect Human All Exon V5+UTR capture, WGS with Illumina TruSeq Nano library preparation, and WGS with KAPA Hyper Prep Kit with a PCR-free workflow. We presented a method, exSTRa, that could detect repeat expansions in seven disease loci, of eight tested, in both WES and WGS.

Our method exSTRa was compared to ExpansionHunter (with an intermediate allele size threshold). In all cohorts, it was difficult to determine which algorithms had higher sensitivity due to the small number of case samples. The pooled results, along with the WES_PCR cohort, suggested exSTRa had better specificity than ExpansionHunter, with similar sensitivity. In each WGS cohort, the algorithms appeared to perform similarly. Detecting repeat expansions in regions of high GC content (such as FRAXA, FRAXE) required a PCR-free library preparation protocol, as GC bias caused low read depth over these loci. Despite this, the algorithmic methods tested still struggled to detect an expansion at the FRAXA locus; in exSTRa, this may be due to the small cohort size with just five controls.

This was the first analysis of repeat expansion detection with WES, along with WGS that involved PCR library preparation. We found that both our method, exSTRa, and the existing method, ExpansionHunter, were able to detect repeat expansions in both of these experiment set ups. A larger number of cases in all cohorts is required to determine which tool performs better, and how this depends on the loci being tested.
Chapter 4

Massively parallel sequencing and clinical data of a family with spinocerebellar ataxia mapping to the SCA25 locus

4.1 Introduction

In this self-contained chapter, we describe a four-generation Australian family affected with spinocerebellar ataxia (SCA) with evidence of anticipation. Four genomic regions were identified by linkage mapping, one of which is contained inside the previously defined SCA25 locus that has no known underlying gene. SCA25 (Online Mendelian Inheritance in Man (OMIM) %608703) was first mapped in 2004 to 2p21-p15 in a single French kindred with seven affected individuals by Stevanin et al. (2004) (cytogenetic location is listed in OMIM as 2p21-p13). The locus was mapped to the region defined by microsatellite markers D2S2174 and D2S2736, (hg19 chr2:45,178,644-59,987,608), with the maximum two-point LOD score of 3.15 at D2S2378. Three of the seven affected individuals were male. The age of onset of affected individuals varied from seventeen months to thirty-nine years; Stevanin et al. did not find strong evidence of anticipation despite its variability. However, evidence of anticipation in disease severity was found, with two affected individuals who required
wheelchairs being from the latest generation and an unaffected obligate carrier aged sixty-one who either displayed incomplete penetrance or very late onset. The other notable clinical feature in the original SCA25 family is sensory neuropathy. No other linkage to SCA25 has been reported in the literature and to date no causative gene has been found.

This chapter describes the methods and results of linkage analysis with SNP chip data of the Australian family, and analysis of exome and genome sequencing that included investigation of single-nucleotide variants (SNV), indels and well as copy-number variants (CNVs). Additionally, investigation of short tandem repeat (STR) expansions is described in detail.

4.2 Methods

4.2.1 Recruitment and diagnosis

This study was approved by the Royal Children’s Hospital Human Research Ethics Committee (Approval number 28097). Informed consent was obtained from all participants or their parent (for minors included in the study). Genomic DNA was extracted from whole blood using the BACC DNA extraction kit (GE Healthcare Life Sciences, Uppsala, Sweden) according to the manufacturer’s protocols. Detailed clinical neurological assessments, brain Magnetic Resonance Imaging (MRI) scanning and nerve conduction studies were performed.

4.2.2 SNP chip linkage analysis

Genetic linkage is the property of DNA sequences that are close together on a chromosome, that are more likely to be inherited together than DNA sequences on different chromosomes (Ott, 1999, p.15). The closer two loci are on a chromosome, the more linked these loci will be. Genetic linkage, or genetic distance is measured in centimorgan (cM), with 1 cM defined as the distance where the expected number of recombination events is 0.01 (Botstein et al., 1980). In humans, 1 cM corresponds to 1 million base pairs on average, but this varies throughout the genome and therefore genetic maps (such as deCODE (Kong et al., 2002) and HapMap (Frazer et al., 2007;
International HapMap 3 Consortium et al., 2010)) are required for accurate conversion from genetic to physical distances and vice versa.

Parametric linkage analysis allows mapping a disease locus given the correct disease model (Abecasis et al., 2002). From the 1990s, linkage analyses were performed with STR markers (usually called ‘microsatellite markers’ in this context), but while these are highly polymorphic, only a comparatively small number of markers can be used in the analysis due to the cost and speed (Evans and Cardon, 2004). For example, the ABI PRISM™ Linkage Mapping Set Version 2 only allowed genotyping of approximately 400 markers (PE Applied Biosystems, 1990s). Today, single nucleotide polymorphisms (SNP) are the preferred marker for linkage analysis due to their abundance, distribution throughout the genome and our ability to genotype many of these markers using massively parallel technologies such as SNP chips (millions of markers) or massively parallel sequencing (MPS) (Smith et al., 2011). In linkage analysis, at each locus being tested, the logarithm of the odds, known as the LOD score, is calculated. The LOD is the logarithm of the ratio of the likelihood the locus is linked to the disease locus (recombination fraction < 0.5), over the likelihood the locus is unlinked to the disease locus (recombination fraction = 0.5) (Morton, 1955).

Genome-wide significance is reached at $LOD = 3$, that is, a 1000 to 1 odds that the locus is not observed by chance. After multiple testing correction, this corresponds to approximately 20 to 1 odds in favour of linkage, or a significance level of approximately 0.05. For single pedigrees, reaching genome-wide significance is usually not possible, but linkage peaks can be useful in filtering the genome to only those regions which fit the chosen disease model (Nyholt, 2000). Linkage analysis can be performed on multiple pedigrees to obtain genome-wide significance. Inspection of inferred haplotypes (such as with Haplopainter (Thiele and Nürnberg, 2005)) is useful to confirm that the identified regions fit the assumed genetic and which individuals carry the susceptibility haplotype.

All available DNA samples ('S' in Figure 4.1) were hybridised to the Illumina HumanCytoSNP-12 v2.1 SNP chip. Eight of these were performed at the Australian Genome Research Facility (AGRF) Melbourne, Australia. The IV-1 sample was collected at a later date, with hybridisation performed at the Victorian Clinical Genetics Services (VCGS), Parkville, Australia (Table 4.1). SNP genotypes were called in Illumina BeadStudio.
Chapter 4. MPS of family with spinocerebellar ataxia mapping to the SCA25 locus

Figure 4.1. Pedigree of family affected with unknown spinocerebellar ataxia

Pedigree with those affected by spinocerebellar ataxia in black, with age of onset of ataxia when available to the lower right of the individual in red. S indicates that SNP chip data with the Illumina HumanCytoSNP-12 v2.1 obtained for that sample and N that next-generation sequencing (WES and WGS) was obtained for that sample.

The best fit HapMap population was chosen with the bestPopTest option of LINKDATAGEN (Bahlo and Bromhead, 2009), then allele frequencies and genetic map positions were generated for the chosen CEU HapMap2 population, removing Mendelian errors. Remaining markers were used to verify relatedness with XIBD (Henden et al., 2016). Markers were further filtered by LINKDATAGEN to keep only the most heterozygous markers within 0.3 cM bins (minimum distance between markers 0.15 cM) to ensure linkage equilibrium assumptions were satisfied (Smith et al., 2011). Inbreeding coefficients were estimated with FEStim (Leutenegger et al., 2003).

Autosomal multipoint linkage analysis was performed with MERLIN (Abecasis et al., 2002) under a 90% penetrant autosomal dominant model allowing for a 5% phenocopy rate with a rare susceptibility allele frequency (0.001%). As SCAs may be a late onset disorder (Paulson, 2009), the three unaffected samples with data, including the obligate carrier II-1, were modelled with unknown affectedness status. The X-chromosome was not considered in the analysis since three of the six affected individuals were female and there was male-to-male genetic inheritance.

LOD score peaks underwent haplotype inspection with Haplopainter (Thiele and Nürnberg, 2005). To refine the linkage regions and breakpoints, analysis of positive regions (LOD>1) was rerun with all markers included. For purposes of filtering variants and STRs, we
Table 4.1. Genomic data generated for ataxia family

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>SNP Chip</th>
<th>WES</th>
<th>WGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-1</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>II-3</td>
<td>Yes</td>
<td>Yes</td>
<td>HiSeq X Ten</td>
</tr>
<tr>
<td>III-1</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>III-2</td>
<td>Yes</td>
<td>Yes</td>
<td>HiSeq X Ten</td>
</tr>
<tr>
<td>III-5</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>III-6</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>III-7</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IV-1</td>
<td>Yes*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IV-2</td>
<td>Yes</td>
<td>Yes</td>
<td>HiSeq 2500</td>
</tr>
<tr>
<td>C-1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

extended the regions by 1 Mb either side to ensure errors in genotyping did not cause erroneous filtering of the underlying mutation.

