

Bioinformatics methods and approaches
to discover disease variants from DNA sequencing data

Harriet Dashnow

ORCID ID: <https://orcid.org/0000-0001-8433-6270>

Doctor of Philosophy

June 2019

School of BioSciences

The University of Melbourne

Bioinformatics

Murdoch Children's Research Institute

Submitted in total fulfilment of the requirements of the degree of
Doctor of Philosophy

Abstract

Next-generation sequencing is increasingly used to diagnose patients with suspected genetic disease. Yet, even after exome or whole genome sequencing, many patients remain undiagnosed. In many cases a genetic diagnosis is not made because we either failed to detect the causal variant, or succeeded in detecting it, but failed to identify it as causative. There is a clear need to develop novel bioinformatics methods and sequencing strategies to address these shortcomings and to increase diagnostic rates. In this thesis I develop several strategies to address these issues.

I propose a pooled-parent exome sequencing approach to prioritise *de novo* variants for genetic disease diagnosis. In this strategy, a set of probands have individual exome sequencing, while the DNA from all the parents of the probands are pooled, exome captured and sequenced together. The variants called in this pool are used to filter out inherited variants in the probands so the remaining list is enriched for *de novo* variants.

Short Tandem Repeat (STR) expansions are a class of disease-causing variants that are frequently missed in short read sequencing data. Here I develop and validate STRetch, a new bioinformatics method to detect STR expansions using STR decoy chromosomes. I show that STRetch can be used to detect both known pathogenic STR expansions, and novel expansions at other annotated STR loci across the genome. I further use STRetch to explore variation across hundreds of individuals to inform our understanding of what is common variation and what is potentially pathogenic, to aid in prioritising STR variants in a gene-discovery setting. Some of the methods that I have developed and describe within this thesis have already been used to help patients receive a genetic diagnosis.

Declaration

This is to certify that

- i. the thesis comprises only my original work towards the PhD except where indicated in the Preface,
- ii. due acknowledgement has been made in the text to all other material used, and
- iii. the thesis is fewer 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed,

A handwritten signature in black ink, appearing to read 'Harriet', written in a cursive style.

Harriet Dashnow

Preface

Collaborations

I am the primary author of all work presented in this thesis. The work presented in chapters 2-4 was done in collaboration with others, specifically:

Chapter 2: Alicia Oshlack, conceived the pooled-parent concept and guided the project. I wrote the software, performed the data analysis and drafted the manuscript, with help from Alicia. I estimate that I contributed 80% of the work for this project. Katrina M. Bell helped with data analysis. Zornitza Stark, Tiong Y. Tan and Susan M. White clinically evaluated and obtained consent from research participants, and provided data.

Chapter 3: I conceived the concept for the STRetch algorithm, wrote the software, performed the majority of the analyses and wrote the manuscript. I estimate that I contributed 80% of the work for this project. Alicia Oshlack advised on the development of STRetch algorithm and helped draft the manuscript. Belinda Phipson and Simon Sadedin contributed to algorithm development. Andreas Halman helped perform data analysis. Andreas, Simon and Andrew Lonsdale helped with software testing. Mark Davis, Nigel G. Laing and Joshua S. Clayton performed experimental validation. Nigel and Phillipa Lamont contributed patient samples. Monkol Lek and Daniel G. MacArthur provided reference samples. Daniel provided sequencing and computational resources. Monkol, Belinda, Andreas, Andrew, Mark, Joshua, Nigel, Daniel and Alicia contributed to the manuscript.

Chapter 4: I planned and performed all the data analyses and wrote the chapter. I estimate that I contributed 90% of the work for this project. Alicia Oshlack advised on analysis and the chapter. Belinda Phipson advised on statistical analyses. Simon Sadedin assisted with running STRetch in the cloud

environment, as well as developing the Bazam tool, which greatly reduced the runtime required for this analysis. Daniel G. MacArthur provided sequencing data and computational resources.

Publication status

Chapters 1, 4 and 5: Unpublished material not submitted for publication.

Chapter 2: Published by bioRxiv on 7 April 2019 as a non-peer-reviewed preprint and submitted for publication.

Dashnow H, Bell KM, Stark Z, Tan TY, White SM, Oshlack A. Pooled-parent exome sequencing to prioritise de novo variants in genetic disease [Internet]. bioRxiv. 2019. Available from:

<https://www.biorxiv.org/content/10.1101/601740v1>

Chapter 3: Published by Genome Biology on 21 August 2018.

Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STretch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biol [Internet]. BioMed Central; 2018 [cited 2019 May 17];19:121. Available from:

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1505-2>

Funding

Harriet was supported by an Australian Government Research Training Program (RTP) Scholarship, a Murdoch Children's Research Institute Top Up Scholarship, and an Australian Genomics Health Alliance PhD Award, funded by NHMRC (GNT1113531).

Publications related to this PhD:

The following are publications to which I contributed that directly relate to this PhD thesis. Where I am the primary author they are included as chapters. Where I am a contributing author they are mentioned in the relevant sections.

Dashnow H, Bell KM, Stark Z, Tan TY, White SM, Oshlack A. Pooled-parent exome sequencing to prioritise de novo variants in genetic disease [Internet]. bioRxiv. 2019. Available from:

<https://www.biorxiv.org/content/10.1101/601740v1>

Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biol [Internet]. BioMed Central; 2018 [cited 2019 May 17];19:121. Available from:

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1505-2>

Sadedin SPSP, Dashnow H, James PAPA, Bahlo M, Bauer DCDC, Lonie A, et al. Cpipe: a shared variant detection pipeline designed for diagnostic settings. Genome Med [Internet]. BioMed Central; 2015;7:68. Available from:

<http://dx.doi.org/10.1186/s13073-015-0191-x>

Stark Z, Dashnow H, Lunke S, Tan TYTY, Yeung A, Sadedin S, et al. A clinically driven variant prioritization framework outperforms purely computational approaches for the diagnostic analysis of singleton WES data. Eur J Hum Genet [Internet]. Nature Publishing Group; 2017 [cited 2019 May 24];25:1268–72. Available from:

<http://www.nature.com/articles/ejhg2017123>

Dashnow H, Tan S, Das D, Eastal S, Oshlack A. Genotyping microsatellites in next-generation sequencing data. BMC Bioinformatics [Internet]. BioMed Central Ltd; 2015 [cited 2019 May 24];16:A5. Available from: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S2-A5>

Publications that were published during this PhD but are not related:

I contributed to the following additional publications during my PhD, however their topics do not directly relate to this thesis, so they are not mentioned in the main text.

Lonsdale A, Penington JS, Rice T, Walker M, Dashnow H, Sietsma Penington J, et al. Ten simple rules for a bioinformatics journal club. PLoS Comput Biol. Public Library of Science; 2016;12:e1004526.

Nunez-Iglesias J, van der Walt S, Dashnow H. Elegant SciPy. Blanchette M, editor. O'Reilly;

Acknowledgements

In writing this, I am overwhelmed by the love and support that made this PhD thesis possible. There are a few people I would particularly like to thank.

First and foremost, I would like to thank my PhD supervisor, Alicia Oshlack. I often reflect on how lucky I was to find you through the tenuous links of chance and connections (although Matthew Wakefield deserves credit here, for singing your praises at a Lorne poster session). You are an absolutely incredible supervisor, mentor and friend. You welcomed me to the world of bioinformatics and inspired me with your passion for science. You lead by example and with force of personality, showing the world that women can be both brilliant scientists and great parents, and you do this while speaking out for and boosting up other women in science. I loved our long chats about everything from science, to careers and parenting. You always made time for me, encouraged me, and pushed me towards independence. I'm going to miss working with you every day, but I'm sure we have so much more science to do together in the future.

I would like to thank everyone in the Bioinformatics group at the Murdoch Children's Research Institute. You are a wonderful bunch of people to work with. I looked forward to coming in every day because of you, and I especially enjoyed our coffee walks and lunch debates. In particular, "Oshlack's Angels": Belinda Phipson, Jovana Maksimovic and Nadia Davidson. You are brilliant scientists and wonderful mothers. Thank you for being there for me through some of the most difficult times in my life. To Simon Sadedin, thank you for answering all of my incessant technical questions and for your genomics expertise. Thank you to Katrina M. Bell for your advice on clinical genomics analysis. Thank you, Belinda Phipson, for

sharing your statistical knowledge. Thank you, Andreas Halman, for your help with STR analyses.

Andrew Lonsdale, thank you for your friendship and support these past years. Thank you for all the writing sessions and debriefs and for your many contributions to making the Australian Bioinformatics community. I truly value your endless enthusiasm for side-projects and have loved working with you on everything from COMBINE and Journal Club to Ten Simple Rules (and countless other side-projects I probably shouldn't enumerate here in case our supervisors are reading!). I'm glad we finally got to write a scientific paper together and I look forward to many more side-projects with you.

Thank you to all the friends from our Bioinformatics student community that I made along the way: Jane Hawkey, Tim Rice, Scott Ritchie, Zoe Dyson, Luke Zappier, Joycelyn Pennington, and Danielle Ingle, to name just a few. Also thank you to the wonderful bioinformaticians and computational biologist I have worked with over these past few years: Juan Nunez-Iglesias, Clare Sloggett, Simon Gladman, Matthew Wakefield, Bernie Pope, and Maria Doyle. Thank you to Andrew Lonie for developing and running the University of Melbourne Bioinformatics Master's program, and for being a great boss and mentor.

I would also like to thank my co-supervisors, past and present. Thank you to Simon Eastal for introducing me to the wonderful world of short tandem repeats, for giving me my first bioinformatics summer project, and for co-supervising my Masters research. Thank you to Kathryn Holt for introducing me to microbial genomics, and for supervising my first bioinformatics software and publication, SRST2. Thank you to John Christodoulou for co-supervising my PhD research.

I had some wonderful collaborators during my PhD. I've mentioned all off your contributions in the relevant chapters. But I'd like to make a few special

mentions. First, Daniel G. MacArthur, thank you for hosting me in your lab in my first year of this PhD. This led to a wonderful collaboration and you have really made much of this research possible by giving me access to sequencing data and computational resources. Thank you also for your support and advice these past years. You have been a pleasure to work with. Nigel G. Laing, thank you for your unwavering enthusiasm and support for the STRetch project. Thank you also for introducing me to Mark Davis, who introduced me to STR diagnostics and validation. There are several other collaborators to whom I am also particularly grateful: Zornitza Stark, Susan M. White, and Monkol Lek.

Bethany Kairys, thank you for your unwavering and honest friendship, for always being there with excellent advice, or to just listen to me rant. Also, thank you for your support and your grammatical expertise.

To my parents, Nadine Cresswell-Myatt and Malcom Walter Gorman, thank you for raising me to believe that I could do anything I set my mind to and filling my childhood with books. Thank you for your endless encouragement, time, and love. Mum, thank you for making sure I had every opportunity to follow my passions. Thank you for inspiring me to write. Dad, thank you for encouraging me to program and for your feedback on this thesis.

To my wonderful, supportive husband, Craig Steven Dashnow. Thank you for all the sacrifices you have made to help me follow my dreams. Thank you for being my cheering squad and my safe harbour. I'm so grateful to have found you. I love you, always.

To my not-so-little-anymore baby, Arthur Henry Gorman Dashnow. This thesis is dedicated to you. My heart bursts with love for you every day.

Table of Contents

Abstract	2
Declaration.....	3
Preface	4
Acknowledgements	8
Table of Contents	11
List of tables	14
List of figures.....	16
Chapter 1 Introduction	19
1.1 The Human genome.....	20
1.1.1 The human reference genome.....	20
1.1.2 Types of genetic variation.....	21
1.1.3 Gene structure and regulation.....	22
1.2 DNA sequencing technologies	23
1.2.1 Short read sequencing.....	24
1.2.2 Long read sequencing.....	25
1.2.3 Sequencing strategies	26
1.3 Clinical genomics and genetic disease research.....	27
1.3.1 Genetic disease	28
1.3.2 Making a genetic diagnosis.....	30
1.3.3 Analysis pipeline: from sequencing reads to disease variant	32
1.4 Short Tandem Repeats	37
1.4.1 STR mutation mechanism.....	39

1.4.2	STR variation impacts gene expression	41
1.4.3	STRs in Human Disease	42
1.4.4	Genotyping STRs by length.....	46
1.4.5	Sequencing STRs	49
1.4.6	Algorithms for genotyping STRs in high-throughput sequencing data	52
1.5	Aims of this thesis.....	56
Chapter 2 Pooled-Parent Exome Sequencing to Prioritise <i>de Novo</i> Variants in Genetic Disease		58
2.1	Abstract.....	59
2.2	Background	60
2.3	Results.....	63
2.3.1	Simulation set up	63
2.3.2	Choice of variant caller and sequencing depth	64
2.3.3	Calling variants from real pools	69
2.4	Discussion	73
2.5	Conclusions.....	75
2.6	Methods.....	75
2.6.1	Simulating pools	75
2.6.2	Variant calling and analysis	77
2.6.3	Samples and sequencing.....	78
2.6.4	Analysis of real pools.....	79
Chapter 3 STRetch: detecting and discovering pathogenic short tandem repeat expansions.....		80
Chapter 4 Frequency and variability of STR expansions.....		94

4.1	Introduction	95
4.2	STR loci in the human reference genome	99
4.2.1	Annotating STR loci	100
4.3	Counting reads assigned to the STR decoy chromosomes.....	102
4.4	Frequency and patterns of STR expansion	108
4.5	STR constraint and prioritising potentially pathogenic expansions	111
4.5.1	STR expansions in set of unaffected individuals.....	116
4.6	The impact of sample choice in STRetch control sets.....	123
4.7	Discussion	132
Chapter 5	Conclusion.....	136
References	143
Chapter 6	Appendices	169

List of tables

Table 1.1: Summary of human inherited diseases caused by tandem repeat expansions.....	43
Table 1.2: Summary of selected STR genotyping methods.....	53
Table 4.1: The top ten STR decoy chromosomes ranked by the number of reads aligned to them across all 362 samples.	104
Table 4.2: STR loci with no STR decoy reads assigned in any sample.....	115
Table 4.3: Number of significant STRetch calls within known pathogenic STR loci across all 125 unaffected samples at $p < 0.05$ and $p < 0.01$	119
Table 4.4: Number of the 59 highly variable samples with significant expansions ($p < 0.01$) for each pathogenic STRs using each of the controls sets. Loci with no significant results are excluded.	127
Table 4.5: Number of unaffected samples with significant expansions ($p < 0.01$) for each pathogenic STRs using each of the controls sets.	130
Table 6.1: Mean recall rate (over three replicates) for all variants as a percentage in simulated pools of two, four, six, eight and ten individuals. ...	170
Table 6.2: Mean recall rate (over three replicates) for singleton variants as a percentage in simulated pools of two, four, six, eight and ten individuals. ...	170
Table 6.3: Mean variants called per pool (over three replicates) in simulated pools of two, four, six, eight and ten individuals.....	171
Table 6.4: Mean false positive rate (over three replicates) as a percentage in simulated pools of two, four, six, eight and ten individuals.....	171
Table 6.5: Mean variants called per individual (diploid) sample used in the simulations for each variant caller.....	172
Table 6.6: Ethnicity of probands, as identified by the patient's family.	172
Table 6.7: SRA run IDs for the 111 exome samples used in pooling simulations.	173

Table 6.8: List of disease genes excluded from the real pooled parents analysis.	174
Table 6.9: The proportion of the hg19 genome covered by various genomic features based on the gencode v19 annotation.	186
Table 6.10: The top ten STR decoy chromosomes ranked by the number of reads aligned to them across all the 303 least variable samples.	187
Table 6.11: Summary of normal, intermediate and pathogenic allele size ranges for known pathogenic loci for which expansions were identified in the unaffected samples.	188