Linkage regions were assessed for overlap with the thirty-nine known dominant SCA loci (as annotated in OMIM) with the R Bioconductor package GenomicsRanges (Lawrence et al., 2013). We also assessed overlap with loci of autosomal recessive spinocerebellar ataxias, SCAR1–26, as it is possible that some mutations may behave dominantly in these genes. Loci that act this way has previously been identified in the gene encoding spectrin beta chain, non-erythrocytic 2 (SPTBN2), causing both SCAR14 (Lise et al., 2012) and SCA5 (Ikeda et al., 2006).

4.2.3 Exome/genome sequencing analysis

Exome sequencing (Agilent SureSelect XT Human All Exon v4 (70x) as well as V5+UTR (50x), sequenced on Illumina HiSeq 2500) of individuals II-3, III-2 (A1-2) and IV-2 (A1-11) was performed at the AGRF Melbourne, Australia (Chapters 2 and 3 names in parentheses).
WGS was performed on individual IV-2 at AGRF Melbourne an Illumina HiSeq 2500 in rapid-run mode (48x (4 lanes), 150 bp paired-end). WGS was also performed for II-3 and III-2 on the Illumina HiSeq X Ten (60x (2 lanes each), 151 bp paired-end), as well as 15 samples from other projects (cohort WGS_PCR_1, Chapters 2 and 3) which were used as controls, Kinghorn Centre for Clinical Genomics (KCCG), Darlinghurst, Australia. The control data were generated in a batch at the same time.

All three WES samples were analysed jointly and all three WGS samples were analysed jointly, but the WES and WGS data were analysed separately. Data were analysed using GATK best practice guidelines (DePristo et al., 2011) up to variant calling. FastQ data were aligned with Bowtie2 (Langmead and Salzberg, 2012) (version 2.2.5) with read groups set at the lane level. Reads were sorted and merged by Novosort v1.03.07 (http://www.novocraft.com/products/novosort/) (including merging the data from the two WES capture platforms, but not WES and WGS). Duplicate marking was performed with Picard v1.117 (Broad Institute of MIT and Harvard, 2016). Local realignment followed by quality score recalibration was performed with GATK v3.2-2 (McKenna et al., 2010). Depth of coverage was assessed with GATK DepthOfCoverage.

Variant calling was performed with the HaplotypeCaller in GATK (DePristo et al., 2011). We verified that the SNP chip, WGS and WES genotypes at HapMap SNPs were identical-by-state to validate sample identity. Variants were annotated to the UCSC Known Genes with ANNOVAR (Wang et al., 2010). Deleterious effects of single nucleotide variants were predicted with SIFT (Kumar et al., 2009) and Polyphen2 (Adzhubei et al., 2010). The Residual Variation Intolerance Score (RVIS) (Petrovski et al., 2015) at 0.1% frequency was annotated for the gene of each variant.

Variants were filtered out if they did not satisfy all the following criteria: within 1 Mb of the linkage regions, alternative allele frequency equal to or less than 0.01 in the 1000 Genomes Project (ANNOVAR 1000g2012feb) (1000 Genomes Project Consortium et al., 2012, 2010), NHLBI 6500 exomes and ExAC (release 0.3) databases, less than 10 observed alleles in 132 in-house control samples, in an open reading frame (ORF) or within 5 bp of a splice site, and not a synonymous change. As the three samples sequenced were all affected individuals we expected heterozygous calls under the dominant disease model. However, to allow for genotyping errors and missing data, we retained all variants where the samples had at least one heterozygous call and not more than one homozygous call. Variants were
Chapter 4. *MPS of family with spinocerebellar ataxia mapping to the SCA25 locus* 114

visually inspected in the Integrative Genomics Viewer (IGV) (Thorvalsdóttir et al., 2013) to validate the genotype.

### 4.2.4 Copy-number variations

Copy-number variations (CNVs) were detected with the SNP chip data using PennCNV (Wang et al., 2007) independently for each sample. PennCNV was used as it has been shown to perform well on SNP chip data (Zhang et al., 2014). CNVs were also called using the WGS HiSeq X Ten data of individuals II-3 and III-2 and 15 in-house controls. CNVs were identified by read coverage depth in 1 kb bins and normalized with a joint LOESS mappability and GC content correction method in the Bioconductor package QDNASeq (Scheinin et al., 2014). The binned data were then merged by DNACopy (Venkatraman and Olshen, 2007) with circular-binary segmentation, with which CGHcall (van de Wiel et al., 2007) made CNV calls. Filtering retained CNVs in DNA from affected individuals if they did not overlap control sample CNVs and were within linkage regions. These CNVs then underwent manual inspection in IGV and compared to known variants in the Database of Genomic Variants (DGV) (MacDonald et al., 2014) in the UCSC Genome Browser (Kent et al., 2002).

### 4.2.5 Validation by Sanger sequencing

Four variants discovered by WES and WGS underwent validation by standard Sanger sequencing in DNA derived from six affected and three unaffected family members (III-6, III-7 and obligate carrier II-1). The genotype of the variant in the gene encoding DNA mismatch repair protein Msh6 (*MSH6*) for the affected individual II-3 was inconsistent with results showing all affected individuals shared the same haplotype. Hence, this variant in II-3 was Sanger sequenced in the sample used for WES and WGS, as well as in an independently collected DNA sample.

### 4.2.6 Analysis of repeats for putative repeat expansions

Twenty-one known repeat expansion disorder loci, as given in Table 1.2, had allele sizes determined with the repeat expansion detection software ExpansionHunter (Dolzhenko et al., 2016) (version 2.0.9) for the two HiSeq X WGS samples and 15 control
samples, noting that this software is designed for WGS with PCR-free library preparation, whereas the data presented here involved a PCR step in library preparation. The loci specification files were obtained by personal communication with Egor Dolzhenko, the lead author of ExpansionHunter. The called allele sizes were classified as either normal, intermediate or expanded, based on known disease repeat sizes.

For loci not yet implicated in disorders, STR expansion detection analysis is laborious, as it requires visual inspection of each STR, and was only attempted for STRs in the linkage regions. STRs with repeat unit sizes of 2–12 bp were identified from the UCSC Genome Browser Simple Repeats track (Benson, 1999). We visually inspected repeats within coding regions (UCSC Known Genes Hsu et al. (2006)) with IGV (with the “Show soft-clipped bases” preference switched on) to look for signs of expansions such as large insertions and soft-clipping of repetitive bases.

The repeat genotyping software lobSTR (Gymrek et al., 2012) was used to call genotypes on genomic STRs with repeat unit sizes of 2–6 bp on WGS HiSeq X data of II-3 and III-2 and 15 control samples. We restricted analysis of STR genotypes to linkage regions. Outlier alleles were identified with the boxplot() function in R (R Core Team, 2016). Filtering required both of the sequenced affected individuals to have the largest outlier alleles with one at least 10 bp larger than the median. Remaining STRs were annotated against the STR Catalogue Viewer (Willems et al., 2014), expression STR (eSTR) significance (Gymrek et al., 2016), and closest gene with its GTEx expression (The GTEx Consortium et al., 2015). We further removed STRs where the largest affected allele was not greater than observed in the STR Catalogue Viewer.

4.3 Results

4.3.1 Clinical details

The proband was patient III-1 who had an onset of ataxia at 10 years. She had scoliosis (curved spine) from age 11 and had Harrington rods (treating scoliosis) inserted at age 13. From about that time, she was unable to walk unaided. At age 20 she was ambulant with assistance, but had upper and lower limb ataxia. There was some restriction of upward and lateral gaze, no nystagmus (involuntary eye movement), distal diminution of vibration sense
and light touch, absent deep tendon reflexes and equivocal plantar responses. At age 31, she had developed titubation (tremor of the head) and nystagmus in the primary position, and had demonstrated central dysautonomia (autonomic nervous system dysfunction) with gastroparesis (partial paralysis of the stomach) on formal testing.

Patient II-1 was seen at 56 years. She had no symptoms of concern. She had very subtle incoordination of upper and lower limbs, with slight gaze-evoked nystagmus. Reflexes and sensation were normal as were nerve conduction studies (measurement of speed of conduction of electrical impulse through a nerve). The pedigree indicates that II-1 is an obligate carrier.

Patient II-3 had an onset of ataxia at 21 years. She was examined at 58 years and required a wheelchair for mobility. She had pes cavus, claw toes, moderate cerebellar dysarthria (motor speech disorder), impaired vestibulo-ocular reflex gain (important for stabilising vision), gaze-evoked horizontal nystagmus, mild vertical nystagmus, hypermetric saccades, gross upper and lower limb ataxia, absent reflexes in upper and lower limbs, extensor plantar responses and diminished sensation in her feet.

Patient III-2 was diagnosed with scoliosis at 14 years and developed ataxia from 20 years. At age 21 there was minimal gait ataxia, slight finger nose-ataxia and ataxia on the heel-knee-shin test. There was no dysarthria, slightly impaired distal sensation, absent reflexes, weakly extensor plantar responses and no nystagmus. At 32 she required the use of a walker and had reduced night vision.

Patient III-5 was asymptomatic at 37 years. Examination findings included diminished upper and lower limb reflexes but no evidence of ataxia. Brain MRI was normal.

Patient IV-1 developed ataxia at 5 years. At age 6 he was unable to tandem walk, with a positive Romberg sign and absent reflexes, nystagmus but no dysarthria.

Patient IV-2 also had an onset of ataxia at 5 years. He required a K-walker from 6 years. Examination revealed upper and lower limb ataxia, absent reflexes and dysarthria.

Nerve conduction studies on individuals III-5, IV-1 and IV-2 revealed a sensory neuropathy, with normal motor studies in III-5 and IV-2 (not assessed in IV-1). All symptomatic individuals (II-3, III-1, III-2, IV-1, IV-2) had cerebellar atrophy on brain MRI scan. The following investigations were normal in one or more individuals: SCA1, 2, 3, 6, 7, 8 and 17, sequencing of the mitochondrial genome, the gene encoding frataxin (FXN)
intron 1 GAA expansions (Friedreich ataxia), vitamin E, phytanic acid, lysosomal enzymes, chromosomal microarray, fragile X DNA study, transferrin isoforms, very long chain fatty acids, cholesterol, lactate, ammonia, urine metabolic screen for amino and organic acids, alpha-fetoprotein and immunoglobulins.

4.3.2 SNP chip linkage analysis

The Illumina estimated gender matched the recorded gender as per the pedigree for all individuals genotyped, and call rates were over 99.2% (Table B.1). The best matched HapMap3 populations were CEU and TSI (that have similar allele frequencies), as expected for a family of European descent (Table B.2). Inferred pairwise relationships were appropriately captured by the given pedigree. FEstim estimated an inbreeding coefficient of 0 (non-consanguineous parents) for all individuals (Table B.3). LINKDATAGEN selected 11,995 SNP markers for linkage analysis. The peak LOD score from MERLIN was 1.66 with four peaks identified on 1q24.2–q25.2, 2p21–p16.1, 10p14-p13 and 15p21.3-q23 (Table 4.2 and Figure 4.2) reducing the search space to 1.5% of the autosome by base pairs. All six affected individuals plus the obligate carrier II-1 were found to carry the susceptibility haplotype relevant for each linked region (Figures 4.4 to 4.7).

![Figure 4.2. Linkage analysis LOD scores for genome](image)

Linkage analysis LOD scores across the autosome. Chromosome names are given along the top.

The linkage regions were found to overlap with only one previously identified SCA locus, SCA25. The 2p21–p16.1 linkage region identified in this study falls entirely within the region delineated by Stevanin et al. (2004) (Figure 4.3).
Figure 4.3. Chromosome 2 linkage analysis LOD scores

The purple shaded area is the SCA25 locus as reported by Stevanin et al. (2004).

Table 4.2. Linkage analysis critical regions

<table>
<thead>
<tr>
<th>Chr</th>
<th>SNP start</th>
<th>SNP end</th>
<th>Cytogenetic location</th>
<th>Genetic map start (cM)</th>
<th>Genetic map end (cM)</th>
<th>Genetic map length (cM)</th>
<th>hg19 start (bp)</th>
<th>hg19 end (bp)</th>
<th>Physical length (bp)</th>
<th>Peak LOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>rs12094807</td>
<td>rs10913665</td>
<td>1q24.2–q25.2</td>
<td>185.536</td>
<td>194.62</td>
<td>9.084</td>
<td>168,461,703</td>
<td>178,945,967</td>
<td>10,484,265</td>
<td>1.6600</td>
</tr>
<tr>
<td>chr2</td>
<td>rs1300537</td>
<td>rs10209275</td>
<td>2p21–p16.1</td>
<td>76.653</td>
<td>86.622</td>
<td>9.969</td>
<td>47,771,677</td>
<td>59,791,233</td>
<td>12,019,557</td>
<td>1.6600</td>
</tr>
<tr>
<td>chr10</td>
<td>rs2248474</td>
<td>rs2997063</td>
<td>10p14–p13</td>
<td>16.773</td>
<td>26.901</td>
<td>10.128</td>
<td>7,285,632</td>
<td>12,389,180</td>
<td>5,103,549</td>
<td>1.6600</td>
</tr>
<tr>
<td>chr15</td>
<td>rs685850</td>
<td>rs12911091</td>
<td>15p21.3–q23</td>
<td>64.754</td>
<td>87.49</td>
<td>22.736</td>
<td>53,751,702</td>
<td>70,443,385</td>
<td>16,691,684</td>
<td>1.6600</td>
</tr>
</tbody>
</table>
Figure 4.4. Haplotypes of chromosome 1 critical linkage region

The disease haplotype is dark blue. Samples with no ataxia diagnosis are not shown for ethical considerations. Grey affectedness status indicates the sample was modelled with unknown affectedness status.
Figure 4.5. Haplotypes of chromosome 2 critical linkage region

The disease haplotype is pink. The location of variants found with WGS and WES is shown with the associated gene. Samples with no ataxia diagnosis are not shown for ethical considerations. Grey affectedness status indicates the sample was modelled with unknown affectedness status.
The disease haplotype is red orange. Samples with no ataxia diagnosis are not shown for ethical considerations. Grey affectionedness status indicates the sample was modelled with unknown affectioned status.
The disease haplotype is dark purple. Samples with no ataxia diagnosis are not shown for ethical considerations. Grey affectedness status indicates the sample was modelled with unknown affectedness status.
4.3.3 Exome/genome sequencing

The WES data had a 6.0% PCR duplicate rate, with the median coverage ranging from 116 to 123 over the UCSC Known Gene coding regions (Table B.4). WGS data from the HiSeq 2500 platform had a duplicate rate of 4.6%, while the HiSeq X Ten data had a much higher duplicate rate of 42% (Table B.5). Despite this, the coverage after duplicate removal was still higher in the HiSeq X Ten than Hiseq 2500 data and above the specified coverage of 60x. The sequencing after filtering PCR duplicates had a median coverage ranging from 80 to 85 for WES over capture regions, 55 for WGS from Illumina HiSeq 2500 and 77 to 85 for WGS HiSeq X Ten data.

The genotypes called from the WES, WGS and SNP chip data on the same samples were over 99.95% concordant (Tables B.6 and B.7) in all three pairwise comparisons.

Almost four times as many variants were found with WGS compared to WES, but after filtering non-exonic and non-splicing variants a similar number remained (WES: 24, WGS: 25) (Table 4.3). After filtering, all variants found in WES were also found in WGS, but two deletions called as heterozygous in only two individuals of WGS were not found in WES. After visual inspection in IGV these were deemed to be likely false positive calls due to 13 and 14 bp poly(A) sequences that can cause polymerase stutter in sequencing.

Four variants of interest remained in the genes encoding DNA mismatch repair protein Msh6 \((MSH6)\) (NM_000179.2:c.2633T>C:p.Val878Ala), stonin-1 \((STON1)\) (variant is also within \(STON1\)-\(GTF2A1L\) readthrough \((STON1\)-\(GTF2A1L)\) (NM_001198595.1:c.1231G>A:p.Glu411Lys), proteasome activator complex subunit 4 \((PSME4)\) (NM_014614.2:c.3400G>A:p.Glu1134Lys) and polynucleotide nucleotidyltransferase 1, mitochondrial \((PNPT1)\) (NM_033109.4:c.2014-3C>G – splice site intron variant) (Table 4.4). There were no variants remaining after filtering in the other three linkage regions.
### Table 4.3. Number of variants after filtering steps

Number of variants detected and remaining after each filtering step, for both WES and WGS data.

<table>
<thead>
<tr>
<th>Cumulative filter applied</th>
<th>WES</th>
<th>WGS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>SNVs</td>
</tr>
<tr>
<td>All</td>
<td>1,651,031</td>
<td>1,447,445</td>
</tr>
<tr>
<td>Linkage</td>
<td>31,535</td>
<td>27,697</td>
</tr>
<tr>
<td>1000 Genomes AAF &lt;= 0.01</td>
<td>4,043</td>
<td>2,459</td>
</tr>
<tr>
<td>ESP 6500 AAF &lt;= 0.01</td>
<td>4,028</td>
<td>2,444</td>
</tr>
<tr>
<td>ExAC AAF &lt;= 0.05</td>
<td>4,017</td>
<td>2,437</td>
</tr>
<tr>
<td>Control allele number &lt;10</td>
<td>4,004</td>
<td>2,426</td>
</tr>
<tr>
<td>Exonic or splicing</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>Splicing or not synonymous</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>Mismatch tolerant disease model</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Mismatch intolerant disease model</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
### Table 4.4. Variants after filtering

Allele frequencies of the alternative allele are given for the Exome Aggregation Consortium (ExAC) and 1000 Genomes. NS SNV is a non-synonymous SNV. Residual Variation Intolerance Score (RVIS) scores (ALL_0.1%) that are lower indicate the gene is less tolerant to variations, with the percentile relative to all genes.