List of figures

Figure 1.1: Typical stages in a clinical genomics pipeline.....	33
Figure 1.2: STR structure and nomenclature using a dinucleotide AC motif as an example.	39
Figure 1.3: Two mechanisms for tandem repeat expansions and contractions theorised to occur within cells.	40
Figure 1.4: During DNA replication, a number of repeat units can transiently dissociate and mis-pair, resulting in deletions or additions of repeat units in the resulting fragments (source: Gemayel et al., 2010). .. Error! Bookmark not defined.	
Figure 1.5: ‘Stutter’ profiles of CA STR amplicons.....	48
Figure 1.6: Size ranges in base pairs of normal, intermediate and pathogenic alleles in STR expansion diseases.	51
Figure 2.1: Simulated pools of two, four, six, eight and ten individuals.	67
Figure 2.2: Overview of variants called in the real pooled parents experiment.	72
Figure 4.1: STR loci counts from the Tandem Repeats Finder annotation of hg19.....	100
Figure 4.2: Percentage of the STR loci in each repeat unit category that overlap a given genomic feature.	101
Figure 4.3: MDS plots of STR decoy counts.....	108
Figure 4.4: STRetch results per locus across all samples.....	109
Figure 4.5: STRetch results per sample.	110
Figure 4.6: STRetch results per locus across the 303 less variables samples.	111
Figure 4.7: Huber’s estimates of median and variance across all samples for each STR locus.	113

Figure 4.8: The Huber’s variance of the log normalised STR locus counts were divided into ten approximately even deciles (x-axis: decile of variance)..... 115

Figure 4.9: Huber’s estimates of median and variance across the unaffected samples for each STR locus. 117

Figure 4.10: STRetch results for all pathogenic loci with results in the 125 unaffected samples. 121

Figure 4.11: Comparing the number of significant ($p < 0.01$) STR loci across the entire set of samples and three different subsets: the previously identified “highvar” samples with more than 10,000 STR loci with any reads, the “lowvar” samples with less than 10,000 STR loci with any reads and the “unaffected” samples. 124

Figure 4.12: Comparing the number of significant STR loci ($p < 0.01$) called in the “highvar”, highly variable samples with more than 10,000 STR loci, using four different control sets: “all” samples, “highvar”, “lowvar” (less than 10,000 STR loci with any reads) and the “unaffected” samples..... 125

Figure 4.13: STRetch p values for all pathogenic loci with results in the 59 highly variable samples using each of the controls sets..... 128

Figure 4.14: Comparing the number of significant STR loci ($p < 0.01$) called in the unaffected samples, using four different control sets: “all” samples, “highvar”, “lowvar” and the “unaffected” samples. 129

Figure 4.15: STRetch p values for all pathogenic loci with results in the 125 unaffected samples using each of the controls sets..... 131

Figure 6.1: Recall for all constant depth simulations from Figure 1, with * to indicate recall for all variants in the real pools (percentage of variants found in all probands that were also detected in the pool). 169

Figure 6.2: Number of significant STR loci at $p < 0.01$ called by STRetch vs. median sequencing depth across the genome for each sample shows not increased rate of significant calls in more deeply sequenced samples. 189

Figure 6.3: Variance vs. Huber's robust variance estimate of log normalised locuscounts for all samples..... 189

Chapter 1 Introduction

Genetic disease is one of the leading causes of death in children [1]. Some genetic diseases have well-characterised causal DNA variants, allowing relatively quick and confident genetic diagnoses. Other disorders may have a poorly understood genetic basis, or symptoms consistent with variants in multiple genes, making them difficult to diagnose using single-gene tests. In the clinical setting, next generation DNA sequencing has become a key tool in the diagnosis of rare genetic disorders. It has become relatively routine to order panel or exome sequencing for patients with a suspected genetic disorder, and increasingly whole genome sequencing is being utilised, especially for cases that are not solved by first line-testing. Yet even with exome or whole genome sequencing, many cases remain unsolved. With diagnostic rates ranging from 31-58% [2,3], there are still many families who do not receive a definitive genetic diagnosis. Even for those patients who do receive a genetic diagnosis, the diagnostic odyssey can be long, frequently lasting months, and in many cases, years.

For many unsolved cases of suspected genetic disease, we suspect that the causal variant has been sequenced, yet we lack either the ability to detect or interpret it. Most clinical sequencing pipelines only consider single nucleotide and short insertion/deletions variants. In contrast, large or repetitive DNA variants are difficult to accurately detect in short read data, and so are not routinely genotyped. For example, Short Tandem Repeats (STRs) are not routinely genotyped in high-throughput sequencing. In fact, until quite recently, there were no bioinformatic methods available to detect STR expansions in short-read sequencing data. This is a critical limitation for the diagnosis of diseases that are known to be caused by either STR expansions or short variants. For example, spinocerebellar ataxia, and associated conditions

are known to be caused by expansions of STRs, SNPs and indels in a number of genes [4].

Even if the causal variant is correctly genotyped, it can be difficult to assign clinical significance to it. The sheer volume of data can be an impediment to finding the causal genetic variant. A given exome sequencing test can generate tens of thousands of variants, which must then be filtered to find those variants most likely to cause disease. Then for some disease variants it can be difficult to clearly link them to the specific disease. They may be in a gene that has not previously been linked to that disease, or a new variant in a gene already known to cause disease.

There is a clear need to develop new bioinformatics methods and sequencing strategies to aid clinicians in making genetic diagnoses and to support the discovery of new disease variants. In this thesis I describe the development of sequencing strategies and bioinformatic methods to support the diagnosis of genetic disease. First, in Chapter 1 I will summarise the current state of clinical genomics, with an emphasis on bioinformatics methods and the detection of STRs. In Chapter 2, I describe a new pooled parent sequencing strategy to aid the diagnosis of individuals with likely *de novo* genetic disease. I describe the considerations with regard to sequencing method as well as bioinformatic analysis of the resulting data. In Chapter 3 I describe STRetch, a new bioinformatic method to detect STR expansions from short-read sequencing data. Finally, in Chapter 4 I apply STRetch to the analysis of a larger population, and describe how this population information can be used to prioritise potentially pathogenic STR expansions.

1.1 The Human genome

1.1.1 The human reference genome

The human reference genome is central to modern genomics. It is the standard against which all individuals are compared and the coordinate system

for communicating sequences and variants. As such, the human reference genome is an incredibly rich resource for the study of human genetics and diversity. Our heavy reliance on the human reference genome also makes it a potential source of bias. Although the original human reference genome included sequences from several donors, it disproportionately represents the genome of just one male individual [5]. It primarily represents European ancestry, making it more difficult to study other populations [6]. Even within Europeans it has limited representation of alternate alleles present in the population, despite the inclusion of alternative haplotypes (e.g. HLA) in the most recent version hg38 [7]. Due to limitations in sequencing technology and genome assembly, the original reference genome concentrated on the euchromatic, less repetitive portions of the genome, with difficult regions such as centromeres included in only the most recent version [8]. It is known to under-represent repetitive regions such as STRs [9].

1.1.2 Types of genetic variation

Single Nucleotide Variants (SNVs) are single DNA base substitutions, for example C to T, where the total length of the DNA fragment does not change. These are often referred to as Single Nucleotide Polymorphisms (SNPs), although this name is not accurate for all single base changes because not all are polymorphic (generally defined as occurring at a frequency of greater than 0.01% in the population [10]). Insertion or deletion variants (indels) are the addition or loss of one or more consecutive DNA bases, for example ATG to AG. In contrast to SNVs, indels change the length of the DNA fragment. Because of their small size and relative simplicity SNVs and short indels are the most commonly characterised DNA change.

Repetitive DNA sequences are often overlooked in sequencing studies because reads arising from them are difficult to uniquely assign to the genome. These sequences include short tandem repeats, minisatellites, larger duplications (e.g. of exons or whole genes), pseudogenes, transposable

elements (e.g. LINEs SINEs Alu), telomeres and centromeres. Short Tandem Repeats (STRs or microsatellites) are short 1-6 bp sequences repeated back to back. As a major focus of this thesis, the biology, clinical importance and genotyping of STRs is explored in greater detail in section 1.4.

Many larger structural variants from copy number variants, large deletions, translocations and whole-chromosome aneuploidy have also proven difficult to identify from sequencing data, although, perhaps due to the increasing prevalence of whole genome sequencing data, the bioinformatics software in these areas is now moving into maturity [11].

1.1.3 Gene structure and regulation

In eukaryotic protein-coding genes DNA is transcribed into unprocessed pre-mRNA. The transcribed region is defined by the Transcription Start Site (TSS) and transcription terminator sequence. The gene sequence also includes some 5' (upstream) and 3' (downstream) regulatory sequences which include the gene promoter as well as enhancer and silencer sequences, which all contribute to the rate at which RNA is transcribed from that gene. After transcription pre-mRNA is then processed to remove introns (a process called splicing), add a 5' cap and a poly-A tail. Specific sequence motifs define the splice donor (GU), acceptor (AG) and branch (A) sites as well as some more variable motifs that guide splicing [12]. The remaining sequences in the mRNA are called exons. The mRNA is then translated: each group of three nucleotides (a codon) codes for a single amino acid, although each amino can be coded for by multiple codons. Together these amino acids are the basic building blocks of proteins. A specific start codon (AUG) defines the first amino acid while any of three stop codons (UAG, UAA, UGA) will end translation. The regions that are ultimately translated are referred to as coding sequences. The regions of the mRNA upstream and downstream of the coding regions are referred to as the UnTranslated Regions (UTRs).

In addition to protein-coding genes, eukaryotic genomes contain many sequences that are transcribed into RNA but not translated into protein. These non-coding genes include essential sequences to produce ribosomal RNA and transfer RNA (key components in the translation process) as well as microRNAs and long non-coding RNAs which have regulatory roles [13].

Understanding the structure of a gene helps us to identify those DNA variants that are more likely impact the function of a gene and potentially lead to a disease phenotype. For example, a SNV in a coding region could cause a change in the codon causing a substitution to a different amino acid (a missense variant), or even obliterate a start or stop codon. Because of redundancy in the assignment of codons to amino acids, a synonymous variant may change the codon but still result in the same amino acid. An indel in a coding region that is not a multiple of three causes a frameshift, that is it changes the reading frame so that all downstream codons are changed.

Changes to the key splicing sequences can cause an intron to be retained or spliced at a different position, making dramatic changes to the protein sequence. Changes to regulatory sequences may be more subtle. For example, they may reduce the affinity of the promoter to the translational machinery and so reduce the rate of gene expression. Ensembl publishes an order of variant consequence severity which categorises variants that cause splicing changes, premature stop codons and frameshifts as amongst the most deleterious, while synonymous and regulatory changes are considered less serious for protein function [14].

1.2 DNA sequencing technologies

At the foundation of modern bioinformatics lies the explosion in DNA sequencing technology. The first major DNA sequencing method was the ‘chain-termination’ or dideoxy technique, developed in 1977 by Frederick Sanger and colleagues and now referred to as “Sanger sequencing” [15]. In the

mid-2000's the first high-throughput short read sequencing machines entered the market. These were dubbed next-generation or second-generation sequencing technologies (Sanger sequencing being the first generation). Sequencing became highly automated and parallelised, resulting in dramatic increases in sequencing output and reductions in the per base pair cost of sequencing over the following decade [16]. After heavy competition, Illumina emerged as the dominant provider of short read sequencing technologies [16]. However, in recent years we have experienced the third wave of DNA sequencing technology, with the development of single-molecule long-read technologies, now marketed by PacBio and Oxford Nanopore. There are a number of considerations when choosing a sequencing technology to use, in particular, read length and cost. I will describe Illumina, PacBio and Oxford Nanopore in more detail, as the technologies most relevant to this thesis.

1.2.1 Short read sequencing

As mentioned, short read sequencing is now dominated by the Illumina sequencing by synthesis (SBS) technology. Briefly, purified DNA is fragmented into smaller pieces. For single-end data, only one end of this fragment is sequenced, while for the (more common) paired-end sequencing strategy a short segment of each of the ends is sequenced. Two reads arising from the same DNA fragment are referred to as a read pair, or pair of reads. The size of the physical DNA fragments is often estimated by calculating the insert size, that is the distance between a pair of reads aligned to the reference genome. The read length of the data is the number of DNA base pairs (bp) that are sequenced (either at one or both ends of the fragment). Most modern Illumina sequencing data described in this thesis has read lengths of 150 bp and insert sizes of approximately 400-500 bp (although there is a large amount of variation within an experiment).

To describe the sequencing process in more detail: adaptors are added to both ends of the fragment, including sequences that allow primer binding,

identification of samples and allow the fragment to bind to short nucleotide sequences attached to the flow cell. Depending on the library preparation type, these fragments may undergo enrichment and/or PCR prior to sequencing (see section 0). Once attached to the flow cell, each DNA fragment is amplified by bridge PCR to produce a cluster of identical fragments. These clusters are then sequenced using the SBS strategy. As each base is added by a DNA polymerase, the base emits a fluorescent signal, with a colour corresponding to the specific A, T, C or G base added.

While the error rate of Illumina sequencing is generally low at $\sim 0.1\%$, it suffers from higher error rates at specific loci such as homopolymer runs and regions of extreme GC [17–19]. Homopolymers are a sequence of identical DNA bases. Using SBS these identical bases all have the same fluorescent signal, so the machine can have difficulty determining how many identical bases were added at a time. Illumina sequencing is ultimately a PCR-based method: even when using PCR-free library preparations the actual sequencing involves a DNA polymerase and bridge-amplification. Therefore, it suffers from the known GC biases arising from PCR. That is regions of extreme GC content tend to be under-represented in the sequencing results, so that these regions tend to have lower coverage [18]. The use of PCR-free library preparations does help to mitigate both GC and homopolymer biases [20,21].

1.2.2 Long read sequencing

Recent advances in sequencing technology have seen a third generation of sequencing technologies enter the market, with dramatically longer read lengths than those seen previously. The two main technologies are the Single-molecule real-time sequencing (SMRT) by Pacific Biosciences (PacBio) and nanopore sequencing by Oxford Nanopore Technologies (ONT), which I will refer to as PacBio and Nanopore sequencing respectively. Both methods work by sequencing individual DNA molecules, in contrast to the amplification-based methods used by Illumina and some other second-generation

technologies. PacBio has average read length of 10-15 kb, with reads greater than 80 kb reported, while ONT claims read lengths are limited by the length of the DNA fragment [22].

Both long read methods also suffer from considerably higher error rates than Illumina. PacBio has a single-pass error rate of $\sim 13\%$, however shorter fragments up to $\sim 1-2$ kb can be sequenced multiple times to generate a consensus read with much higher accuracy [23]. ONT reports an error rate of approximately 15% [24], falling to $\sim 3\%$ if both strands are sequenced [22].

Despite their high error rates, the long-read lengths generated by third generation technologies make them ideal for detecting larger variants such as structural re-arrangements. In addition, these methods are particularly useful for resolving repetitive sequences such as STRs, gene duplications and distinguishing pseudogenes. The relative expense, lower throughput and high error rates of long read technologies have largely prevented their adoption by the clinical genomics community thus far, however there have been some forays into using long reads for targeted sequencing of difficult to resolve regions such as HLA and STR loci [25].

1.2.3 Sequencing strategies

In concert with the explosion in DNA sequencing technology, the scientific community has developed a variety of strategies to make efficient use of sequencing. In particular, whole exome and other targeted sequencing strategies have had a huge impact on clinical genomics [26]. Rather than sequencing the entire human genome, hybridisation probes are designed to capture and enrich for DNA sequences of interest, which are then PCR-amplified. This can result in dramatic cost-savings, by only sequencing the desired regions. For whole exome sequencing the probes are designed to target most known protein-coding regions of the genome, or in some cases

selected UTRs as well. Smaller target regions are also relatively common, for example cardiac gene panels.

Another advance in the Illumina sequencing library protocol is the introduction of “PCR-free” library preparation. Where sufficient input DNA is available, sequencing can be performed without first amplifying the library. This reduces PCR induced biases such as GC bias and homopolymer errors [20].

1.3 Clinical genomics and genetic disease research

The increasing use of high-throughput DNA sequencing technologies in the clinical setting has given rise to the field of clinical genomics. In many hospitals, patients with rare, likely genetic conditions are now routinely being offered testing with next-generation sequencing. This has dramatically expanded our ability to diagnose a wide range of single-gene disorders [26]. Disease gene panel sequencing and exome sequencing are the most common approaches, with whole genome sequencing still primarily being used only in cases where initial testing fails to make a genetic diagnosis.