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Start</th>
<th>HapMap2 genetic map position (cM)</th>
<th>Ref</th>
<th>Alt</th>
<th>Gene</th>
<th>Change</th>
<th>Protein change</th>
<th>dbSNP146</th>
<th>ExAC</th>
<th>1000 Genomes</th>
<th>Polyphen2</th>
<th>SIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr2</td>
<td>48,027,755</td>
<td>77.35</td>
<td>T</td>
<td>C</td>
<td>MSH6</td>
<td>NS SNV</td>
<td>NM_000179.2:p.Val878Ala</td>
<td>rs2020912</td>
<td>5.3E-03</td>
<td>1.0E-02</td>
<td>Benign</td>
<td>Tolerated</td>
</tr>
<tr>
<td>chr2</td>
<td>48,809,003</td>
<td>77.56</td>
<td>G</td>
<td>A</td>
<td>STON1</td>
<td>STON1-GTF2A1L</td>
<td>NS SNV</td>
<td>NM_001198595.1:p.Glu411Lys</td>
<td>rs760676566</td>
<td>8.1E-06</td>
<td>-</td>
<td>Probably damaging</td>
</tr>
<tr>
<td>chr2</td>
<td>54,127,040</td>
<td>82.87</td>
<td>C</td>
<td>T</td>
<td>PSME4</td>
<td>NS SNV</td>
<td>NM_014614.2:p.Glu1134Lys</td>
<td>rs142064799</td>
<td>1.2E-03</td>
<td>5.0E-04</td>
<td>Benign</td>
<td>Tolerated</td>
</tr>
</tbody>
</table>

### Table 4.5. CNV variants found in WGS with QDNAseq, DNAcopy and CGHcal

Pr(state) gives the probability of the copy-number state given by CGHcall. CN=copy-number. DGV=database of genomic variants. Grey indicates a CNV called only in III-3, green/orange a greater/smaller position than the CN call for II-3. *Outside the strict linkage region, but included when extended by 1 Mb. #Total sample count includes multiple DGV features.

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Start</th>
<th>end</th>
<th>width</th>
<th>log2 ratio</th>
<th>Pr(state)</th>
<th>start</th>
<th>end</th>
<th>width</th>
<th>log2 ratio</th>
<th>Pr(state)</th>
<th>ID</th>
<th>Observed</th>
<th>N</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>54,659</td>
<td>54,665</td>
<td>6</td>
<td>-0.96</td>
<td>0.719</td>
<td>54,659</td>
<td>54,665</td>
<td>6</td>
<td>-1.22</td>
<td>0.719</td>
<td>-</td>
<td>-</td>
<td>esv3587975</td>
<td>808L</td>
</tr>
<tr>
<td>chr2</td>
<td>52,748</td>
<td>52,785</td>
<td>37</td>
<td>-0.83</td>
<td>0.717</td>
<td>52,750</td>
<td>52,785</td>
<td>35</td>
<td>-0.85</td>
<td>0.717</td>
<td>-2</td>
<td>13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>chr2</td>
<td>52,748</td>
<td>52,785</td>
<td>37</td>
<td>-0.83</td>
<td>0.717</td>
<td>56,655</td>
<td>56,655</td>
<td>3</td>
<td>-5.54</td>
<td>0.934</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>esv2660641</td>
</tr>
<tr>
<td>chr10</td>
<td>12,347</td>
<td>12,380</td>
<td>33</td>
<td>-1.31</td>
<td>0.719</td>
<td>12,347</td>
<td>12,380</td>
<td>33</td>
<td>-1.29</td>
<td>0.719</td>
<td>0</td>
<td>15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>chr15</td>
<td>55,585</td>
<td>55,604</td>
<td>19</td>
<td>-1.66</td>
<td>0.72</td>
<td>55,585</td>
<td>55,604</td>
<td>19</td>
<td>-1.65</td>
<td>0.72</td>
<td>0</td>
<td>15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>chr15</td>
<td>55,585</td>
<td>55,604</td>
<td>19</td>
<td>-1.66</td>
<td>0.72</td>
<td>55,585</td>
<td>55,604</td>
<td>19</td>
<td>-1.65</td>
<td>0.72</td>
<td>0</td>
<td>15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>chr15</td>
<td>58,443</td>
<td>58,448</td>
<td>5</td>
<td>-1.05</td>
<td>0.719</td>
<td>58,442</td>
<td>58,448</td>
<td>6</td>
<td>-0.94</td>
<td>0.719</td>
<td>1</td>
<td>8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>chr15</td>
<td>62,533</td>
<td>62,588</td>
<td>55</td>
<td>-0.77</td>
<td>0.716</td>
<td>62,533</td>
<td>62,588</td>
<td>55</td>
<td>-0.74</td>
<td>0.716</td>
<td>0</td>
<td>15</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Outside the strict linkage region, but included when extended by 1 Mb. #Total sample count includes multiple DGV features.
4.3.4 Copy-number variations

Analysis of the SNPchip data with PennCNV detected 30 unique CNV loci across the six DNA samples derived from affected individuals, with a median of 8 CNVs called per sample. None of these were within the linkage regions. PennCNV found no CNVs common to all affected samples for the genome.

Analysis of the WGS with the QDNAseq pipeline detected at least 1,454 (26 gains, 1,428 losses) and 1,452 (32, 1,420) distinct CNVs in the II-3 and III-2 sample autosomes respectively. PennCNV calls with confidence scores of at least 50 were concordant in the direction of the change (gain or loss) in the QDNAseq pipeline but called single copy deletions as double copy deletions (copy-number 0). After filtering WGS CNVs present in controls, 31 (6 gains, 26 deletions) and 35 (5 gains, 30 deletions) CNVs were found, with only a single deletion found in both affected individuals at chr5:27,606,001-27,639,000. Due to haplotype sharing this CNV was likely to have been inherited from I-1 or I-2. This CNV was at a similar locus to the rare deletion dgv1589e212 (7 losses in 873 (0.8%) samples) (Figure 4.8(a)) in the Database of Genomic Variants (DGV) (MacDonald et al., 2014).

Within the linkage regions there were 7 (0 gains, 7 deletions) and 9 (0 gains, 9 deletions) CNVs in the II-3 and III-2 samples (Table 4.5), but after filtering CNVs found in controls, only one copy-number deletion was detected in III-2 uniquely, but this CNV overlapped with a common deletion (Figure 4.8(b)) (DGV esv3587975 with observed loss of 808 in 2504 (32%) samples 1000 Genomes Project Consortium et al. (2015).

No novel CNVs were detected in all affected individuals in WGS or SNP chip data.

4.3.5 Validation by Sanger sequencing

Sanger sequencing confirmed the four WES and WGS variants were called correctly. The three variants discovered in \textit{STON1} (\textit{STON1-GTF2A1L}), \textit{PNPT1} and \textit{PSME4} (Table 4.6) were confirmed to segregate with disease in DNA derived from affected individuals. Conversely, the \textit{MSH6} found to be wildtype in the affected individual II-3 (Figure 4.9).

The \textit{MSH6} variant was shown to segregate with disease only in II-2 and her descendants (six affected individuals), while the genotype was homozygous reference in the affected individual II-3. This result was inconsistent with the result that all affected
individuals shared the haplotype at the \textit{MSH6} variant, including after close inspection of SNP genotypes. Two replicates confirmed the genotype of II-3. Sample swaps appeared unlikely given the XIBD relatedness results. No evidence of a copy-number deletion in II-3 at this locus could be found in PennCNV or QDNAseq results, nor was there any apparent loss of coverage in WGS when inspected in IGV. Furthermore, heterozygous variants were found within 2000 bp either-side of this variant indicating no loss of heterozygosity that would be indicative of a large deletion. This may be a recurrent mutation or a small segmental uniparental isodisomy event but these hypotheses cannot be tested with our data.
4.3.6 Analysis of repeats for putative repeat expansions

All alleles of known repeat expansion disorder loci called with ExpansionHunter were within normal ranges (Figure 4.10) as expected, since none of the known repeat expansion loci were located in any of the four linkage regions, including the SCA25 region. The called alleles also did not appear as outliers when compared to the 15 control samples.
Table 4.6. Variants after filtering and Sanger verification

Ref = “reference allele”, Alt = “alternative allele”. Allele frequencies of the alternative allele are given for the Exome Aggregation Consortium (ExAC) (Lek et al., 2016) and 1000 Genomes. Residual Variation Intolerance Score (RVIS) scores that are lower indicate the gene is less tolerant to variations, with the percentile relative to all genomic genes.

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Position (hg19)</th>
<th>HapMap2 genetic map position (cM)</th>
<th>Ref</th>
<th>Alt</th>
<th>Gene(s)</th>
<th>Protein change</th>
<th>dbSNP146</th>
<th>ExAC</th>
<th>1000 Genomes</th>
<th>Polyphen2</th>
<th>SIFT</th>
<th>RVIS (percentile)</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr2</td>
<td>48,809,003</td>
<td>77.56</td>
<td>G</td>
<td>A</td>
<td>STON1, STON1-GTF2A1L</td>
<td>NM_001198595.1:p.Glu411Lys</td>
<td>rs760676566</td>
<td>8.1 x 10^{-6}</td>
<td>Probably damaging</td>
<td>Tolerated</td>
<td>1.6 (96), 1.2 (95)</td>
<td></td>
</tr>
<tr>
<td>chr2</td>
<td>54,127,040</td>
<td>82.87</td>
<td>C</td>
<td>T</td>
<td>PSME4</td>
<td>NM_014614.2:p.Glu1134Lys</td>
<td>rs142064799</td>
<td>1.2 x 10^{-3}</td>
<td>Benign</td>
<td>Tolerated</td>
<td>-1.1 (7.6)</td>
<td></td>
</tr>
<tr>
<td>chr2</td>
<td>55,870,351</td>
<td>83.78</td>
<td>G</td>
<td>C</td>
<td>PNPT1</td>
<td>Splice site</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.14 (64)</td>
</tr>
</tbody>
</table>

Figure 4.10. ExpansionHunter alleles for affected individuals sequenced with Illumina HiSeq X, by locus and sample

Blue bars are the largest repeat number usually observed in unaffected individuals, and red bars the minimum repeat number usually observed in affected individuals. ExpansionHunter was designed for PCR-free data, but as our WGS used a PCR step in library preparation ExpansionHunter often called more than two alleles for a diploid sample. Alleles sizes were either derived from reads that span the STR, with source SPANNING, while those that overlapped but did not span contributing to FLANKING allele size calls. FLANKING alleles were given a confidence interval, here indicated by a horizontal bar at the data point, but as these intervals were only 1 repeat number wide they are difficult to visualise here.
A total of 5,995 STRs were found in the linkage regions (1,533 in SCA25 overlap region), of which twelve were within coding regions as putative functional STRs, however, none of these STRs appeared to be expanded when inspected in IGV (Figure 4.11(a), Table 4.7), with insertions no more than one repeat unit longer than the reference while soft-clipped bases did not consist of the repeat.

After filtering of lobSTR calls, six STRs remained (Figure 4.11(b), Table 4.8). Four of these STRs were filtered as the allele lengths were observed in the 1000 Genomes Project data (assessed in the STR Catalogue Viewer). One pentanucleotide repeat 18 kb from the gene encoding vesicle-associated membrane protein 4 (VAMP4) did not have an entry in the STR Catalogue Viewer nor was it assessed as a potential eSTR, thus this repeat was retained as a candidate. The remaining STR that was within intron 1 of the gene encoding pappalysin-2 (PAPPA2) was not significant as an eSTR (Table 4.9). Both remaining STRs were inspected in IGV for the IV-2 sample with WGS on the Illumina HiSeq 2500 platform that had reads supporting an expanded allele for both STRs.

In GTEx, both VAMP4 and PAPPA2 show expression in the cerebellar regions and generally across brain regions. VAMP4 shows the highest expression in the “cerebellar hemisphere” region compared to all other 52 human tissues sampled. PAPPA2 was only the most highly expressed in the “cerebellar hemisphere” region of the 13 brain regions sampled in GTEx and is most highly expressed in the “kidney - cortex” tissue.
Table 4.7. Coding STRs of 2–12 bp within linkage regions

STRs of 2–12 bp from the Simple Repeats track (Tandem Repeats Finder) of UCSC Genome Browser found within linkage regions and coding regions of genes. STRs were inspected within IGV for indications of repeat expansions and noted. Columns from tandem repeats finder include the STR period, copy number, consensus size, percentage match, percentage indel, score, percentage of A/C/G/T and entropy score. Genes encode the proteins SUN domain-containing ossification factor (SUCO), uncharacterized protein KIAA0040 (KIAA0040), F-box only protein 11 (FBXO11), lutropin-choriogonadotropic hormone receptor (LHCGR), neurexin-1-beta (NRXN1), reticulon-4 (RTN4), B-cell lymphoma/leukemia 11A (BCL11A), transcription initiation factor TFIID subunit 3 (TAF3), hepatocyte nuclear factor 6 (ONECUT1), DNA-binding protein RFX7 (RFX7), sodium/potassium/calcium exchanger 1 (SLC24A1) and SKI family transcriptional corepressor 1 (SKOR1).

Table 4.7.
<table>
<thead>
<tr>
<th>Chrom</th>
<th>Start</th>
<th>End</th>
<th>Motif</th>
<th>Repeat number</th>
<th>Gene</th>
<th>IGV inspection</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>172,555,054</td>
<td>172,555,087</td>
<td>TCAACTCTGTT</td>
<td>2.8</td>
<td>SUCO</td>
<td>Appeared normal</td>
</tr>
<tr>
<td>chr1</td>
<td>175,129,914</td>
<td>175,129,962</td>
<td>TCT</td>
<td>16</td>
<td>KIAA0040</td>
<td>9 bp deletion and 3 bp insertions</td>
</tr>
<tr>
<td>chr2</td>
<td>48,132,690</td>
<td>48,132,800</td>
<td>GGGTCCTGTCGC</td>
<td>9</td>
<td>FBXO11</td>
<td>Soft clipped reads not representative of repeat</td>
</tr>
<tr>
<td>chr2</td>
<td>48,982,756</td>
<td>48,982,790</td>
<td>GCTGCAGCA</td>
<td>3.9</td>
<td>LHCGR</td>
<td>6 bp insertion in III-2 and IV-2</td>
</tr>
<tr>
<td>chr2</td>
<td>60,688,535</td>
<td>60,688,594</td>
<td>CTC</td>
<td>20</td>
<td>BCL11A</td>
<td>Appeared normal</td>
</tr>
<tr>
<td>chr10</td>
<td>8,006,417</td>
<td>8,006,448</td>
<td>CCAAGAGCC</td>
<td>3.6</td>
<td>TAF3</td>
<td>Appeared normal</td>
</tr>
<tr>
<td>chr15</td>
<td>53,081,666</td>
<td>53,081,713</td>
<td>TGG</td>
<td>16</td>
<td>ONECUT1</td>
<td>Appeared normal</td>
</tr>
<tr>
<td>chr15</td>
<td>56,387,062</td>
<td>56,387,088</td>
<td>GTGGGT</td>
<td>4.5</td>
<td>RFX7</td>
<td>Appeared normal</td>
</tr>
<tr>
<td>chr15</td>
<td>65,943,068</td>
<td>65,943,157</td>
<td>GAG</td>
<td>30</td>
<td>SLC24A1</td>
<td>Appeared normal</td>
</tr>
<tr>
<td>chr15</td>
<td>68,119,717</td>
<td>68,119,743</td>
<td>GGGCCCGGCAGC</td>
<td>2.2</td>
<td>SKOR1</td>
<td>Soft clipped bases in II-3 and IV-2 samples not representative of repeat</td>
</tr>
</tbody>
</table>

Table 4.8. Annotated lobSTR STR calls

Those remaining after filtering by larger outliers. *The observed allele in the affected individuals was larger than observed in the STR catalogue.

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Start</th>
<th>End</th>
<th>Motif</th>
<th>Genome reference</th>
<th>Mode</th>
<th>Range</th>
<th>Mode</th>
<th>Range</th>
<th>A1-11</th>
<th>A1-2</th>
<th>Name</th>
<th>Distance to exon</th>
<th>eSTR p-value</th>
<th>Controls</th>
<th>STRCat</th>
<th>Closest gene(s)</th>
<th>Closest gene GTEX RPKM</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>169,150,364</td>
<td>169,150,430</td>
<td>CTAT</td>
<td>67</td>
<td>63</td>
<td>51-75</td>
<td>63</td>
<td>51-83+</td>
<td>67/79</td>
<td>75/79</td>
<td>NME7 (intron)</td>
<td>9.5 kb</td>
<td>0.4</td>
<td>3.452</td>
<td>3.203</td>
<td>3.221</td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>171,651,720</td>
<td>171,651,765</td>
<td>CTTTT</td>
<td>46</td>
<td>51</td>
<td>46-66</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>ATP1B1</td>
<td>48 kb</td>
<td>0.01</td>
<td>49.907</td>
<td>54.486</td>
<td>54.486</td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>176,479,941</td>
<td>176,479,985</td>
<td>AC</td>
<td>45</td>
<td>45</td>
<td>37-53</td>
<td>45</td>
<td>31-57+</td>
<td>41/67*</td>
<td>45/65*</td>
<td>VAMP4</td>
<td>18 kb</td>
<td>-</td>
<td>6.969</td>
<td>15.298</td>
<td>12.421</td>
<td></td>
</tr>
<tr>
<td>chr2</td>
<td>50,635,976</td>
<td>50,636,013</td>
<td>TC</td>
<td>38</td>
<td>36</td>
<td>22-52</td>
<td>22</td>
<td>16-60</td>
<td>38/58</td>
<td>44/60</td>
<td>NRXN1 (intron)</td>
<td>57 kb</td>
<td>-</td>
<td>5.935</td>
<td>15.136</td>
<td>10.766</td>
<td></td>
</tr>
<tr>
<td>chr15</td>
<td>57,504,106</td>
<td>57,504,148</td>
<td>TC</td>
<td>43</td>
<td>29</td>
<td>27-43</td>
<td>29</td>
<td>24-45</td>
<td>29/45</td>
<td>29/45</td>
<td>ZNF280D (intron)</td>
<td>37 kb</td>
<td>0.6</td>
<td>2.246</td>
<td>4.438</td>
<td>3.741</td>
<td></td>
</tr>
<tr>
<td>chr15</td>
<td>57,666,737</td>
<td>57,666,789</td>
<td>TG</td>
<td>53</td>
<td>47</td>
<td>43-53</td>
<td>47</td>
<td>23-59+</td>
<td>47/61</td>
<td>47/59</td>
<td>CGNL1</td>
<td>2.0 kb</td>
<td>0.08</td>
<td>5.283</td>
<td>0.972</td>
<td>1.538</td>
<td></td>
</tr>
</tbody>
</table>