The needs of the clinical diagnostic community differ critically from genetic disease research in a number of ways. In the clinical setting time to diagnosis is crucial, with test results typically returned in the scale of weeks, or even days [27]. In contrast, researchers may spend years validating a novel disease gene. In the clinical setting, a variant will only be accepted as a genetic diagnosis if it meets strict criteria. For example the American College of Medical Genetics (ACMG) has published recommended standards and guidelines interpreting sequence variants, which describe the commonly used set of criteria [28].

More generally we can think of clinical genomics as a process of reaching a specific diagnosis for a patient by detecting one or more high-confidence variants in an established disease gene. In contrast, in the genetic disease research context, we are often looking for novel variants in genes that have

not previously been linked to that disease. There is a continuum between the clinical and research contexts. Patients who fail to receive a genetic diagnosis in the clinical context may be moved into a research context. Conversely, new methods, and evidence that specific variants and genes are responsible for disease, are constantly being developed in the research context and translated into clinical use.

Much of the work in this thesis straddles the divide between research and clinical genomics. For example, the pooled parent exome sequencing approach described in Chapter 2 is designed to be quickly translated into clinical use, as it integrates well with the strategies already used in clinical practice. The STRetch algorithm described in Chapter 3 is able to detect known, high confidence, pathogenic STR expansions, and so could be integrated directly into clinical pipelines. However, it is also able to detect STR expansions at all STR loci across the genome, making it suitable for disease gene discovery in a research context. Gene discovery is also the emphasis of Chapter 4, as I consider how we can prioritise STR expansions that may be more likely to be pathogenic.

1.3.1 Genetic disease

This thesis focuses on the genomics of rare genetic disease. These diseases are typically caused by a single genetic variant in a single gene, which has a large impact on the phenotype of the affected individual. Many of these are inherited, termed Mendelian diseases. Well known examples include Cystic Fibrosis, Beta Thalassemia and Huntington's disease. In contrast, complex genetic diseases can be caused by a larger number of genetic loci, each contributing a relatively small risk of disease, for example type II diabetes.

1.3.1.1 Inheritance patterns

Genetic diseases typically follow one of a small set of inheritance patterns. In the autosomal dominant inheritance pattern, only one copy of the disease

allele is required to cause disease, so families show a pattern of affected individuals having at least one affected parent, and the disease does not skip generations. For an autosomal recessive disease, two copies of the disease allele, one inherited from each parent, are required to cause disease. These can be the same variant, often found in consanguineous families, or two different variants in the same gene (such individuals are termed compound heterozygotes). In recessive inheritance an affected individual may have two unaffected parents who are carriers, as they are heterozygous for the causal variant. Both these cases are referred to as autosomal because the causal genetic variant is carried on one of the autosomes (chromosomes 1-22) not the sex chromosomes (X, Y) or in the mitochondrial genome.

Sex-linked inheritance occurs when the disease variant is located on one of the sex or mitochondrial chromosomes, typically the X chromosome. X-linked dominant disease is passed from an affected parent to their daughter, but males can only inherit from their mother because they do not receive an X chromosome from their father. X-linked recessive disorders are more frequently seen in males because they have only one X chromosome, termed hemizygous. X-linked recessive disease is also possible in females, however they would have to have inherited from an affected father, while their mother may be a carrier. There are Y-linked diseases, however due to the relatively small size of the Y chromosome, there are few of these. Mitochondrial genetic disorders are generally inherited from mothers, as the egg contributes most of the mitochondria to the zygote.

Finally, *de novo* genetic diseases are those caused by a novel genetic variant in the proband that is not inherited from the parents. These are often caused by DNA mutations in the sperm or egg, or in the early stages of embryonic development. *Do novo* diseases are often dominant, and can be autosomal or sex-linked. Although *de novo* variants are rare, they contribute disproportionately to the number of diagnosed cases, for example one study of monogenic

disease in infants found 36% of their diagnosed variants were *de novo*, compared to 13% dominant and 47% recessive [3]. Genetic mosaicism may complicate the detection of *de novo* variants. Mosaicism occurs when a mutation occurs during embryonic development, resulting in an individual with cell populations harbouring different genotypes at a given locus. An unaffected parent may be mosaic for a causal variant such that their gonads harbour the variant, while their blood does not. This could lead to the appearance that a variant is *de novo*, when it was in fact inherited. It is estimated that ~10% of apparent *de novo* variants may in fact be the result of post-zygotic mosaicism [29]. Mosaicism can also lead to phenotypic differences, depending on which tissues contain the disease variant.

1.3.2 Making a genetic diagnosis

Making a genetic diagnosis means finding one or more DNA variants with a high confidence of causing the condition. A genetic diagnosis is valuable for a number of reasons. It can inform treatment, for example several single-gene disorders can be treated with available medications [30]. In addition, a genetic diagnosis can critically inform prognosis. This may inform management, give patients and their families a clearer expectation of disease progression and in some cases prevent the application of unnecessary invasive treatments. Many families also report that having a clear disease or gene name can allow them to find other families with the same condition, and form support networks. Finally, understanding the genetic basis of the disease may allow the family to make informed decisions about conceiving future children. Genetic variants may be inherited, or arise *de novo* in the patient. Knowing the inheritance pattern allows clinicians to counsel families on the risk to other family members and future children, and may allow preimplantation or prenatal testing for the specific DNA variant.

In children, genetic diagnostic rates from exome sequencing are typically in the range 31-39% [2]. These studies included many patients who were

sequenced only after low throughput testing failed to make a diagnosis, so tend to be enriched for difficult cases. When exome sequencing is used as a first line test in a prospective cohort, diagnostic rates of up to 57.5% have been reported (this was part of the Melbourne Genomics Health Alliance, a study for which I performed bioinformatic analysis) [3].

When trying to obtain a genetic diagnosis for a likely Mendelian disease there are two common approaches: singleton and trio sequencing. When using a singleton strategy, only the proband is sequenced, while in a trio strategy the parents of the proband are also sequenced. There is also a diversity of other family sequencing patterns, such as including siblings. The advantage of trio sequencing, or of sequencing any number of additional family members is that it allows a large number of potential disease variants to be filtered out based on inheritance patterns. For example, if a disease is thought to be *de novo* dominant then any variants found in the parents can be filtered out. If a disease appears to be recessive, then we can exclude variants that are homozygous in either parent. Most likely due to these powerful filtering strategies, a trio sequencing approach outperforms the singleton strategy, with diagnostic rates of 33% for trios and 22% for singletons [2]. However, the trio approach is generally more expensive, sequencing two additional samples, so overall the singleton approach may result in more genetic diagnoses for a given sequencing budget because more affected individuals can be sequenced.

Note that one limitation of this argument is that diagnostic rates are heavily impacted by differences in the analysis methods used in these studies as well as the specific diseases that were assessed. Therefore, diagnostic rates should be interpreted with caution when comparing between studies. In addition, assigning a specific condition to any of the above categories can be further complicated by differences in phenotype between individuals with the same genotype. For example reduced penetrance occurs when only a proportion of individuals with a given genotype have the disease, while variable expressivity

refers to differences in symptoms that can occur between individuals with the same genotype.

1.3.3 Analysis pipeline: from sequencing reads to disease variant

In recent years bioinformatic pipelines for analysing clinical sequencing data have tended to converge on a core set of essential steps and tools. There have been a number of standard analysis pipelines developed for clinical genomics. I helped develop one such pipeline, Cpipe [31], which was developed for the Melbourne Genomics Health Alliance project, and has since been integrated into clinical practise here at the Victorian Clinical Genetics Services (VCGS) and at other clinical centres.

Briefly, the key stages of most clinical genomics pipelines are as follows (Figure 1.1). A DNA sample from the patient is sequenced to produce raw sequencing reads (FASTQ files). These sequence reads are aligned to the reference genome to produce a BAM or CRAM file. The reads are then compared to the reference genome at each position to assess the evidence that the individual has a different DNA sequence (a variant) compared to the reference genome (creating a VCF file). This full list of variants that differ from the reference is processed by a series of steps including variant annotation, filtering and prioritisation. Lower quality variants may be removed, while information about the genomic context, predicted consequences and other characteristics of variants are used to filter or prioritise them, working towards a short list of candidate disease variants. At this stage, one or more likely pathogenic variants may be found, and a diagnosis made. However, the case may remain unsolved and/or move into a research gene discovery paradigm. Quality control (QC) metrics are generally collected and checked at all stages in this pipeline. I will describe the main pipeline stages in more detail, highlighting the key methods and considerations at each point.

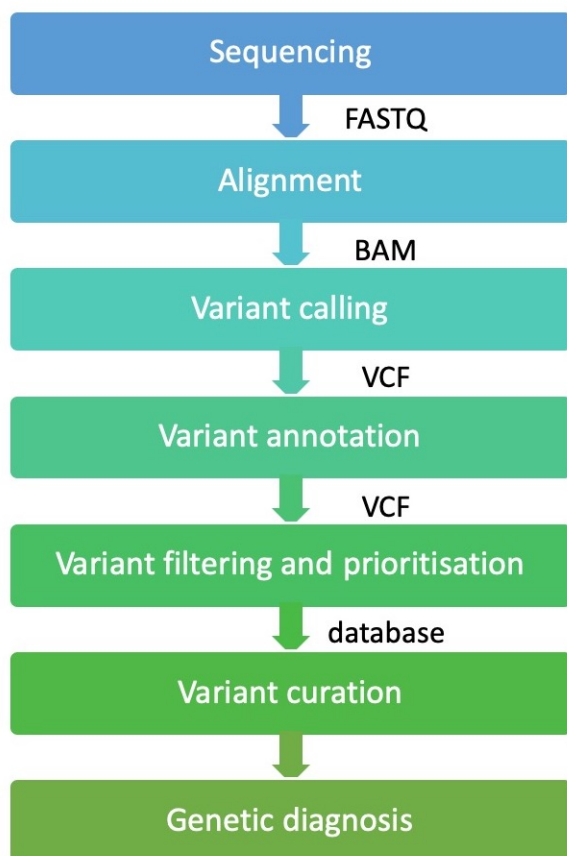


Figure 1.1: Typical stages in a clinical genomics pipeline.

1.3.3.1 Alignment

DNA sequencing in the clinical context is most frequently performed with paired-end short read Illumina sequencing technology. This generally produces a pair of FASTQ files, containing the forward and reverse reads. These reads are aligned to the human reference genome. While there are many alignment algorithms, BWA-MEM [32] has emerged as the most frequently used algorithm in this context, as it is more sensitive to larger indels than other methods [33]. It generates a BAM file describing the position of each read on the reference genome and any differences between the read and the reference.

1.3.3.2 Variant calling

Variant calling of SNVs and short indels has converged on the Genome Analysis Toolkit HaplotypeCaller algorithm [34]. Most labs use this algorithm in the context of the GATK Best Practices recommendations for germline SNV and short indel variant calling, which includes a number of intermediate processing steps [35]. Likely duplicate reads are identified and flagged, and base quality scores are recalibrated. Variant calling may be performed on individual samples; however, the guide recommends performing joint calling, integrating variant information across individuals. The resulting variants are then filtered and refined using quality metrics. While the recommended joint calling can be used to weight the confidence of variant calls based on their frequency in a set of samples, this may be undesirable in a clinical context. In clinical diagnostic labs there is emphasis on the reproducibility of results, therefore using a method that draws information across samples may generate unwanted variability.

1.3.3.3 Variant annotation, filtering and prioritisation

One difficulty that arises when using exome or whole genome sequencing in a clinical setting is the sheer number of variants discovered by these technologies. For example, an average individual exome has more than 20,000 variants. Typically 10,000-12,000 of these variants will be non-synonymous (changing the amino acid sequencing of a protein), 120 will be protein truncating and 54 of the variants will have been previously reported as pathogenic [36,37]. Therefore, it is critical to filter and prioritise this set of variants to produce a more manageable list.

Variant annotation is an area where there is much more variability between different analysis pipelines and different diagnostic laboratories. Currently, the two main tools used to perform variant annotation are Annovar [38] and VEP [39]. These tools enable annotation of variants based on a large set of different criteria and integrate other annotation tools or metrics. One of the

main goals of variant annotation is to determine the genomic context of the variant. For each variant the tool reports if it overlaps a known transcript, or other genomic feature such as a promoter or splice site (many of these features and their implications are summarised in section 1.1.3). These tools then make a consequence prediction. For example, a variant that produces a stop codon in a protein coding transcript might be predicted to have a protein truncating consequence. These tools may also optionally add information from other consequence-prediction methods, for example SIFT [40] uses homology information to predict if an amino acid substitution will impact protein function, while PolyPhen [41] makes similar predictions but using physical characteristics. Condel [42] and CADD [43] scores integrate information from other scores to provide a consensus deleteriousness prediction. Previously reported pathogenic variants can also be annotated by drawing on databases such as ClinVar [44]. Databases such as Matchmaker Exchange [45] can be used to determine if a variant has been seen in an individual with a similar phenotype, even if that variant has not yet been clearly established as pathogenic.

Variants are also commonly annotated with their population allele frequency. There are several databases of variant frequencies, such as Exome Variant Server [46], 1000 Genomes [6] and the Genome Aggregation Database (gnomAD, previously known as Exome Aggregation Consortium or ExAC) [37,47]. With the largest number of individuals, gnomAD is currently the most used. It also provides gene-level scores estimating tolerance to Loss of Function (LoF) variants, the pLI score, and more recently the LoF observed/expected ratio. Allele frequencies can be used to filter variants that are common in the population, and therefore are unlikely to cause severe early-onset disease. Typical variant frequency thresholds are 0.01 for a “rare” variants, and 0.0005 to be considered a “very rare” [31]. The threshold used may depend on the severity of symptoms and the pattern of inheritance. For

example, dominant diseases tend to be caused by rarer variants than recessive diseases. As allele frequencies can differ drastically between populations, it is important to use a variant frequency database that is matched to the individual. For example gnomAD provides allele frequencies for the entire set as well as 7 specific sub-populations.

Although many diagnostic laboratories and research groups use broadly similar analysis pipelines (especially with regard to alignment and variant calling), they often make different choices with regard to variant filtering and prioritisation that can have a large impact on which variants remain or are prioritised at the end of the pipeline. For example we have previously compared the use of clinically-determined gene lists with Exomiser [48], CADD and Condel scores to determine variant ranking, and found dramatic differences in the number of known causal disease variants that were ranked at the top of the list [49].

1.3.3.4 Variant curation and diagnosis

In the final stage, the set of annotated variants is typically imported into a variant database with a graphical user interface to allow manual curation of the results. Common database choices are LOVD [50] and Seqr [51]. These databases may also link other sources of information about the variants, for example gene-level disease information from Online Mendelian Inheritance in Man (OMIM) [52]. They may also incorporate variant calls from family members to allow filtering based on an inheritance model.

The process of variant curation typically involves applying the ACMG guidelines [28], or other laboratory-specific strategies, to determine if there is a sufficient evidence that a particular variant is pathogenic. This process will generally draw information from all the variant annotations that I have described. For example, considering if the variant is likely to impact protein function, if it is rare in the population and if it is in a gene that is consistent

with the patient's phenotype. In many cases variant curation will also require searches of other databases and of the scientific literature. It is worth noting that the ACMG guidelines offer substantial scope for interpretation, which hinders automation and makes this stage a highly labour-intensive process.

Variant curation can be lengthy process. Therefore, it is critical to use filtering to limit the number of variants that must be curated for each patient. In many cases variants will be ranked, and then curation will continue from the top of the list onwards until either a likely causal variant is found, or until a specified limit of variants have been checked. To have the best chance of making a genetic diagnosis it is critical to rank the best candidate variants first.

Finally, in a clinical diagnostic setting, a clinical report will be produced. It describes any relevant variants, or reports that there were no confident disease-causing variants, or that there were variants of uncertain significance found. Clinicians then use this report as evidence towards making a genetic diagnosis.

1.4 Short Tandem Repeats

As mentioned previously, most genomic testing pipelines only genotype SNVs and short indels, so there are a range of genetic variants that go completely unobserved. Some of our failures to reach a genetic diagnosis stem from our inability to detect the causal variant in short read sequencing data. One type of genetic variant that is rarely tested is short tandem repeat expansions. For example, in Chapter 3 I describe a patient who underwent multiple rounds of genetic testing, culminating in whole genome sequencing. This patient did not receive a diagnosis until years later when we detected a known pathogenic STR expansion using STRetch. As STR expansions are a major focus of this thesis, this section summarises the STR literature.

Short tandem repeats (STRs), also known as microsatellites, simple sequence repeats, or simple sequence length polymorphisms, are short DNA sequences

of one to six base pairs that are repeated consecutively [53]. In contrast, minisatellites (also known as variable number tandem repeats) contain longer – 10 to 60 base pair – repeat unit motifs [54]. Discussion of repeats with motifs of seven to nine base pairs is rare in the literature, and these are generally grouped under the general term “tandem repeats”. Tandem repeats (including STRs and minisatellites) show a large degree of variation between individuals, with mutation rates 10 to 100,000 times higher than average mutation rates in other parts of the genome such as SNPs [55].

The terms STR and microsatellite are used interchangeably, with the choice often depending on the context. For example, in population studies the term microsatellite is often preferred, while in human context STR is more commonly used. I will use a range of terms to describe the qualities and components of a STR. Figure 1.2 summarises this nomenclature. The repeat units of a STR are its building blocks and have a characteristic motif (in this case AC). An STR locus consists of a number of repeat units which are arranged in tandem (head-to-tail). Repeat length refers to the number of repeat units in a given allele. Repeat length can vary between individual chromosomes at a given STR locus, so that an individual with a heterozygous STR genotype has two alleles with different numbers of repeat units and thus with different repeat lengths. An STR locus has upstream and downstream flanking sequences. Most STRs in the human genome are simple STRs. These contain a simple DNA sequence repeated in tandem. Compound STRs contain two or more adjacent simple STRs. In addition to simple and compound STRs, it is possible to find complex STRs, which are compound

STRs that are interrupted by short non-variable sequences [56].

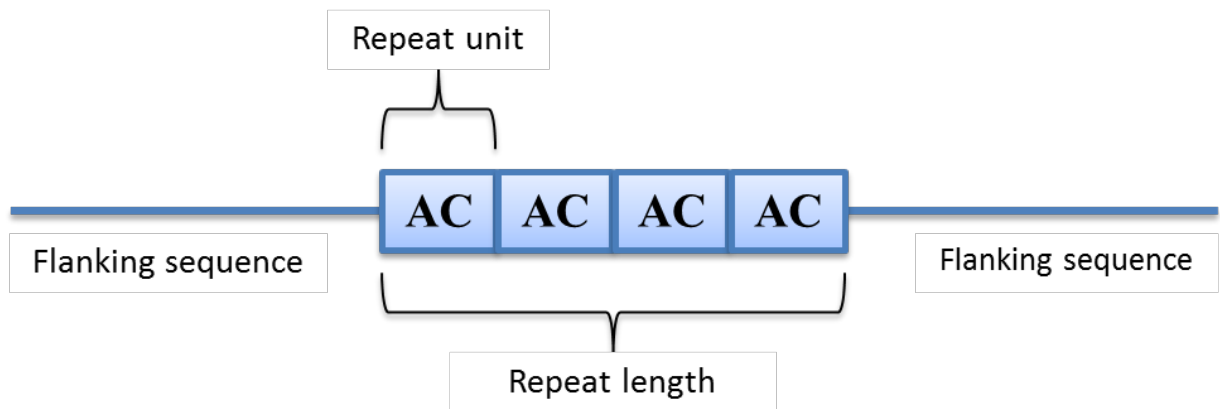


Figure 1.2: STR structure and nomenclature using a dinucleotide AC motif as an example.

Approximately 3% of the human genome consists of STRs [57]. Of these, 5353 loci overlap with exons and 537 genes have at least one STR overlapping with their coding region (see section 4.2). Dinucleotide STRs are the most common [58], however STRs with tri-nucleotide (3bp) and hexa-nucleotide (6bp) repeat units are enriched within exonic regions, while other repeat unit sizes are predominately found in non-coding regions [57]. STR expansions and contractions which are a multiple of three would produce in-frame deletions, and thus are likely to have a lower impact on protein coding regions than STRs with repeat units that are not a multiple of three.

1.4.1 STR mutation mechanism

Tandem repeat loci are prone to frequent mutations and high polymorphism, with mutation rates 10 to 100,000 times higher than average mutation rates in other parts of the genome [55]. Tandem repeats have been reported to have mutation rates in the order of 10^{-3} to 10^{-7} mutations per cell division (reviewed in Gemayel et al., 2010). STR mutations rates vary substantially depending on the repeat unit length, the total repeat length and purity [59]. In general, shorter repeat units, longer repeat lengths and high repeat purity all increase mutation rates.

Most mutations in STR loci are the addition or deletion of one or more complete repeat units. SNPs, and indels involving partial repeat units are much less common [55]. There are two major models for natural tandem repeat variation are recombination and DNA polymerase slippage [55]. Recombination (Figure 1.3) is theorised to occur within a tandem repeat locus on a single chromosome (intra-repeat recombination), or between homologous loci on two chromosomes (inter-repeat recombination). In both cases this can result in the addition or subtraction of many repeat units in a single event. The other model is DNA Polymerase slippage. During DNA replication, the repeat units may mis-pair, causing the DNA Polymerase to “slip” and copy fewer or additional repeat units. DNA Polymerase slippage is also theorised to cause the “stutter” observed in PCR (see section 1.4.5).

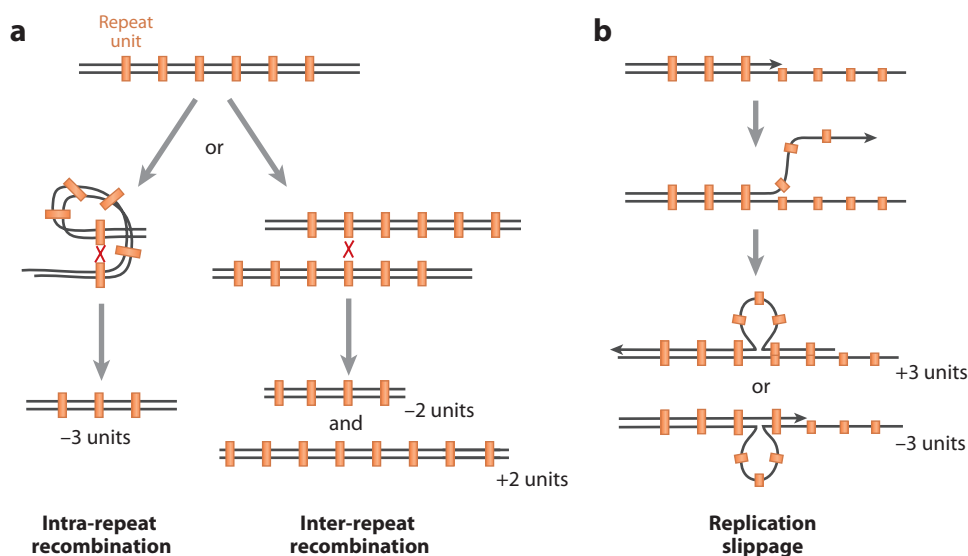


Figure 1.3: Two mechanisms for tandem repeat expansions and contractions theorised to occur within cells.

a) Recombination within a single chromosome (intra-repeat recombination) or unequal crossing over between a pair of chromosomes (inter-repeat recombination) can result in the gain or loss of multiple repeat units. b) During DNA replication

the two strands can mis-pair and result in the DNA polymerase adding or failing to add repeat units (replication slippage). Figure from [55].

1.4.2 STR variation impacts gene expression

There is substantial evidence that STRs impact gene expression and contribute to variation in gene expression within and between species [55]. This has been shown experimentally; mutating STRs in the promoter region impacts reporter gene expression in human cells [60]. STRs are found within many coding and regulatory regions of genes. There is evidence of STRs in promoter regions modulating gene expression in bacteria, yeast and mammals [61]. It has been observed that the presence of STRs in promoter regions are often highly conserved in mammals, with higher conservation observed closer to the TSS, suggesting selection for the presence of STRs in human promoter regions [62].

Perhaps the most famous example of STRs influencing gene expression comes from the prairie vole (*Microtus ochrogaster*), a monogamous rodent species found in the grasslands of the central United States and Canada. Voles form 'pair bonds', where individuals preferentially affiliate and copulate with their partner as well as sharing a nest and rearing offspring together [63]. STR variation in the promoter region of the *Avpr1a* gene influences expression of that gene, which encodes arginine vasopressin receptor 1a. Longer STR alleles were associated with monogamous behaviours, specifically pup grooming behaviour in males [64]. STR variation in the corresponding human gene has been associated with differences in monogamous behaviour in men [65] as well as a range of other social phenotypes. STR variation has been shown to influence *AVPR1A* gene expression levels *in vitro* [66] and in *post mortem* brains [67,68].

There are a number of mechanisms by which STRs are hypothesised to influence gene expression. STRs consisting of alternating pyrimidines and purines promote Z-DNA formation [55]. This unusual left-handed form of

DNA (normally right-handed) has been observed to act as a regulatory element for nearby genes [69]. Z-DNA is more likely to form near promoters because these regions more commonly experience negative torsional strain due to RNA polymerase helicase activity as well as chromatin remodelling. Z-DNA can prevent nucleosome binding or can block the movement of RNA polymerase activity proteins [61]. In rats, it has been shown that the length of a (AC)_n STR just upstream of the TSS is associated with changes in gene expression, due to Z-DNA formation [70].

Many transcription factor binding sites are STRs, therefore changing the length of these STRs can change the number of binding sites [55]. Other STRs fall between functional elements in promoters so that changes to their repeat length changes the spacing between these sites [71]. STRs can also influence promoter activity by altering the structure of chromatin [72].

In humans, variation at over 2,000 STR loci has been associated with changes in gene expression [73]. It is clear that through a variety of mechanisms, STRs modulate gene expression over many species and may be important in the regulation of a substantial number of genes.

1.4.3 STRs in Human Disease

Tandem repeats have been implicated in dozens of human diseases from neuromuscular and neurodegenerative diseases – such as Huntington’s disease, multiple forms of ataxia and fragile X syndrome – to cancer.

1.4.3.1 Mendelian disease

Tandem repeat expansions have been implicated in more than 30 Mendelian human diseases (a subset of these are summarised in Table 1.1). Many of these conditions affect the nervous system, for example Huntington’s disease, spinocerebellar ataxias, spinobulbar muscular atrophy, Friedreich ataxia, fragile X syndrome and many of the polyalanine disorders [74]. There are eleven STR loci for which expansions are known to cause spinocerebellar

ataxia and these diseases can also be caused by SNVs or indels at other loci. Most tandem repeat expansion disorders show dominant inheritance (Table 1.1). STR expansions can cause disease through a variety of mechanisms, including polyglutamine aggregation, changes to methylation, RNA toxicity, and repeat associated non-ATG (RAN) translation [75].

Table 1.1: Summary of human inherited diseases caused by tandem repeat expansions.

AD = autosomal dominant, AR = autosomal recessive, XD = X-linked dominant, XR = X-linked recessive, *intermediate = premutation/incomplete penetrance/uncertain pathogenicity. Sources: [52,75–77]

disease	gene	repeat unit	inheritance	type	normal	intermediate*	pathogenic
CANVAS	RFC1	AAGGG	AR	intronic	0*		400-2000
CCHS	PHOX2B	GCN	AD	coding	20	24	25-33
DM1	DMPK	CAG	AD	coding	5–34	35-49	50-2000
DM2	ZNF9	CCTG	AD	intronic	11–26	27-74	75-11,000
DRPLA	ATN1	CAG	AD	coding	3–35		49-93
FRA12A	DIP2B	CGG	AD	5'UTR			
FRAXE	AF2/FMR2	CGG	XR	5'UTR	6-25		200-2000
FRDA	FXN	GAA	AR	intronic	5–33	34-65	66 to 1700
FTDALS1	C9orf72	GGGGCC	AD	intronic	3–25	20-60	>60
FXS/FXTAS/ POF1	FMR1	CGG	XD	5'UTR	5–44	45-200	200-2000
HD	HTT	CAG	AD	coding	6–26	27-39	40–250
HDL2	JPH3	CAG	AD	3'UTR	6–28	2-39	41–58
HFG	HOXA13	GCN	AD	coding	12-18		18-32
OPMD	PABPN1	CGN	AD/AR	coding	10		11-17
SBMA	AR	CAG	XR	coding	9–34	36-37	38–68
SCA1	ATXN1	CAG	AD	coding	6–35	36-38	39–88
SCA10	ATXN10	ATTCT	AD	intronic	10–32	280-850	800-4500
SCA12	PPP2R2B	CAG	AD	5'UTR	4–32		51–78
SCA17	TBP	CAG	AD	coding	25–40	41–48	49 to 66
SCA2	ATXN2	CAG	AD/AR	coding	14–32	31-34	33–200
SCA3/MJD	ATXN3	CAG	AD	coding	12–44	45-59	60-87
SCA31	BEAN	TGGAA	AD	intronic	0*		
SCA36	NOP56	GGCCTG	AD	intronic	3 to 14	15-649	650-2500
SCA6	CACNA1A	CAG	AD	coding	4–18	19	20–33
SCA7	ATXN7	CAG	AD	coding	4–19	28-36	37–460
SCA8	ATXN8/ ATXN8OS	CAG	AD	untranslated exon	15–50	50-70	71-1300

Tandem repeat expansion diseases are characterised by a mode of inheritance called genetic anticipation. Many of these diseases show increased expressivity through generations, showing greater severity and earlier age of onset as the tandem repeat expands [76]. In some diseases, the penetrance (probability that a given individual is affected) also increases with number of repeat units and hence in later generations, and in some cases depending on the gender of the parent who transmitted the repeat expansion [78]. The number and position of imperfect repeat units may also influence penetrance. Together, these features produce genetic anticipation; the disease worsens through generations.

For most STR disease loci allele sizes are characterised into normal, intermediate and pathogenic ranges (Table 1.1). Intermediate alleles generally do not cause disease, however in some cases they may cause an intermediate phenotype or convey reduced risk of a disease. For example, expansions at the FMR1 locus of 200 CGG repeat units cause fragile X syndrome (FXS) [77]. Alleles, between 55 and 200 repeat units at the same locus increase the risk of primary ovarian insufficiency (POI) in females and fragile X-associated tremor/ataxia syndrome (FXTAS) in both sexes, with males at greater risk.

For many of these diseases, alleles in the intermediate range are termed “premutation” alleles because they become unstable when passed on, so that the children of an individual with a premutation allele are at risk of inheriting a pathogenic allele [79]. For several STR diseases, premutation alleles have been matched to founder haplotypes, suggesting a shared genetic origin rather than premutations occurring *de novo* in multiple unrelated individuals [80–83].

The most notable class of STR expansion diseases are the PolyQ disorders. These are caused by CAG repeat expansions in coding regions, resulting in a polyglutamine (Q) expansion in the protein product. These include

Huntington's disease, Kennedy's disease and several of the spinocerebellar ataxias [79].

Huntington's disease (HD) is the most common of the polyQ disorders, and has been most extensively researched, with findings likely to be applicable to other polyQ disorders (especially those causing neurodegeneration). HD is caused by a CAG STR expansion in the coding region of the Huntingtin gene (HTT), which codes for the protein Huntingtin (Htt), eventually resulting in a range of later-onset cognitive, psychiatric and motor symptoms [79].

Along with about 100 other human proteins, Htt has repetitive polyglutamine tract. When this polyQ tract exceeds a given length (approximately 35 glutamines for most diseases), this can result in protein aggregation, leading to amyloid-like fibrils. This triggers the protein misfolding pathway, and can result in cellular toxicity and eventually neurodegeneration [84].