Chapter 4. MPS of family with spinocerebellar ataxia mapping to the SCA25 locus
**4.4 Discussion**

We have identified a family with six individuals affected with an autosomal-dominant progressive ataxia. Known repeat expansion SCA loci tested negative, and this result was concordant with negative results from ExpansionHunter on Illumina HiSeq X Ten data of the two affected individuals.

Linkage analysis reduced the search space for a causative variant to 1.5% of the autosome that included the SCA25 locus, originally mapped in a French family in Stevanin et al. (2004), and thus may be the second reported SCA25 family in the literature. In support of this family having SCA25 was the marked variability in age of onset and severity of symptoms as well as the presence of sensory neuropathy. The gastroparesis in III-1 accorded with the gastrointestinal symptoms in the French family, and scoliosis was found in two patients in this study and two patients in the French family. It is unknown whether these families are related, which would indicate a common genetic cause, or if the phenotypes in the two families are caused by two different mutations. Identification of a causal variant will be required to confirm that the family presented in this chapter and the French family in Stevanin et al. (2004) have the same underlying genetic aetiology. It is important to note that we do not know for certain which of the four haplotypes is carrying the disease mutation, but suspect the 2p21–p16.1 region due to overlap with the SCA25 locus and common symptoms with the French family.

For ethical reasons, we have not presented haplotype or variant results for descendants of I-1 and I-2 who currently are asymptomatic. As the disorder had a variable age of onset, (with one person, II-1, an obligate carrier of the family mutation being essentially asymptomatic at 56 years), haplotype or variant information could indicate predisposition to late-onset in asymptomatic individuals.
Across all four linkage regions, we detected three rare variants segregating with disease. No variants remained after filtering in the chromosome 1, 10 and 15 linkage regions, with all three in the SCA25 locus. The two nonsynonymous variants were found to be tolerated by SIFT. The variant in \textit{STON1} (\textit{STON1-GTF2A1LI}) was the only variant found to be probably damaging by PolyPhen2, but RVIS indicated that rare variants are overrepresented in \textit{STON1} and \textit{STON1-GTF2A1LI}. The nonsynonymous variant in \textit{PSME4} was found in a heterozygous form in 152 of the ExAC samples thus appearing unlikely to be causative given the rarity and severity of this ataxia. The variant in \textit{PNPT1} is intronic and within 3 bp of a splice junction, therefore likely to affect splicing of \textit{PNPT1}. Variants in \textit{PNPT1} have previously been associated with either combined oxidative phosphorylation deficiency-13 (COXPD13, OMIM \#614932) causing early-onset encephalomyopathy or autosomal recessive deafness-70 (DFNB70, OMIM \#614934), but have not been associated with ataxia, while ClinVar (Landrum et al., 2016) lists this variant as having uncertain significance (RCV000200258.1). The SNVs in \textit{STON1} (\textit{STON1-GTF2A1L}) and \textit{PNPT1} may be causal but do not have a clear mechanism of pathogenesis; a second-hit in either of these genes in an independent family affected by ataxia is required to support causality.

In light of recent findings identifying modifier genes for Huntington disease (Genetic Modifiers of Huntington’s Disease (GeM-HD) Consortium, 2015) and SCA (Bettencourt et al., 2016), including mismatch repair genes and pathways, we examined the rare variant in the mismatch repair gene \textit{MSH6} more closely. The variant was found in five of six affected individuals and an obligate carrier, but without apparent effect on the severity of SCA. This variant has been classified as benign in the International Society for Gastrointestinal Hereditary Tumours (InSiGHT) database (Thompson et al., 2014). The InSiGHT database classifies variants in four Lynch Syndrome genes associated with mismatch repair. This further suggests that the \textit{MSH6} variant is not associated with this SCA.

There was evidence of anticipation in this family. Therefore, with the knowledge that many SCAs are caused by repeat expansions, it was important to search for novel disease causing repeat expansions. We identified two STRs within the linkage regions close to \textit{VAMP4} and \textit{PAPPA2} that had expanded alleles with respect to internal controls and an external database based on Phase I of the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2012). However, the result of \textit{PAPPA2} must be interpreted with caution, as the STR Catalogue viewer may not have annotated larger...
alleles due to the use of shorter reads than this study. Our STRs are 18 kb and 45 kb from gene exons respectively, further from exons than any other repeat expansion disorder to date; SCA10 is the most extreme reported example, caused by an intronic ATTCT repeat 12 kb from an exon of \textit{ATXN10}. The CCTTT STR close to \textit{VAMP4} was not assessed in the STR Catalogue Viewer and is only one repeat unit longer than the largest of the alleles in the controls, and therefore together with its distance from exons we deemed this unlikely to be pathogenic. Thus far, no dinucleotide alleles have been implicated in repeat expansion disorders in humans and intronic repeat expansions have involved alleles that are hundreds of bp longer than normal sized alleles, rather than the 7 bp increase observed here; therefore, the AC repeat in the intron of \textit{PAPPA2} also appeared unlikely to be causative.

Here we have attempted to find small repeat expansions with lobSTR and signals of larger expansions by manual inspection of reads, similar to the discovery of the FTDALS1 locus. Repeat expansion detection will benefit from longer-read technologies that allow reads to span large repeats.

4.5 Conclusion

We have extensively investigated a spinocerebellar ataxia family for SNVs, CNVs and repeat expansions but have been unable to find a strong candidate as to the genetic cause. Linkage results overlapped with the SCA25 locus and thus two families are now published with linkage to this locus, yet with no identified genetic mechanism. In both families, sensory neuropathy is a prominent feature and there is marked clinical variability ranging from onset in young children through to individuals who are essentially asymptomatic in their sixties. Scoliosis and gastrointestinal symptoms were also found in some patients of both families.
Chapter 5

Discussion and summary

5.1 Use of MPS to determine the cause of genetic disorders

The use of massively parallel sequencing (MPS) to determine the cause of genetic disorders is now well established. MPS is now being integrated into healthcare, such as at the Melbourne Genomics Health Alliance (MGHA) and Australian Genomics Health Alliance (AGHA), where MPS is used as a diagnostic tool. This is often faster and more efficient than attempting to diagnose with disease-specific genetic tests, that may return many negative results before a hit is found (Stark et al., 2016; Tan et al., 2017). Additionally, the causes of many rare disorders have been discovered with MPS, expanding the entries in OMIM. Other OMIM entries have been merged after discovery with MPS that these were the same disorders. MPS can also be used as a general screening tool; while MPS screening is not routinely performed, this will likely become more common in the future (including screening with newer long-read technologies).

The greater affordability of whole-genome sequencing (WGS) has made it more likely to be chosen for analysis over whole-exome sequencing. One barrier to performing WGS is that the datasets are larger, by an order of magnitude, than WES, and hence WGS requires greater computational and storage capacities. WGS has enabled better detection of non-coding variants, but methods to identify these variants as pathogenic still require much more development. Methods for detecting single-nucleotide variants (SNVs), short-indels and copy number variations (CNVs) are well developed. The development of methods to
detect repeat expansions of short tandem repeats (STRs) in MPS is in its infancy and has yet to be tested on large-numbers of cases to compare sensitivity and specificity.

Another advantage of MPS is that it can genotype modifiers of repeat expansion disorders. At present, only two genome-wide significant loci with repeat expansion disorder modifiers have been identified (Genetic Modifiers of Huntington’s Disease (GeM-HD) Consortium, 2015; Bettencourt et al., 2016), but further studies may find more loci and identify the causal variants of existing loci. These modifier loci were found to influence factors such as the age of onset and the severity of the symptoms, and appear to be related to STR (microsatellite) instability. They may also be useful in developing a more powerful test statistic for the detection of repeat expansions.