Polyalanine disorders are another class of diseases caused by GCN STR expansions in coding regions, resulting in a string of alanines in the protein product (listed in Table 1.1). Most of these disorders involve loss of function to transcription factors required during development [85]. The pathogenesis of polyaniline disorders is not yet well understood.

A substantial number of tandem repeat expansion disorders are caused by STRs in non-coding regions, particularly promoters, 5' and 3' UTRs, and introns (Table 1.1). These include several of the spinocerebellar ataxias, the myotonic dystrophies and Fragile X and associated syndromes. The pathogenesis of many of these disorders has been identified as abnormal gene expression or changes in RNA structure or function [79]. Interestingly, of disease-causing repeat expansion disease mutations in non-coding regions, most are also triplet repeats; although it is difficult to determine if this reflects a biological process or simply confirmation bias.

1.4.3.2 Cancer: Microsatellite instability

Microsatellite instability (MSI) is when the somatic mutation rate of many STRs across the genome increases due to a defect in the DNA mismatch repair system [86]. MSI is an indicator that the cancer has a mutator phenotype, and is more prone to mutations in multiple oncogenes and/or tumour suppressor genes [87].

These STR mutations can contribute to carcinogenesis through the subsequent inactivation of tumour-suppressor genes or changes to regulatory sequences in proto-oncogenes or other genes [88]. STRs can inactivate genes through frameshift mutations, activate genes by enhancing transcription factor binding within promoter regions, or any other of the myriad ways STRs can impact gene expression, as discussed in section 1.4.2.

MSI has primarily been used as a prognostic marker, to identify cancers with mutator phenotypes, and thus guide diagnosis and treatment. The assignment of STR-stable vs. STR-unstable cancer phenotypes is usually based on PCR electrophoresis of a set of STR loci in the Bethesda panel [89,90].

Several software tools have been developed to detect MSI from short-read sequencing data, for example MSIseq [91], MSIsensor [92] and mSINGS [93]. In general, these tools look for signatures of genome-wide STR instability in tumour compared to normal samples, rather than genotyping specific alleles.

1.4.4 Genotyping STRs by length

Traditionally, STR variation has been identified using gel electrophoresis [94]. Polymerase chain reaction (PCR) is performed using primers, which are complementary to unique sequences flanking the STR. An amplicon is the DNA product of a PCR reaction. It consists of many copies of the DNA between a pair of primers. Amplicons are separated on the basis of size by capillary gel electrophoresis to determine the sequence length, which reflects the number of repeat units.

There are a number of problems with genotyping STRs using PCR and electrophoresis. Each STR genotyping assay must be individually developed, optimised and validated before it can go into production. Although many of these protocols have been scaled to handle dozens of samples, it is still labour intensive and costly.

Further, while PCR has been found to give relatively accurate length estimates for short alleles it can become inaccurate for or even completely miss longer alleles. STRs are particularly prone to allelic dropout in PCR reactions [95,96]. This is where one allele fails to be amplified and so the individual appears to be homozygous for the other allele. For example, GC rich repeats and secondary structure may hamper the amplification of STRs. The longer allele is more likely to experience dropout, because the shorter is preferentially amplified, and many polymerases may not be capable of spanning some of the larger alleles. However, either allele can be affected by dropout due to mutations in the primer binding sites. For this reason, many clinical tests for pathogenic STR expansions test for two normal length alleles as a screening method.

Another disadvantage of PCR and electrophoresis genotyping is that it is only able to reveal the overall length of the STR. For example, a compound allele $(TC)_n(AC)_m$ with 33 repeat units, may actually be any of five cryptic alleles ranging from $(TC)_6(AC)_{27}$ to $(TC)_{10}(AC)_{23}$. There are several examples of pathogenic STR loci that are compound STRs. For example, the Huntington's disease locus and one of the Muscular Dystrophy loci (DM2) are both directly adjacent to at least STR locus with a different repeat unit. In both these cases, the adjacent STRs are not known to cause disease.

Another issue arising during the PCR is that the polymerase has a tendency to 'slip' on the STR during replication [97]. This is theorised to occur when the repetitive sequences mis-pair with neighbouring identical sequence (Figure

1.3b). DNA polymerase slippage has the potential to produce daughter fragments that are a number of repeat units longer or shorter than the original fragment. Over the course of a PCR, these artificial mutations can be replicated and further insertions and deletions may occur, amplifying the effect.

The insertion and deletion mutations that occur during PCR manifest as a characteristic ‘stutter’ in the gel electrophoresis. The degree of stutter increases with repeat length (Figure 1.4) and more stutter bands are observed in STRs with shorter repeat units [97]. Stutter has proven a challenge to STR analysis algorithms.

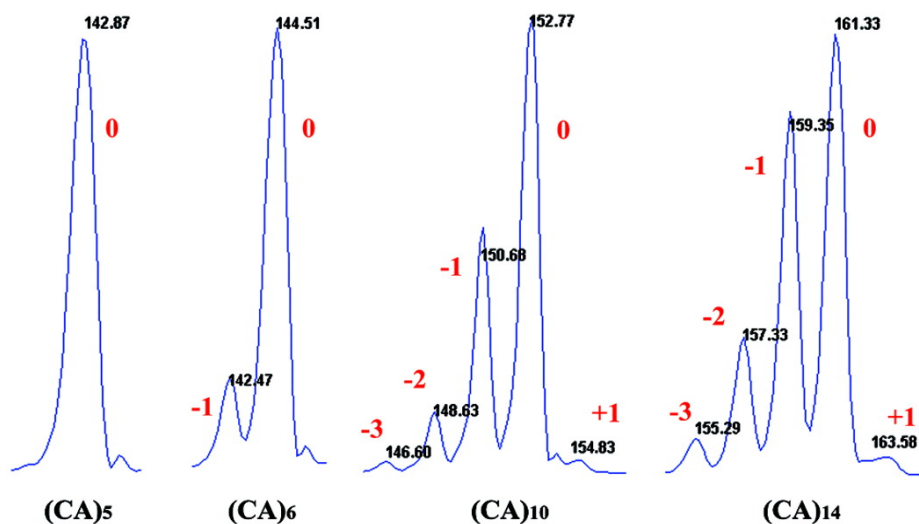


Figure 1.4: ‘Stutter’ profiles of CA STR amplicons.

The number of CA repeat units varies from 5 to 14 (horizontal axis). The ‘0’ peak is the inferred original allele size. The ‘+’ and ‘-’ peaks represent additions and deletions of one CA repeat unit due to stuttering. Peak heights (vertical axis) indicate number of amplicons present. The electrophoresis migration distance of amplicons, which reflect repeat length, are indicated for each peak (adapted from Shinde et al., 2003).

Because of the expense and difficulty of PCR assays for STRs, diagnostic laboratories only typically test a small number of loci. For example, even though expansions in more than 10 STR loci known to cause ataxia [98], our

collaborators at PathWest diagnostic lab usually only test a few of the most common loci. Any additional STR testing would require custom assay development, which can take days or weeks per locus (and which PathWest have kindly done for selected STR expansions predicted by STRetch in Chapter 3). With the increasing application of exome and whole genome sequencing to rare disease, there is a clear need to detect potentially pathogenic STR expansions in next-generation sequencing data.

1.4.5 Sequencing STRs

1.4.5.1 Sanger sequencing

Sanger sequencing has also been used to determine genotypes in compound and complex STRs, where the length of the allele is not a sufficient measure of the genotype. Although the long reads in Sanger sequencing are useful for spanning long STR loci, in addition to its expense, this method also suffers from the issue of stutter. The final step in Sanger sequencing is capillary gel electrophoresis of the sequencing reaction. This involves measuring a large population of DNA fragments – each with different amounts of stutter – that cannot be easily disentangled. While stutter can make Sanger sequencing difficult to interpret, heterozygosity causes an even greater problem. Because the alleles are of different lengths, the repeat units closest to the end of the longer allele are overlaid with the flanking sequence of the shorter allele, making it difficult to confidently determine the lengths of either allele.

1.4.5.2 Short read sequencing

The use of next generation sequencing has the potential to overcome many of the problems with traditional STR genotyping assays. Sequencing enables the detection of compound and complex STRs and thus detection of the full range of variation at STR loci. In contrast to Sanger sequencing, next generation methods characterise each DNA fragment individually and so do not suffer as drastically from stutter or heterozygosity. However, because

many of these methods involve an initial PCR amplification step, it may still be necessary to include a model of stutter in any analysis algorithm.

There are a number of considerations when choosing a sequencing technology to use to genotype STR polymorphisms, in particular, read length and cost. To accurately determine the length of STR alleles, most genotyping software tools require reads that fully sequence the locus from one end to the other.

Pathogenic STR alleles typically range from 100-5,000 bp, with some reports of allele as large as 40,000 bp (Figure 1.5). Additionally, tools that look at STRs within reads require some flanking sequence to provide the specificity needed for sequence mapping. For even moderately sized STR expansion of around 100-200 bp read length of at least 200-300 base pairs is likely to be required. This immediately precludes many short-read technologies for longer STR loci, although they may be useful for genotyping shorter loci. Both Ion Torrent and Roche 454 offer sufficient read length, however Ion Torrent is considerably cheaper (and still being supported). Illumina MiSeq machines can now sequence beyond 300 base pairs, making them a cost-effective choice for genotyping longer loci. Yet the fact remains that the vast bulk of DNA sequencing data is produced with Illumina technology with read lengths of 150 bp. While one might wish for longer reads, the reality is we need to develop algorithms for the data that exists.

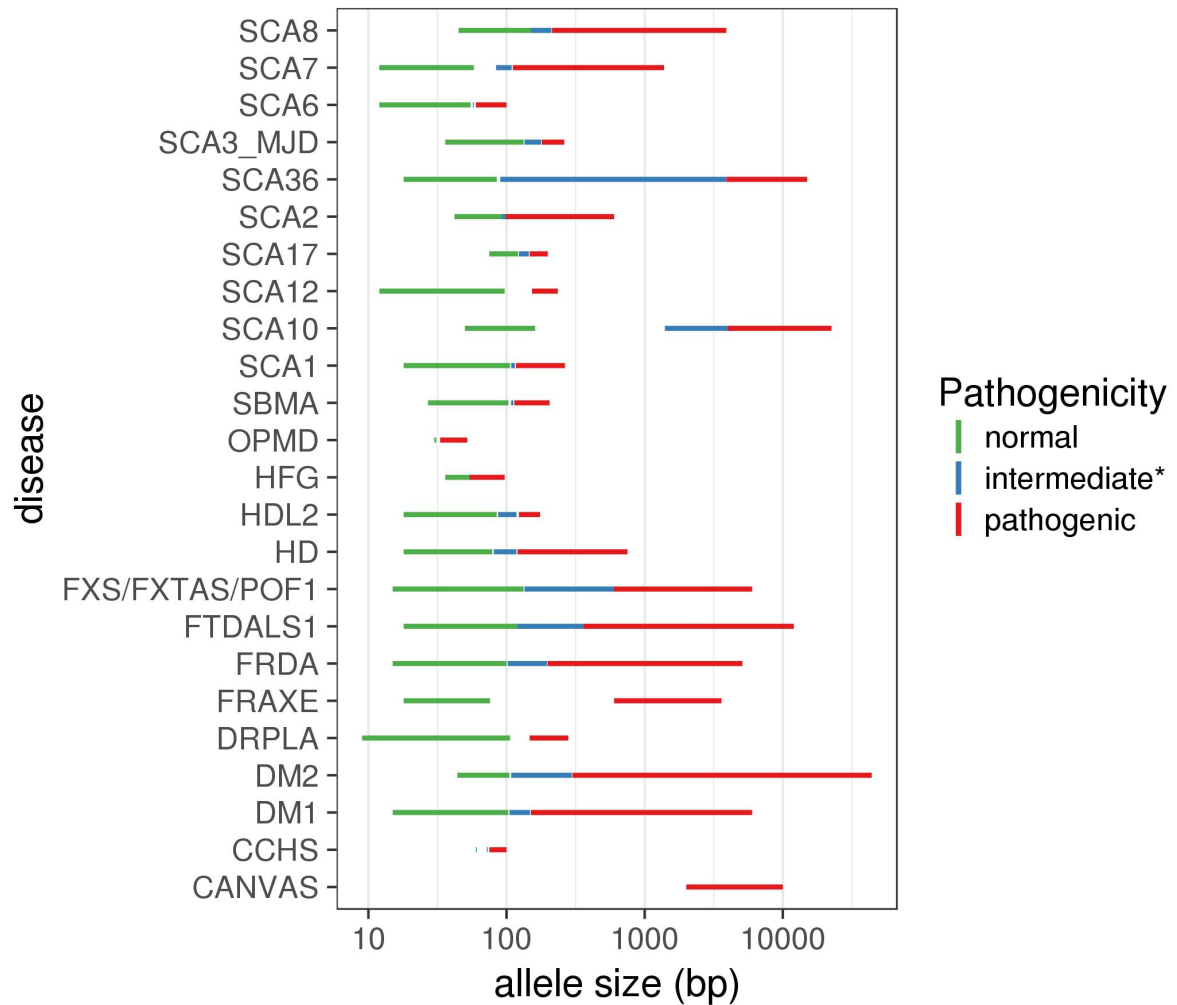


Figure 1.5: Size ranges in base pairs of normal, intermediate and pathogenic alleles in STR expansion diseases.

***intermediate refers to any premutation range, incomplete penetrance or in some cases uncertain pathogenicity. Sources: [52,77]**

Perhaps the most common targeting technique in the context of human genomics is the exome capture. Here, the hybridisation probes are designed against the coding regions of most genes, and so these are preferentially amplified. Many STRs known to cause human disease (discussed in 1.4.3.1) are located in exons and are therefore amendable to genotyping from exome sequencing data. The drawback of exome sequencing is that it will obviously miss the many STRs that occur outside the coding region, and these (particularly those in promoters, 5' and 3' UTRs and introns) are likely

candidates for new disease-causing variants. Another disadvantage of exome sequencing as compared with whole genome sequencing, is that the coverage can be quite uneven, and sequencing artefacts more common. Compared with exome sequencing, whole genome sequencing provides better coverage of target genes, and more even coverage overall, as well as more equal coverage of both alleles at heterozygous sites [99].

1.4.5.3 Long read sequencing

Single-molecule long read sequencers PacBio and Oxford Nanopore offer excellent read lengths, and have the advantage of not requiring a PCR amplification step. Although, for targeted sequencing a PCR step may still be necessary, as it is difficult to isolate a sequence of interest without selective amplification. However, both these sequencing technologies are more expensive per base sequenced, and have substantially higher error rates than short read sequencers. Although there has been a proof of principle with whole genome analysis performed using PacBio data [100], these tools are unlikely to be regularly applied to STR sequencing at a whole genome scale in the near future. It is, however, likely that these sequencers will be used for targeted sequencing of a smaller number of loci, for example there has been success genotyping the CGG-STR in the human fragile X gene using PacBio sequencing [101]. Both PacBio and ONT have also been used to confirm STRs genotypes found by other methods [102–104].

1.4.6 Algorithms for genotyping STRs in high-throughput sequencing data

A number of software tools already exist to detect simple STR polymorphisms in next generation sequencing data. The most well-known are summarised in Table 1.2. Most software for STR genotyping falls into one of the following categories:

- Software that runs on short read, paired end Illumina data to detect:

- STRs that are shorter than the read length
- STRs that are longer than the read length
- Software that runs on long reads to detect STRs that are shorter than the read length

There are also several domain-specific STR software tools especially in forensics, e.g. STRait Razor [105,106], MyFLq [107]. The most well-known general-purpose STR genotypers are summarised in Table 1.2.