5.2 Developments in this thesis

In this thesis, we have developed one of the first methods, exSTRa, alongside ExpansionHunter and STRetch, that can be used to detect expansions with MPS. We showed that exSTRa can be used to identify expansions in seven different disorders, including five ataxias that can be difficult to distinguish without genetic tests. For detecting Fragile-X syndrome (FRAXA), a polymerase chain reaction–free (PCR-free) library preparation protocol was required to mitigate low-coverage bias caused by high-GC content. In this thesis, we had too few control cases with PCR-free data to determine if exSTRa could detect FRAXA expansions. Other repeat expansion disorder loci will also likely be detected with exSTRa, and verifying this is future work.

In Chapter 4 we described investigations to find the cause of a spinocerebellar ataxia in a family with apparent anticipation, by a combination of linkage and MPS analysis. The disorder mapped to four genomic locations, one of which overlapped with the SCA25 locus (OMIM %608703). We found two rare variants of unknown significance in the genes STON1 (or read-through gene STON1-GTF2A1L) and PNPT1; finding a second hit in either of these genes in unrelated spinocerebellar ataxia patients (without previous diagnosis) would greatly increase the evidence for the role of that gene in the disease. We applied some of our insights into expansions of STRs to attempt to find a causal variant with some heuristic methods, but this family as yet remains unsolved.
5.3 Further directions

At the start of this thesis, no algorithm was published for the detection of repeat expansions with MPS data. As of August 2017, two additional algorithms to exSTRa have been published on the preprint server bioRxiv (pronounced bio-archive), being ExpansionHunter (December 2016) and STRetch (July 2017). At present, ExpansionHunter, exSTRa and STRetch are based on widely-varying paradigms, such that it may be possible the best expansion calls will be made using a combination of these methods. Our method did not make use of mapping-quality or base-quality scores, which could further improve the reliability of exSTRa expansion calls. The exSTRa software may be erroneously counting reads that do not actually overlap the STR; ensuring we do not count these may be an improvement.

We presented the first published analysis of repeat expansions in MPS data with a PCR step in library preparation in WGS and WES. We showed that exSTRa and ExpansionHunter could find expansions using MPS with PCR-free library preparation. This experimental setup has not yet been assessed with STRetch.

The methods presented in this thesis could be extended to interrogate all genomic STR loci, as exSTRa is a case-control method that makes no assumptions on the underlying allele sizes. Our software only requires knowledge of the STR location and its repeat motif, such that hundreds of thousands more STRs can be included in the exSTRa input. Similarly, both ExpansionHunter and STRetch can be extended in a similar way. This would be highly applicable to the family in Chapter 4 where a repeat expansion is suspected but as yet not identified. Mitigating the effects of multiple-testing of thousands of STR loci within linkage regions may be problematic as we may lack power to overcome the effect of multiple-testing correction. Currently, no software is attempting to call STR expansions without an STR database, such as the Tandem Repeats Finder (Benson, 1999) or a manually curated repeat expansion disorder database as used in this thesis.

exSTRa may not be able to easily distinguish expanded and normal samples for loci where the intermediate range is small. Our method does not explicitly detect the purity of an STR, but STRs that are more pure will lead to higher repeat scores; validation of detected STR expansions with PCR can be used to assess purity, as well as visual inspection of MPS reads with IGV. We only presented 26 case samples from 17 individuals, with at most 3
samples with the same expanded locus (HD and FRAXA). Due to lack of case samples we were unable to assess exSTRe’s performance in another 13 repeat expansion disorders. As larger sample sizes are required for a solid assessment of exSTRe, we plan to further test on larger, population-based cohorts at 30x coverage as well as additional samples from known repeat expansion loci. It is important to note that exSTRe is only designed as an initial screening tool, and all hits would require validation with current gold standard diagnostic methods.

Metcalf et al. (2017) demonstrated that fragile X syndrome, being the most common form of inherited intellectual and developmental disability, should undergo population carrier screening (with informed consent). Clearly, a more refined test for fragile X syndrome (FRAXA and FRAXE) is desirable from exSTRe, allowing screening in MPS cohorts. MPS would also allow the detection of rarer, non-repeat associated mutations causing fragile X syndrome such SNV nonsense mutations (Grønsvok et al., 2011), that are not included in standard screening.

In a recent exciting advancement, Batra et al. (2017) developed an RNA-targetting Cas9 system that would cause degradation of repeat expansion RNAs in vitro. This could be potential treatment for repeat expansion disorders that are caused by a protein or RNA with toxic gain-of-function effects, such as in myotonic dystrophy (DM1 and DM2), frontotemporal dementia and/or amyotrophic lateral sclerosis 1 (FTDALS1), and polyglutamine repeat expansion disorders such as Huntington disease and several spinocerebellar ataxias.

One of the ultimate goals of this study is to integrate repeat expansion detection into standard clinical MPS pipelines. Thousands of patients are tested with WGS each year in AGHA, MGHA, and in similar initiatives throughout the world. Including interrogation of STRs in their diagnostic pipelines alongside SNVs, indels and CNVs will increase diagnosis rates leading to better management and treatment.
Bibliography


Appendix A

Repeat score ECDFs of all 21 repeat loci for all cohorts

This appendix shows the empirical cumulative distribution function (ECDF) of the repeat scores given to each read for 21 repeat expansion loci, for each of the four MPS cohorts. The repeat score was defined as the number of bases covered by a repeat motif; this is described in detail in Section 2.1.1. Some of these ECDFs are presented in Chapter 3, Figures 3.33 to 3.36 and 3.38.
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.1. WES_PCR cohort repeat score ECDFs
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.2. WES_PCR cohort repeat score ECDFs
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.3. WES_PCR cohort repeat score ECDFs
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.4. WES_PCR cohort repeat score ECDFs
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.5. WGS_PCR_1 cohort repeat score ECDFs
Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.6. WGS_PCR_1 cohort repeat score ECDFs
Figure A.7. WGS_PCR_1 cohort repeat score ECDFs
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.8. WGS_PCR_1 cohort repeat score ECDFs
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.9. WGS_PCR_2 cohort repeat score ECDFs
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.10. WGS_PCR_2 cohort repeat score ECDFs
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

![Figure A.11. WGS_PCR_2 cohort repeat score ECDFs](image-url)
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.12. WGS_PCR_2 cohort repeat score ECDFs
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.13. WGS_PF_3 cohort (PCR-free) repeat score ECDFs
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

**Figure A.14.** WGS_PF_3 cohort (PCR-free) repeat score ECDFs

FRDA (intron_1 GAA) norm: 6 (20bp), exp: 200 (600bp) score ECDF

FTDALS (intron_1 GGGGCC) norm: 3 (23bp), exp: 250 (1500bp) score ECDF

HD (coding CAG) norm: 19 (59bp), exp: 36 (108bp) score ECDF

HDL2 (exon_variably_spliced CTG) norm: 14 (42bp), exp: 66 (198bp) score ECDF

SBMA (coding CAG) norm: 34 (103bp), exp: 38 (114bp) score ECDF

SCA1 (coding CAG) norm: 30 (91bp), exp: 39 (117bp) score ECDF
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.15. WGS_PF_3 cohort (PCR-free) repeat score ECDFs
Appendix A. Repeat score ECDFs of all 21 repeat loci for all cohorts

Figure A.16. WGS_PF_3 cohort (PCR-free) repeat score ECDFs
## Appendix B

Supplementary tables and figures for Chapter 4

**Table B.1. Estimated sex and call rates from BeadStudio**

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Illumina</th>
<th>Estimated Gender</th>
<th>Call Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-1</td>
<td></td>
<td>Female</td>
<td>99.37%</td>
</tr>
<tr>
<td>II-3</td>
<td></td>
<td>Female</td>
<td>99.30%</td>
</tr>
<tr>
<td>III-1</td>
<td></td>
<td>Female</td>
<td>99.26%</td>
</tr>
<tr>
<td>III-2</td>
<td></td>
<td>Female</td>
<td>99.40%</td>
</tr>
<tr>
<td>III-5</td>
<td></td>
<td>Male</td>
<td>99.93%</td>
</tr>
<tr>
<td>III-6</td>
<td></td>
<td>Male</td>
<td>99.92%</td>
</tr>
<tr>
<td>III-7</td>
<td></td>
<td>Male</td>
<td>99.87%</td>
</tr>
<tr>
<td>IV-1</td>
<td></td>
<td>Male</td>
<td>99.95%</td>
</tr>
<tr>
<td>IV-2</td>
<td></td>
<td>Male</td>
<td>99.83%</td>
</tr>
</tbody>
</table>