Table 1.2: Summary of selected STR genotyping methods

Purpose	Sequencing data	Software name	Notes	Reference
STRs shorter than the read length	Illumina short reads	LobSTR	Includes an STR aligner, or can use mapped BAMs	[108]
		RepeatSeq		[109]
		Genotan		[110]
		STR-FM	Pipeline via Galaxy toolshed	[111]
		HipSTR	Phases STRs with SNPs/indels	[112]
STRs longer than the read length	Illumina short reads	STRViper	Requires tight insert size distribution, and therefore specialised sequencing prep.	[113]
		GangSTR		[103]
		ExpansionHunter	Limited to specified loci (disease loci supplied)	[114]
		STRetch	Uses a custom reference genome with STR decoy chromosomes	[104]
		TREDPARSE	Limited to specified loci (disease loci supplied)	[96]
STRs shorter than the read length	PacBio long reads	PacmonsTR	Does not estimate allele length	[115]
				[100]

1.4.6.1 Short read STR genotypers

There are two main strategies for genotyping STRs using short read technology (such as Illumina). The first is to genotype STRs that are shorter than the read length. The major STR genotypers in this category include LobSTR, RepeatSeq, HipSTR, Genotan, and STR-FM (Table 1.2). I also developed an algorithm for genotyping standard and compound STRs with interruption SNVs within reads as part of my Master's project, however the code was not released as a fully developed tool [116]. Finally, STR variants can also be detected by many small variant callers, as these variants can be described as sort indels. For example GATK HaplotypeCaller [34] detects these variants quite well.

The key limitation of all these tools is that they cannot genotype STR alleles greater than the read length, and most pathogenic STR expansions have allele sizes greater than the typical read lengths of 150bp (Figure 1.5). When using these tools, to be informative a read must completely span the STR and contain some unique flanking sequence at both ends (e.g. a minimum of 8 bp at each end is required for LobSTR) to enable the read to be mapped uniquely and to detect the ends of the STR. Variation in an STR can be detected by looking for indels in the reads relative to the reference.

Since existing tools to genotype STRs within the read length cannot detect most pathogenic STR expansions, in recent years a number of tools have been released that use alternate strategies to detect large STR expansions from short reads (Table 1.2). The first such method, STRViper, detects a shift in the insert size between pairs of reads. In general, this strategy will only detect larger indels relative to the reference, however its key advantage is that it may be able detect longer STR expansions, which are more likely to be pathogenic. However, upon testing this method, it failed to detect confirmed large STR expansions. From correspondence with the author we confirmed that STRViper requires a tight insert size distribution, which is not typically found

in modern sequencing data without specialised library preparation. It would also fail to detect many pathogenic expansions that are greater than in the insert size.

GangSTR, ExpansionHunter and TREDPARSE share relatively similar genotyping strategies. They each integrate information from multiple informative reads, including those containing partial STR sequences, shifts in insert size (as above) and reads where one in the pair is contained in the STR, while the other maps to a unique flanking sequence. One limitation of all these tools is that they rely on informative reads correctly aligning to the reference genome. However, we know that reads arising from large STR expansions often align to the incorrect STR locus (see Chapter 3 for more details). Although some of these tools estimate the likely off-target mapping positions of these reads, this is reference genome-specific, or manually entered. A limitation of ExpansionHunter and TREDPARSE is that they are not genome-wide. They require a set of user-specified STR loci, with a set of about 30 known pathogenic loci provided.

Another tool, exSTRa, uses reads containing partial STR sequences. While this has been shown to detect many known pathogenic expansions, it does not estimate their allele size. Therefore, secondary testing would be required to determine if the allele is in the pathogenic range.

Finally, our method STRetch uses STR decoy chromosomes to deal with the issue of reads from expanded STRs mis-aligning to the reference genome. It considers the number of read pairs where one in the pair aligns to the decoy chromosome, and the other aligns to the STR flanking sequence. This method is described in detail in Chapter 3.

To date there have been no independent reviews of short-read STR genotyping methods. As each tool was published, they often compared to select previous tools. While a small number of general reviews and

comparisons of STR genotyping tools have been published, they have all been written by one of the authors of the tools [117,118].

1.4.6.2 Long read STR genotypers

In general, these genotypers are similar conceptually to short read genotypers, which genotype STRs that are shorter than the read length. In general, these reads will span even the longest STRs. However, these technologies have higher error rates and tend to be sequenced at much lower coverage, so these genotypers have strategies to account for such errors. The only genotyper I am aware of in this category is PacmonsTR [100].

1.5 Aims of this thesis

As I have described in this chapter, there are two main ways in which a patient fails to receive a genetic diagnosis. Either the true causal genetic variant is genotyped, but its significance is not understood (for example the variant is not highly ranked, or cannot be interpreted), or the causal variant is not detected at all. In this thesis I make contributions to both these issues:

1. I address the problem of prioritising a large number of variants in two ways: pooled-parent exome sequencing allows filtering out variants seen in the parents, and the STR variation work helps to prioritise STR expansions that are more likely to be pathogenic.
2. I also address the issue of large STR expansions not being genotyped, by developing STRetch, a method to detect large STR expansions from short read data.

In addition to these general aims, I will summarise the specific aims of each thesis chapter.

In Chapter 2 I propose a new sequencing strategy for prioritising *de novo* variants using pooled parent exome sequencing. I use simulation to consider several different possible strategies to perform pooled parent sequencing. In

particular, I consider the questions of how many parents to pool and how deeply to sequence to optimise the number of variants recalled and the cost. I also consider which variant caller is best suited to pooled parent sequencing data. I then show the effectiveness of the pooled parent strategy at filtering inherited variants in a real analysis. This chapter has been published a preprint on bioRxiv [119].

In Chapter 3 I develop a new method, STRetch, to detect pathogenic STR expansions in whole genome sequencing data. I validate this method on individuals with known STR disease loci, and also confirm predicted expansions at other STR loci that are not known to be pathogenic. I also assess the rate of STR expansions in known pathogenic loci in a set of almost 100 samples. The STRetch code and reference data are made freely available to the research and clinical communities. This chapter is in the form of a publication from 2018 [104].

In Chapter 4 I extend the work from Chapter 3 to explore the frequency of STR expansions called by STRetch in more than 300 individuals. I first consider the attributes of more than 300,000 STRs that are annotated in the human reference genome, then consider the evidence for expansions at all these loci. To enhance our ability to interpret and rank these variants I consider the value of genomic context and constraint to prioritise STRs in a locus-discovery setting. I also look at counts of reads aligned to the STR decoy chromosomes to detect STR repeat units that are observed more frequently in the samples than we would expect based on the reference genome. I then assess the frequency of expansions in known pathogenic STR loci and comment on the choice of controls samples when running STRetch in research and clinical contexts.

Finally, in Chapter 5, I will draw together the concepts and findings of this thesis and consider them in their broader scientific context.

Chapter 2 Pooled-Parent Exome Sequencing to Prioritise *de Novo* Variants in Genetic Disease

This chapter is adapted from the following preprint [119], and has also been submitted for publication:

Dashnow, Harriet, Katrina M. Bell, Zornitza Stark, Tiong Y. Tan, Susan M. White, and Alicia Oshlack. 2019. “Pooled-Parent Exome Sequencing to Prioritise *de Novo* Variants in Genetic Disease.” *BioRxiv*. doi:10.1101/601740.

Supplementary materials can be found in Appendix A.

Acknowledgement

I would like to thank my collaborators for all their contributions to this paper. In particular, I would like to thank Alicia Oshlack, who conceived the pooled-parent concept and guided the project. Katrina M. Bell helped with data analysis. Zornitza Stark, Tiong Y. Tan and Susan M. White clinically evaluated and obtained consent from research participants, and provided data.

2.1 Abstract

In the clinical setting, exome sequencing has become standard-of-care in diagnosing rare genetic disorders, however many patients remain unsolved. Trio sequencing has been demonstrated to produce a higher diagnostic yield than singleton (proband-only) sequencing. Parental sequencing is especially useful when a disease is suspected to be caused by a *de novo* variant in the proband, because parental data provide a strong filter for the majority of variants that are shared by the proband and their parents. However, the additional cost of sequencing the parents makes the trio strategy uneconomical for many clinical situations. With two thirds of the sequencing budget being spent on parents, these are funds that could be used to sequence more probands. For this reason, many clinics are reluctant to sequence parents.

Here we propose a pooled-parent strategy for exome sequencing of individuals with likely *de novo* disease. In this strategy, DNA from all the parents of a cohort of unrelated probands is pooled together into a single exome capture and sequencing run. Variants called in the proband can then be filtered if they are also found in the parent pool, resulting in a shorter list of prioritised variants. To evaluate the pooled-parent strategy we performed a series of simulations by combining reads from individual exomes to imitate sample pooling. We assessed the recall and false positive rate and investigated the trade-off between pool size and recall rate. We compared the performance of GATK HaplotypeCaller individual and joint calling, and FreeBayes to genotype pooled samples. Finally, we applied a pooled-parent strategy to a set of real unsolved cases and showed that the parent pool is a powerful filter that is complementary to other commonly used variant filters such as population variant frequencies.

2.2 Background

De novo Mendelian diseases are single-gene disorders where the causal variant is found in the proband, but not in the somatic tissues of either of their parents. Such conditions are usually dominant, as the probability of two mutations affecting the same gene is low. *De novo* variants have been identified as the cause of monogenic disorders such as congenital heart disease [120,121], deafness, metabolic disease and a range of syndromic disorders, reviewed in [122]. *De novo* variants are rare, occurring at a rate of $\sim 1.1 \times 10^{-8}$ per position, or approximately 70 new mutations in each diploid human genome [123]. Yet, in a meta-analysis of diagnostic next-generation sequencing in children, *de novo* variants accounted for the majority of genetic diagnoses in non-consanguineous families [2]. It has also been shown that *de novo* variants are a major cause of neurodevelopmental disorders in non-consanguineous populations [124–127]. In addition, *de novo* mutations are also recognised as contributing to a number of complex conditions such as intellectual disability, autism-spectrum disorders and schizophrenia [128]. Finding the causal genetic variant of a disease can provide diagnosis, prognosis and guide treatment or management [129], yet conditions caused by *de novo* variants can be difficult to diagnose because there is no family history of that condition.

Exome next-generation sequencing (NGS) has become a key tool to discover disease-causing variants. There are two common strategies to finding a genetic diagnosis with exome sequencing: singleton and trio sequencing. In the singleton strategy only the proband is sequenced, while for trio analysis both the proband and their parents are sequenced. The trio approach is particularly powerful in the context of *de novo* mutations (e.g. [124]) where variants observed in the parents can be used as a filter to prioritise those variants in the proband that are likely to be *de novo*. While the trio approach significantly outperforms the singleton approach in terms of diagnostic rate [2], the trio

approach is substantially more costly, as it requires library preparation and sequencing of three individuals rather than one. The advantage of the singleton strategy is that while diagnostic rates may be lower, three times as many affected individuals can be sequenced for the same cost, allowing for increased capacity and so more cases overall to be solved. For example, if we sequence 100 exomes and assume a 22% diagnostic rate for singletons and 33% for trios [2], we would expect to solve 22 of 100 cases vs 11 of 33 trios.

In addition to the cost of exome capture and sequencing, we must consider the cost of variant curation. Rather than a specific fee for service, variant curation is usually a limited resource; that is the analyst may only have time to consider a limited number of variants per patient before they must declare that case unsolved and move on to the next patient. Therefore, when diagnosing patients from exome sequencing, a key consideration is the number of variants that need to be assessed in each individual. An average individual exome has 10,000-12,000 non-synonymous variants [36], 120 protein truncating variants, and ~54 variants previously reported as pathogenic (although not necessarily relevant to the given phenotype) [37]. This is clearly too many variants for curators or clinicians to assess, so some prioritisation and filtering strategies are necessary. A common strategy is to filter or prioritise variants by their population frequency based on the assumption that highly penetrant pathogenic variants will be rare or absent in unaffected individuals. Frequencies in datasets such as Exome Variant Server [46], 1000 Genomes [6] and the Genome Aggregation Database (gnomAD, previously known as Exome Aggregation Consortium or ExAC) [37,47] are commonly used. For example a frequency of 0.01 might be considered rare, and a frequency of 0.0005 to be very rare [31]. More detailed variant assessment would then consider known pathogenic variants (e.g. from ClinVar), variant consequence prediction (e.g. VEP [39] or Condel [42]), evolutionary conservation and clinically-informed gene lists [49]. Even with all

of these filtering and prioritisation tools, typically hundreds of variants still remain to be curated and the role of inheritance information is vital in reducing this list. One reason that the diagnostic rate for trios is often higher is that inheritance information can be used to filter out large numbers of variants. Studies that perform trio or other family sequencing can use an inheritance model to select variants that fit with the expected pattern (e.g. *de novo*, dominant, recessive, sex-linked). Yet as we have described, trio sequencing carries a high cost for a modest increase in diagnostic rate.

Here we propose and evaluate a compromise between the singleton and trio strategies: pooled-parent exome sequencing. In this strategy, probands are still sequenced individually. For a given batch of unrelated probands, we pool all their parental DNA, then perform exome capture and sequencing on the pool. Variants called from this pool can then be used as a powerful filter for prioritising *de novo* variants in the probands. Because the exome capture is a substantial portion of the cost (currently ~60%), this strategy provides a dramatic cost saving over a standard trio, while still allowing the majority of parent alleles to be filtered out when analysing the probands.

Pooled sequencing strategies have been used successfully for assessing population allele frequencies for genome-wide association studies and other applications, reviewed in [130]. More recently pooling has been used in for rare variants in Mendelian disease. For example a recent study used pools of 12 probands to identify *de novo* causes of neurodevelopmental disorders and so were able to detect relevant likely-pathogenic variants in 28% of cases at a greatly reduced cost [131].

In this study we assess the novel strategy of pooled-parent exome sequencing as a method to filter variants from proband exome sequencing. We assume that we are looking for rare *de novo* variants in the probands and therefore the causative variants will not be found in any of the parents. Thus, we can take

the list of variants called in the probands and filter out any variants observed in any of the parents. We first performed a series of simulations to assess the feasibility of using pooled parents, and to explore the effect of factors such as the number of parents in the pool and sequencing depth. In addition, we compare the performance of two common variant callers on pooled data: GATK HaplotypeCaller [34] and FreeBayes [132]. Finally, we present exome sequencing analysis of four probands with suspected *de novo* causal variants, and a pool of their eight parents. We assess the utility of using the pooled parents as a filter to prioritise *de novo* variants, and compare this strategy to commonly used variant filters, in particular population allele frequency. We show that the pooled-parent filter is a powerful and complementary filter to other strategies.

2.3 Results

2.3.1 Simulation set up

In order to test the utility of pooled parents for prioritising *de novo* variants we performed a series of simulations. We selected a set of 111 parents from Simons Simplex Collection that had undergone individual exome sequencing [133]. This particular subset of samples was chosen to be matched for DNA sequencer, read length and exome capture kit sequencing depth (see Methods). Only samples with at least 64X median depth over the capture region were retained in order to both remove low quality samples and to limit the range of depths such that if the two samples with the most extreme depths were combined then their reads would appear in a 40:60 ratio. Any other pair of samples combined would have depth ratios between 40:60 and 50:50. This places the simulation within the range of sample ratios likely to be seen with errors in DNA concentration and volume quantification when pooling. The final set consisted of 111 samples, 52 females and 59 males, with median depths ranging from 64X to 97X (median 78X).

We randomly sampled reads from these individuals in various combinations to simulate pooling, then calculated recall rates by comparing variants called on the original exomes to those called on the simulated pools. Full details of our simulation strategies can be found in Methods. Briefly: we simulated pools of two, four, six, eight and ten individuals, generated by extracting reads from a random subset of the 111 exomes. We simulated two different strategies for sequencing depth; constant depth and additive depth. In the constant depth simulations, the overall amount of sequencing per pool was kept constant at twice the average depth of the source BAMs. For example, to generate a pool of four 75X samples, half the reads would be drawn from each individual to produce a pool of 150X depth. Since the total depth remains constant, the number of reads sampled from each individual decreases as the number of individuals in the pool increases. For additive depth the total sequencing depth was proportional to the number of samples, so each sample was sequenced to the same depth no matter the pool size, with the total sequencing depth increasing for larger pools. For example, if a pool contained four individuals each with 75X depth, all the reads from all four individuals would be added together to create a 300X pooled sample.

2.3.2 Choice of variant caller and sequencing depth

When selecting a variant caller for pooled sequencing data, previous comparisons have primarily considered sensitivity and false positive rate [134,135]. For the purposes of a pooled-parent study design we contend that the most important feature of a variant caller is the recall, that is, the number of variants called in the individual that are also detected in the pool. In this application, recall is the most important metric because it affects the number of variants that are able to be filtered from the probands. In contrast to pooling for the detection of pathogenic variants in the pooled samples, here false positive calls are less important. However, we also assess false positives: those variants called in the pool that are not called in any of the individuals.

False positives are only problematic in the very unlikely event that they happen to coincide with the causal variant in the proband. So, while previous papers choose their variant caller based on sensitivity, specificity and false positives, we aim to optimise the recall (sensitivity).

A key issue with calling variants in pooled samples is that most variant callers assume the reads come from a single diploid individual. They expect to observe approximately 0%, 50% or 100% of reads supporting a given allele. In our pooled samples we expect many variants to differ dramatically from these ratios. In addition, more than two variants can be present at a single locus. Therefore, using a variant caller that supports setting ploidy can be advantageous in pooled samples. Both FreeBayes [132] and GATK UnifiedGenotyper [136,137] have been proposed as good variant callers for pooled sequencing data [131,134,135]. Both provide the ability to set ploidy when calling variants, allowing more than two alleles to be called at each locus, and explicitly modelling the lower read counts expected to support rare alleles in pooled data. However UnifiedGenotyper has since been deprecated in favour of GATK HaplotypeCaller [138]. More recently GATK HaplotypeCaller has introduced the option to set ploidy, with values up to 21 possible before reaching performance limitations [139]. Since this change is relatively recent, we are not aware of any published assessment of HaplotypeCaller on pooled data. In addition, HaplotypeCaller is currently the preferred genotyper for unpooled genomes/exomes and variation references such as gnomAD [37,47]. One advantage of using HaplotypeCaller for calling variants in the pool is that it is already the standard analysis tool for individual exomes, and by using the same variant caller for both the proband and the parent pool we reduce the chance of technical artefacts.

The GATK Best Practices now recommends individual calling of samples using HaplotypeCaller followed by joint calling with GenotypeGVCFs [35]. Although it is possible to set ploidy in conjunction with joint variant calling,

this is well beyond the intended use for this tool. As such we experienced errors when using joint calling in conjunction with ploidy of 16 and 20 (i.e. our simulated pools of eight and ten). We therefore performed two different analyses with HaplotypeCaller: 1) diploid joint calling on each pool and the individuals that made up that pool and 2) individual variant calling with ploidy set as appropriate for the pool size, with the individuals genotyped separately in diploid mode (see Methods). In addition, we compared the performance of these calling modes with FreeBayes.

We compared variants called from the pool to all variants called on the original exomes and calculated recall and false positive rates. For all analysis scenarios the recall across all variants for a pool of two was greater than 94% (Appendix Table 6.1). Looking across all variants, the overall trend was for recall to diminish as the pool size increased (Figure 2.1A). HaplotypeCaller had greater recall than FreeBayes for all simulated pool sizes. This difference was small for pools of two individuals, becoming dramatic in the larger pools. Overall individual calling with HaplotypeCaller performed slightly better than joint calling, especially for larger pools.

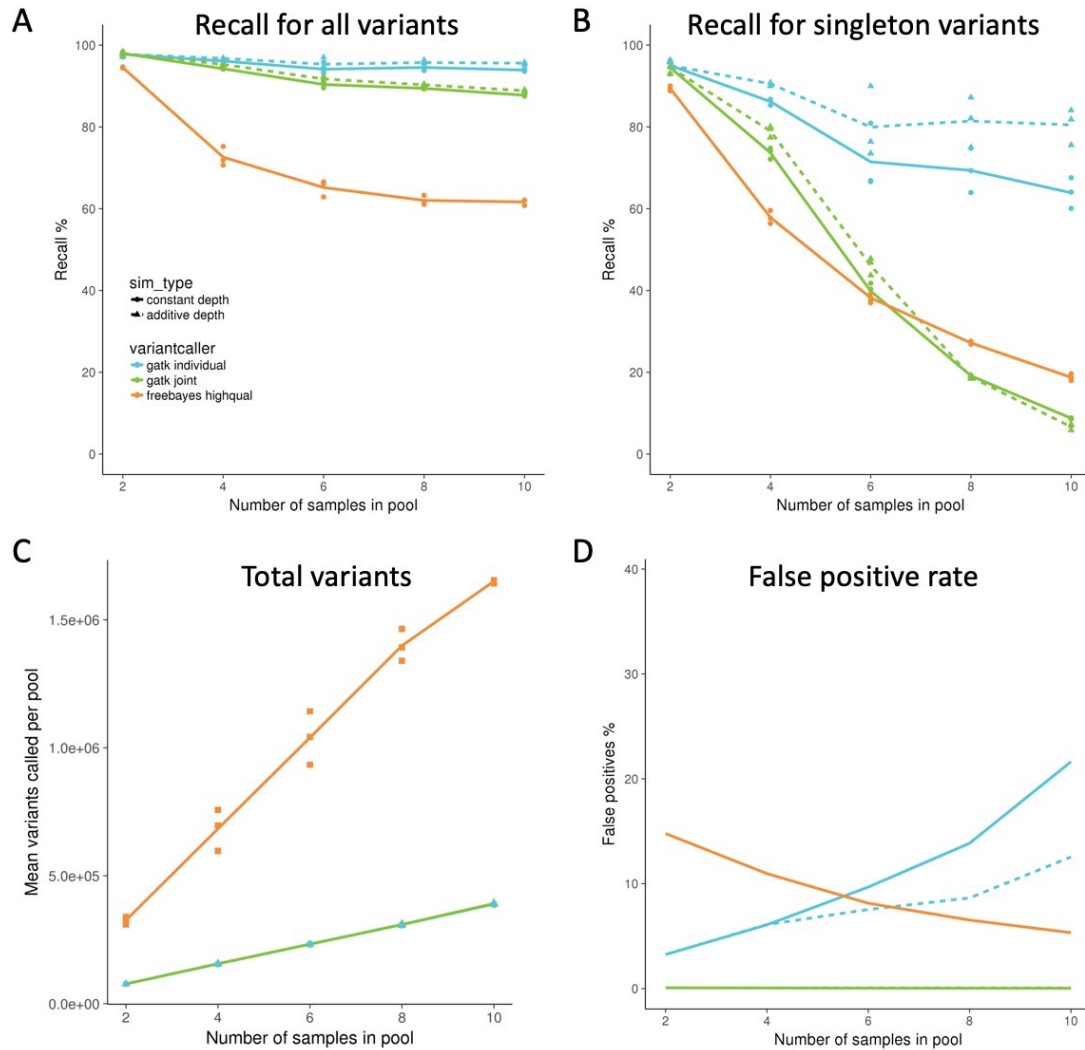


Figure 2.1: Simulated pools of two, four, six, eight and ten individuals.

Each point is a simulation (three replicates) with a different randomly selected set of individuals from a possible 111 individuals. Recall % is the percentage of variants called in all individuals that make up the pool that are also called in the pool. False positive % is the percentage of variants called in the pool that are not called in any of the individuals that make up that pool. A) Recall for all variants B) Recall for variants with an allele present in one copy in one of the individuals C) Total variants called in the pool (for constant depth simulation). HaplotypeCaller individual and joint calling produced similar numbers of variants so are difficult to distinguish at this scale. D) False positive rate for all variants. Mean values for plots A, B, C and D can be found in Appendix Table 6.1, Table 6.2, Table 6.3 and Table 6.4 respectively.

When using parents to filter for potential de novo variants, the most important and difficult class of variants to recall in the pool are those that are rare and only one allele is likely to be present in the pool. We therefore considered the recall rate for these so-called singleton variants. All calling

methods had lower recall for singleton variants (Figure 2.1B). This is expected as these variants will generally have fewer reads supporting them and lower allelic depth. Surprisingly, for HaplotypeCaller, individual calling performed dramatically better than joint calling on these variants. Joint calling showed a pronounced loss of recall with increasing pools size, especially for six or more individuals in a pool. Joint calling draws evidence across samples, thus increasing support for variants found in multiple samples and decreasing support for those variants only found in one sample. This may explain why joint calling performs poorly for singleton variants. It should be noted that FreeBayes reported approximately five to nine times as many variants as HaplotypeCaller, which may contribute to the lower recall rate (Figure 2.1C, Appendix Table 6.3).

For individual calling, HaplotypeCaller saw a steady increase in false positives with pool size (Figure 2.1D). For joint calling the overall false positive rate is less than 1% and, although it increases slightly with pool size, this increase is insignificant (Appendix Table 6.4). So, while individual calling with HaplotypeCaller provides superior recall, joint calling better controls the false positive rate. FreeBayes showed a decrease in false positive rate with increasing pool size which is likely why it was previously recommended for variant calling in pooled samples in previous studies.

We further performed an additive depth simulation where all reads from each individual are combined in the pool. In general, increasing the depth increased recall rate and decreased false positive rate. The only exception to this is for singleton joint HaplotypeCaller variants in pools eight and ten. The increase in recall with additive depth was most dramatic for HaplotypeCaller individual calling of singleton variants, where in the largest pool the recall rate increase from 63.9% to 80.5% (Figure 2.1B, Appendix Table 6.2). This indicates that increasing the sequencing depth may be useful in pooled samples. Many of the additive depth pools could not be called using FreeBayes due to massive

memory requirements on such large depth samples, so FreeBayes is not included for this simulation.

In summary, HaplotypeCaller individual calling was found to have superior performance in terms of recall rate, especially of singleton variants. In addition, it is more commonly used in clinical variant calling, making this approach more compatible with existing clinical pipelines and reducing the risk of technical bias by using multiple variant callers. Based on our simulations, we expect a pooled parent strategy to provide a useful, cheaper alternative to trio sequencing for *de novo* cases. For example, by simply pooling the two parents we get 98% of variants recalled with a low false discovery rate, even with ten in pool we calculated an average recall of 94% (Appendix Table 6.1). We found that increasing the sequencing depth can also increase the recall rate.

2.3.3 Calling variants from real pools

We performed a real pooled-parent sequencing experiment. We had previously exome sequenced four individual probands likely to have *de novo* disease that were still unsolved after the initial variant analysis. We then performed exome sequencing of a pool of all eight of their parents. For the probands, exome capture was performed with the Nextera v1.2 Rapid Exome Capture Kit, and all the libraries were sequenced to ~100x depth over the target region. The parent pool was captured using the Agilent SureSelect QXT Clinical Research Exome kit and sequenced to a median on-target depth of 119x (or ~15x per parent). While the probands and parents were sequenced using different exome captures, the SureSelect Clinical exome is much larger than Nextera, and it mostly covers the same regions so should be able to recover most positions called in the probands. Based on the results of our simulation study we performed variant calling using GATK HaplotypeCaller on each of the samples individually. The probands were genotyped with default (diploid) ploidy and the pool with ploidy set to 16. We then used the

parent pool variant calls to filter out variants in the probands. We additionally performed variant annotation with VEP and added gnomAD frequencies (see Methods).

We calculated recall on this real data set as the percentage of variants found in all probands that were also called in the pool. Our calculated recall rate for all variants was 81.3%. This is a little lower than what we expected based on our simulations (Appendix Figure 6.1). The recall rate for variants with one allele found in one proband was 72.6%. This is consistent with the singleton rate observed in simulations (69.4%), however it should be noted that for the real pooled experiment we would expect each proband to have half the genetic material from each parent. Therefore, a variant found once in the probands could occur more than once in the parent pool if the variant also appeared in the untransmitted allele. Hence if a variant is found once across all the probands, it could plausibly have an allele frequency of anywhere from 1/16 to 9/16 in the parent pool and so is not directly comparable with the simulations.

We next used gnomAD to prioritise rare variants by filtering out variants with allele frequency over 0.0005. Of the remaining variants the recall in the pooled sample was 50.8%. This shows while both strategies filter many of the same variants, the parent pool provides substantial filtering beyond gnomAD.

We further assessed the power of using variants called in the parent pool as a strategy for filtering variants in the probands. Here the goal is to minimise the number of variants that need to be curated for each individual by removing those that are unlikely to cause *de novo* disease. We compare using the parent pool as a filter to some of the standard approaches, namely filtering low quality variants and common variants. Figure 2.2 summarises our variant filtering approach and shows the number of variants remaining after each filter is applied. We defined low quality variant calls as those with QD

(QUAL/DP) < 2 [140]. Since our probands have different phenotypes, we expect each to be caused by a different *de novo* variant. So, we filtered out variants observed more than twice across the probands, as unlikely to be causal. We also filtered out common variants to retain only very rare variants i.e. those observed at a frequency of greater than 0.0005 in gnomAD. We compared this to filtering out variants called in the parent pool. We found that filtering with the parent pool alone resulted in fewer variants than using the gnomAD frequencies alone. Importantly however, we found that the pooled parent filter was a complementary filter to other strategies and reduced the number of variants to less than 45% of the gnomAD only filter (Figure 2.2A). In particular we note that while gnomAD is useful to filter out variants observed frequently in the general population (specifically those populations included in gnomAD), the pool was able to filter out variants observed in the “private population” made up by these families. This may be particularly important when considering patients from populations that are not well represented in gnomAD. Most of our probands identified as European (Appendix Table 6.6), populations which are generally well represented in gnomAD. The gnomAD filter was slightly less effective for the Pacific Islander proband (Proband 3). 95.3% of all raw variants could be filtered using gnomAD for this individual, compared with 96.1-96.4% for the three European probands.

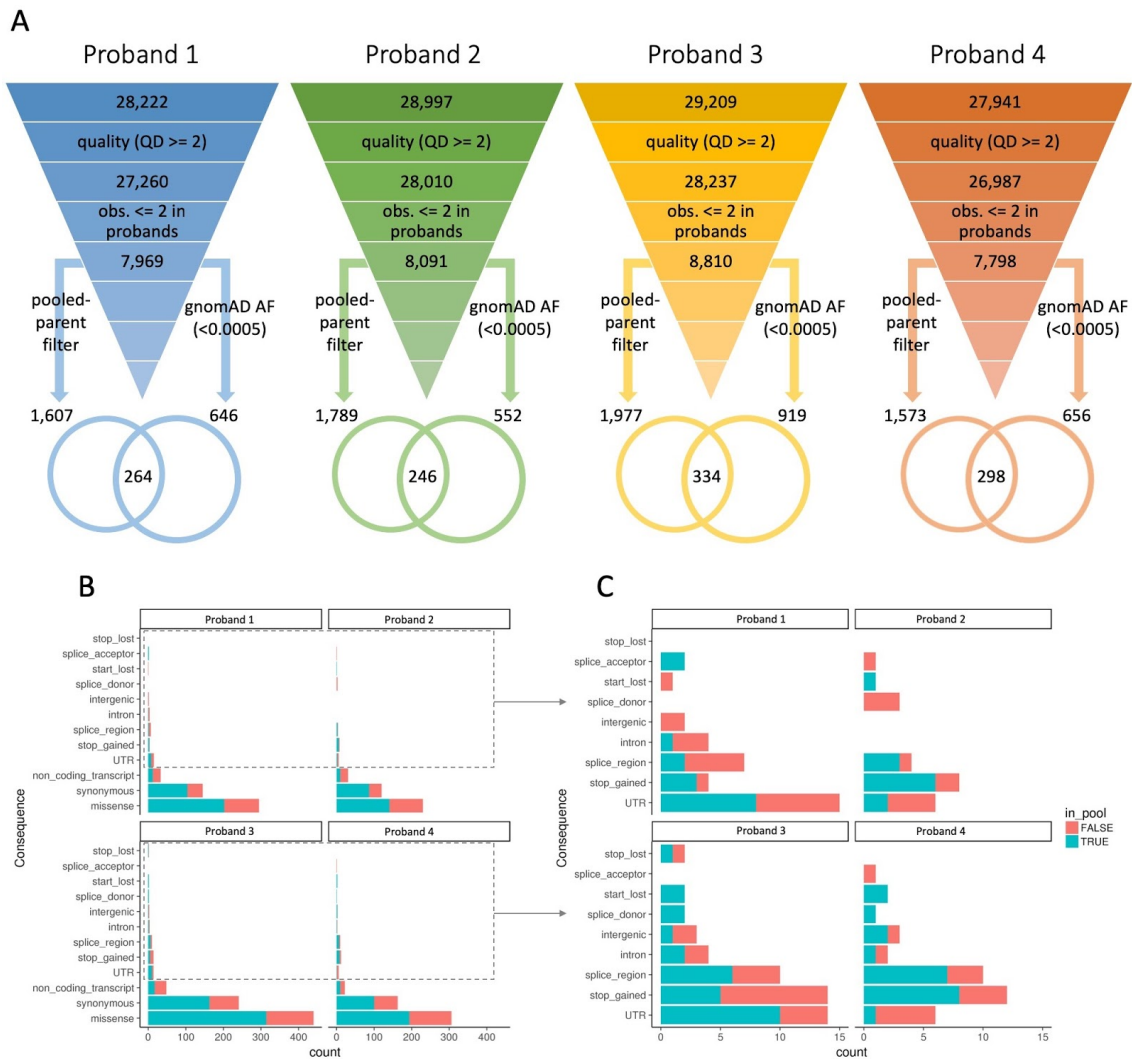


Figure 2.2: Overview of variants called in the real pooled parents experiment.