**Table B.2. Chi-squared values of SNP frequencies of samples**

Compared to HapMap3 populations. Generated by the bestPopTest option of LINKDATAGEN. Lower values indicate a closer match to the given population.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>CEU</th>
<th>ASW</th>
<th>CHB</th>
<th>CHD</th>
<th>GIH</th>
<th>JPT</th>
<th>LWK</th>
<th>MEX</th>
<th>MKK</th>
<th>TSI</th>
<th>YRI</th>
<th>BEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-1</td>
<td>6.9</td>
<td>397.1</td>
<td>333.5</td>
<td>357.4</td>
<td>54.5</td>
<td>296.8</td>
<td>551.5</td>
<td>57.7</td>
<td>373.8</td>
<td>11.2</td>
<td>500.7</td>
<td>CEU</td>
</tr>
<tr>
<td>II-3</td>
<td>5.7</td>
<td>443.4</td>
<td>361.7</td>
<td>330.8</td>
<td>74.0</td>
<td>329.1</td>
<td>543.7</td>
<td>99.6</td>
<td>341.0</td>
<td>9.1</td>
<td>608.2</td>
<td>CEU</td>
</tr>
<tr>
<td>III-1</td>
<td>0.8</td>
<td>377.7</td>
<td>310.4</td>
<td>282.9</td>
<td>50.4</td>
<td>312.9</td>
<td>526.9</td>
<td>28.1</td>
<td>296.2</td>
<td>2.2</td>
<td>537.6</td>
<td>CEU</td>
</tr>
<tr>
<td>III-2</td>
<td>0.2</td>
<td>344.5</td>
<td>319.7</td>
<td>363.8</td>
<td>44.0</td>
<td>331.7</td>
<td>503.2</td>
<td>56.9</td>
<td>285.0</td>
<td>4.1</td>
<td>491.9</td>
<td>CEU</td>
</tr>
<tr>
<td>III-5</td>
<td>3.6</td>
<td>414.9</td>
<td>260.7</td>
<td>287.0</td>
<td>41.7</td>
<td>283.8</td>
<td>551.3</td>
<td>24.0</td>
<td>336.9</td>
<td>0.0</td>
<td>528.2</td>
<td>TSI</td>
</tr>
<tr>
<td>III-6</td>
<td>2.5</td>
<td>311.2</td>
<td>287.1</td>
<td>286.9</td>
<td>38.9</td>
<td>355.4</td>
<td>456.9</td>
<td>32.7</td>
<td>290.5</td>
<td>0.8</td>
<td>476.5</td>
<td>TSI</td>
</tr>
<tr>
<td>III-7</td>
<td>0.5</td>
<td>399.5</td>
<td>342.1</td>
<td>288.4</td>
<td>55.7</td>
<td>288.7</td>
<td>513.4</td>
<td>44.1</td>
<td>377.6</td>
<td>0.3</td>
<td>562.0</td>
<td>TSI</td>
</tr>
<tr>
<td>IV-1</td>
<td>1.0</td>
<td>400.3</td>
<td>313.2</td>
<td>336.7</td>
<td>58.6</td>
<td>335.1</td>
<td>554.8</td>
<td>63.4</td>
<td>335.0</td>
<td>3.9</td>
<td>633.3</td>
<td>CEU</td>
</tr>
<tr>
<td>IV-2</td>
<td>1.1</td>
<td>355.7</td>
<td>340.3</td>
<td>274.6</td>
<td>47.0</td>
<td>325.7</td>
<td>496.6</td>
<td>47.9</td>
<td>347.3</td>
<td>0.8</td>
<td>537.6</td>
<td>TSI</td>
</tr>
</tbody>
</table>

**Number of SNPs**: 9671

**Mean heterozygosity**: 0.497

177
Table B.3. FEstim estimates

The inbreeding coefficient is F and IBD rate of change is A.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>F</th>
<th>F StdE</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-1</td>
<td>0</td>
<td>0</td>
<td>0.021</td>
</tr>
<tr>
<td>II-3</td>
<td>0</td>
<td>0</td>
<td>0.032</td>
</tr>
<tr>
<td>III-1</td>
<td>0</td>
<td>0</td>
<td>0.019</td>
</tr>
<tr>
<td>III-2</td>
<td>0</td>
<td>0</td>
<td>0.022</td>
</tr>
<tr>
<td>III-5</td>
<td>0</td>
<td>0</td>
<td>0.015</td>
</tr>
<tr>
<td>III-6</td>
<td>0</td>
<td>0</td>
<td>0.014</td>
</tr>
<tr>
<td>III-7</td>
<td>0</td>
<td>0</td>
<td>0.015</td>
</tr>
<tr>
<td>IV-1</td>
<td>0</td>
<td>0</td>
<td>0.013</td>
</tr>
<tr>
<td>IV-2</td>
<td>0</td>
<td>0</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Table B.4. WES alignment statistics

Each pair is counted as two reads. Duprem=reads remaining after duplicate removal.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>All</th>
<th>Aligned</th>
<th>% of all</th>
<th>Unique</th>
<th>% of all</th>
<th>Duprem</th>
<th>% of all</th>
<th>Unique</th>
<th>% of all</th>
<th>Total</th>
<th>% of all</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-3</td>
<td>142,151,822</td>
<td>141,133,056</td>
<td>99.28%</td>
<td>141,100,207</td>
<td>99.26%</td>
<td>133,380,692</td>
<td>93.83%</td>
<td>133,348,409</td>
<td>93.81%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>III-2</td>
<td>149,567,994</td>
<td>148,451,102</td>
<td>99.25%</td>
<td>148,415,373</td>
<td>99.23%</td>
<td>140,712,734</td>
<td>94.08%</td>
<td>140,677,627</td>
<td>94.06%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV-2</td>
<td>139,639,924</td>
<td>138,598,554</td>
<td>99.25%</td>
<td>138,561,546</td>
<td>99.23%</td>
<td>131,372,055</td>
<td>94.08%</td>
<td>131,335,624</td>
<td>94.05%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>431,359,740</td>
<td>428,182,712</td>
<td>99.26%</td>
<td>428,077,126</td>
<td>99.24%</td>
<td>405,465,481</td>
<td>94.00%</td>
<td>405,361,660</td>
<td>93.97%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B.5. WGS alignment statistics

Each pair is counted as two reads. Duprem=reads remaining after duplicate removal.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>All</th>
<th>Aligned</th>
<th>% of all</th>
<th>Unique</th>
<th>% of all</th>
<th>Duprem</th>
<th>% of all</th>
<th>Unique</th>
<th>% of all</th>
<th>Total</th>
<th>% of all</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-3</td>
<td>2,631,819,642</td>
<td>2,563,588,730</td>
<td>97%</td>
<td>2,557,416,462</td>
<td>97%</td>
<td>1,604,182,282</td>
<td>61%</td>
<td>1,600,013,709</td>
<td>61%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>III-2</td>
<td>2,644,311,696</td>
<td>2,593,031,162</td>
<td>98%</td>
<td>2,587,837,166</td>
<td>98%</td>
<td>1,441,490,697</td>
<td>55%</td>
<td>1,437,787,592</td>
<td>54%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV-2</td>
<td>1,024,678,520</td>
<td>1,014,927,818</td>
<td>99%</td>
<td>1,013,292,173</td>
<td>99%</td>
<td>978,875,386</td>
<td>96%</td>
<td>977,548,880</td>
<td>95%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6,300,809,858</td>
<td>6,171,547,710</td>
<td>98%</td>
<td>6,158,545,801</td>
<td>98%</td>
<td>4,024,558,365</td>
<td>64%</td>
<td>4,015,350,181</td>
<td>64%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B.6. Markers with matching genotypes in SNP chip data

Percentage of markers with matching genotypes in SNP chip data compared to whole-genome sequencing. Correctly matched samples should be close to 100%.

<table>
<thead>
<tr>
<th></th>
<th>WGS</th>
<th>SNP chip</th>
<th>II-3</th>
<th>III-2</th>
<th>IV-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-3</td>
<td>99.97%</td>
<td>63.04%</td>
<td>61.67%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>III-2</td>
<td>63.05%</td>
<td>99.97%</td>
<td>64.40%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV-2</td>
<td>61.65%</td>
<td>64.42%</td>
<td>99.96%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table B.7. WES and WGS samples matching at HapMap 2 locations

WES and WGS samples compared at HapMap 2 locations for matching genotypes. The orange cells give the results for matching samples from WES to WGS that we expect to be close to 100% if no sample swaps have occurred. This table includes comparisons for the same technology across different samples.

<table>
<thead>
<tr>
<th></th>
<th>WES</th>
<th>WGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-3</td>
<td>100.00%</td>
<td>75.66%</td>
</tr>
<tr>
<td>III-2</td>
<td>75.66%</td>
<td>100.00%</td>
</tr>
<tr>
<td>IV-2</td>
<td>74.87%</td>
<td>100.00%</td>
</tr>
<tr>
<td>II-3</td>
<td>99.96%</td>
<td>75.65%</td>
</tr>
<tr>
<td>III-2</td>
<td>75.65%</td>
<td>99.96%</td>
</tr>
<tr>
<td>IV-2</td>
<td>74.86%</td>
<td>76.75%</td>
</tr>
</tbody>
</table>