A) Schematic of variant filtering in each proband. Proband IDs are indicated above each inverted triangle. The first figure is the raw number of variants called in each proband. We then filter out low quality variants ($QD \geq 2$) and variants observed 2 or more times across the set of probands, in each case showing the remaining number of variants. At this point variants are filtered with either gnomAD allele frequency or the pooled parent variants, and then the intersection of these two filters is shown as a Venn diagram. **B)** VEP worst consequence annotations for the variants remaining after filtering by quality, frequency in the probands and gnomAD allele frequency. Variants that can be further filtered using the parent pool are indicated as “in_pool”. **C)** Magnification of lower frequency consequence categories indicated by the dotted rectangles in B.

We also performed annotation with Variant Effect Predictor (VEP) to aid interpretation of potential disease variants. Figure 2.2B and Figure 2.2C show the worst consequences annotated by VEP for each variant for only those variants that passed the quality and gnomAD frequency filters. The pooled

parent filter dramatically reduces the number of potentially deleterious variants that variant curators might need to consider. For example, in all of the probands the parent pool was able to filter out more than half of the missense variants, while in several probands it removed all start lost and splice site donor or acceptor variants.

The pooled parent strategy enabled us to filter out the majority of variants from the probands. Importantly it was complementary, removing a different set of variants to those filtered based on population allele frequency. It removed a number of potentially pathogenic variants, reducing the variant curation load.

2.4 Discussion

The pooled-parent exome sequencing strategy we propose here is a powerful and cost-effective way to prioritise *de novo* variants in the search for causal disease variants. Pooling DNA from parents is dramatically cheaper than full trio exome sequencing. While our simulations indicate that increasing pool size does reduce recall rates, the pooling strategy still allowed filtering of 94-98% of variants (for pool sizes of ten and two respectively). The reduced cost of this strategy allows funds to be reallocated to sequencing of more probands. We assessed variant calling strategies for pooled sequencing, and found that while the GATK HaplotypeCaller joint calling strategy provided the best recall rate overall, HaplotypeCaller individual calling had higher recall for the critical singleton variants. We also found that increasing the sequencing depth for pools was able to increase the recall rate, particularly for singleton variants. Generally increasing the sequencing depth is cheaper than performing additional exome captures. In a real analysis of four probands with undiagnosed likely *de novo* disease we were able to use a pooled-parent strategy to filter over 81% of variants. This strategy was complementary to population allele frequency filtering using gnomAD and resulted in reducing

the final list of variants to less than 54% compared to just using gnomAD. Unfortunately, the variants responsible for causing disease in these probands remain unknown.

One reason that the pooled-parent filtering strategy may perform particularly well when compared to population filtering, is that the parent pool is in essence an exquisitely matched population to the probands. The parents are the precise populations from which these probands arose and therefore is an excellent strategy for underrepresented populations. In contrast, gnomAD populations are weighted towards specific populations, particularly individuals of European ancestry [141]. If gnomAD is not a good representation of the population from which the proband arose, then the pooled-parent strategy may perform particularly well in comparison to population filtering. The gnomAD filter was less effective in the Pacific Islander proband compared to the European probands, while the pooled parent filter was more effective for this individual, supporting this hypothesis.

We have seen the cost of DNA sequencing decrease over time, while the cost of exome capture has remained relatively high, both in reagents and because it is a labour-intensive task. So, for exome sequencing the pooled-parent strategy is actually becoming increasingly cost-effective over time. However, as the cost of sequencing drops still further, clinicians may increasingly move to whole genome sequencing instead. For whole genome sequencing, the per-sample preparation is a relatively small proportion of the overall cost, so the economics of pooled-parent sequencing are not as compelling. Therefore, we expect the pooled-parent strategy to be most useful for exome and other targeted sequencing strategies such as smaller gene panels.

One limitation of this study is the simulations are performed on randomly selected individuals rather than trios. This does not truly reflect the pooling of parents, but rather a comparison of individual samples with those same

samples pooled together. However, having the individuals and the pools contain the same samples is a key advantage because the true allele frequencies of variants are known and this design allows false positives to be called. This was particularly useful when assessing the impact of increasing ploidy on the quality of variant calls and in evaluating different variant calling strategies.

2.5 Conclusions

Pooled-parent sequencing is a powerful strategy to filter out inherited variants to allow the analysis to focus on possible *de novo* variants. It is dramatically cheaper than full trio sequencing, allowing additional budget to sequence more probands. Importantly, our analysis shows the pooled parent variant filter is complementary to other standard approaches, in particular, filtering out different variants to using gnomAD population frequencies.

2.6 Methods

All code used for the simulations and analysis of these data sets can be found at <https://github.com/Oshlack/pooled-parents-paper>.

2.6.1 Simulating pools

To simulate pooled exome sequencing experiments of various sizes, we combined reads from a set of separately sequenced individuals. We selected parents from the Simons Simplex Collection [133]. These samples were chosen as the largest subset of this collection that were relatively technically homogeneous individuals within the publicly available sequencing data from this project. Specifically, this set is matched for DNA sequencer (Illumina GAIIx), read configuration (74 bp paired end reads) and exome capture kit (Nimblegen EZ Exome V2.0). We also removed samples with less than 64x median depth over the capture region in order to both remove low quality samples and to limit the range of sample depths such that $\text{min.depth}/(\text{min.depth}+\text{max.depth}) = 0.4$. I.e. if any two samples were combined then their reads would appear in a ratio between 40:60 and 50:50.

The final set used for simulation consisted of 111 samples, 52 females and 59 males, with median depths ranging from 64 to 97 (median 78). SRA run IDs are listed in Appendix Table 6.7. Re-use of public data was approved under HREC/49895/RCHM-2019.

Pipelines were written in Bpipe [142]. Raw FASTQ reads were downloaded for each of these samples, then aligned to `gatk.ucsc.hg19.fasta` with BWA MEM version 0.7.17 [143] and indexed with Samtools version 1.8 [144] (script: `genotype_individuals.groovy`).

In the constant depth simulation strategy, pools of two, four, six, eight and ten were simulated by selecting individuals at random from the 111 above, then randomly sampling a proportion of reads from the raw (not deduped) BAM files using `samtools view -s` (scripts: `pooled_sim_bpipe.groovy` and `pooled_sim_joint.groovy`). The proportions of reads were chosen such that the resulting pool would have twice the average depth of the source BAMs. The sampled bam files are merged with MergeSamFiles (Picard Tools version 2.18.11 [145]), then the read group and sample information from the original samples are removed using Picard `AddOrReplaceReadGroups` so that – for downstream processes - the BAM will appear to have originated from a single sample. These BAM files were deduplicated using Picard `MarkDuplicates`.

In the additive depth simulation strategy, pools were simulated so that the total sequencing depth for the pool is proportional to the number of samples. This was achieved by simply combining all reads for the samples in each pool. The simulation steps (scripts: `pooled_depth_sim_bpipe.groovy` and `pooled_depth_joint.groovy`) were the same as for the first simulation with the exception that instead of sampling reads from the individual BAMs, all randomly selected BAMs were merged together.

For both the constant and additive simulations we performed three replicates of each pool size using different random seeds and therefore different input

samples, resulting in a total of 30 bam files. The code for generating all these simulations can be found at <https://github.com/Oshlack/pooled-parents-paper>.

2.6.2 Variant calling and analysis

All variant calling was performed against hg19 and after deduplication with Picard Tools as above. For individual variant calling with GATK HaplotypeCaller version 4.0.10.1 [34], all individual samples were genotyped with default diploid ploidy, while for the pools, ploidy was set to two times the number of samples in the pool. Version 138 of dbSNP was used for all HaplotypeCaller commands. For joint calling, variants were called with GATK HaplotypeCaller -ERC GVCF to generate GVCFs, with default ploidy. Each pool GVCF was combined with the GVCFs of all the individual samples used to create that pool using GATK CombineGVCFs. Joint calling on the pool and its constituent samples was performed with GATK GenotypeGVCFs, to produce a final multi-sample VCF with genotype calls for loci that were called as variant in the pool or any of its individuals. FreeBayes version 1.2.0 [132] variant calling was performed on the individuals and pools from the constant simulation strategy only, as the high depth samples from the additive simulation caused excessive memory consumption. Individual calling was performed with default settings (script: `freebayes.groovy`), while for pooled variant calling (script: `freebayes_pool.groovy`) we set the relevant ploidy and ran in pooled-discrete mode with `use-best-n-alleles = 4`. FreeBayes reports all potential variants, including many of questionable quality, so in both the individuals and the pool we implemented the recommended `QUAL > 20` filter [146] using the `vcffilter` script included with FreeBayes.

To assess recall and false positive rates we compared variants called in each pool to the all variants called on the individuals that made up that pool. For HaplotypeCaller individual calling and FreeBayes we matched up specific

variants across VCF files by creating a unique string representing the position and reference/alternate alleles. To do this we created a variant identifier: CHROM_POS_REF_ALT (or ALT1/ALT2 etc if multi-allelic and use this to uniquely match variants across VCF files (filter_individualVCF.py). For the joint calling the pool and individuals were already represented at the same locus in a single VCF file, so we compared a variant across samples in the same VCF (filter_multiVCF.py). The recall rate is then calculated per pool as the number of alleles recalled in the pool divided by the total number of non-reference variants called in all individuals that made up that pool. If an allele was called in the pool but not in any of the individuals, we consider it to be a false positive. False positives rate is then the number of false positives called in the pool divided by the total number of non-reference alleles called in that pool. VCF parsing was accomplished using Python 3.6.8 and PyVCF version 0.6.8 [147] and further analysis and plots were generated using dplyr and ggplot2 in R version 3.5.0. Manipulation of exome capture target bed files was performed with Bedtools v2.27.1 [148].

2.6.3 Samples and sequencing

Four unsolved probands were selected from the Melbourne Genomics Health Alliance Childhood Syndromes project [3] as being good candidates for dominant *de novo* disease based on clinical assessment. This project was approved under Human Research Ethics Committee approval 13/MH/326. Parents provided written informed consent after genetic counselling regarding the testing. As part of the demonstration project these patients received exome sequencing alongside standard of care. DNA was extracted from peripheral blood, and exome sequenced used Nextera v1.2 Rapid Exome Capture Kit on a HiSeq 2500 at the Australian Genome Research Facility to 100X to a median on-target depth of 100x. We additionally sequenced a pool of all eight of these probands' parents using the Agilent SureSelect QXT Clinical Research Exome capture kit, and 151 bp paired end reads on an

Illumina HiSeq4000 to a total median on-target depth of 119x, or on average ~15x per parent. Parent DNA samples were quantified and pooled in approximately equimolar concentrations before exome capture.

2.6.4 Analysis of real pools

The probands and pool were analysed using a similar strategy to the simulations above (scripts: `genotype_individuals.groovy` and `pooled_joint_analysis.groovy`). Reads were aligned to `gatk.ucsc.hg19.fasta` with BWA MEM, indexed and deduplicated. Variant calling was performed with GATK HaplotypeCaller. Each sample was genotyped individually: the probands with default ploidy, and the pool with ploidy = 16. Proband variants were annotated with allele frequencies from gnomAD version r2.0.2 [37,47] using `vcfanno` version 0.2.9 [149]. VEP was also used to annotate the most severe consequence for each variant.

As for the simulations, we calculated recall as the number of alleles called in any proband that were also called in the pool over all genomic regions (script: `filter_individualVCF.py`). Multiallelic variants were split to allow individual annotation with gnomAD allele frequencies and VEP consequences. We performed a series of filtering steps. We performed light filtering for variant quality, by filtering out variants with $QD < 2$ ($QD = QUAL/DP$), as recommended by the GATK documentation [140]. We also removed variants observed in more than one proband, as they have different diseases so these shared variants are unlikely to be causal. We removed variants with an allele frequency of more than 0.0005 in gnomAD. Finally, we filtered out any variants called in the parent pool. Before examining any individual variants in detail, we excluded a set of genes known to cause high penetrance early onset disease to avoid secondary findings in line with Melbourne Genomics ethics requirements (Appendix Table 6.8).

Chapter 3 STRetch: detecting and discovering pathogenic short tandem repeat expansions

This chapter contains the publication [150]:

Dashnow, H. *et al.* STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* **19**, 121 (2018).

Supplementary materials can be found in Appendix B.

Acknowledgement


I would like to thank my collaborators for all their contributions to this paper. In particular, I would like to thank Alicia Oshlack, for her guidance and advice in developing the STRetch algorithm. Belinda Phipson and Simon Sadedin contributed to algorithm development. Andreas Halman helped perform data analysis. Andreas, Simon and Andrew Lonsdale helped with software testing. Mark Davis, Nigel G. Laing and Joshua S. Clayton performed experimental validation. Nigel and Phillipa Lamont contributed patient samples. Monkol Lek and Daniel G. MacArthur provided reference samples. Daniel provided sequencing and computational resources. Monkol, Belinda, Andreas, Andrew, Mark, Joshua, Nigel, Daniel and Alicia contributed to the manuscript.

METHOD

Open Access



STRetch: detecting and discovering pathogenic short tandem repeat expansions

Harriet Dashnow^{1,2}, Monkol Lek^{3,4}, Belinda Phipson¹, Andreas Halman^{1,5}, Simon Sadedin¹, Andrew Lonsdale¹, Mark Davis⁶, Phillipa Lamont⁷, Joshua S. Clayton⁸, Nigel G. Laing⁸, Daniel G. MacArthur^{3,4} and Alicia Oshlack^{1,2*} 

Abstract

Short tandem repeat (STR) expansions have been identified as the causal DNA mutation in dozens of Mendelian diseases. Most existing tools for detecting STR variation with short reads do so within the read length and so are unable to detect the majority of pathogenic expansions. Here we present STRetch, a new genome-wide method to scan for STR expansions at all loci across the human genome. We demonstrate the use of STRetch for detecting STR expansions using short-read whole-genome sequencing data at known pathogenic loci as well as novel STR loci. STRetch is open source software, available from github.com/Oshlack/STRetch.

Background

Short tandem repeats (STRs), also known as microsatellites, are a set of short (1–6 bp) DNA sequences repeated consecutively. Approximately 3% of the human genome consists of STRs [1]. These loci are prone to frequent mutations and high polymorphism, with mutation rates 10–100,000 times higher than average rates throughout the genome [2]. Dozens of neurological and developmental disorders have been attributed to STR expansions [3]. STRs have also been associated with a range of functions such as DNA replication and repair, chromatin organization, and regulation of gene expression [2, 4, 5].

STR expansions have been identified as the causal DNA mutation in almost 30 Mendelian human diseases [6]. Many of these conditions affect the nervous system, including Huntington's disease, spinocerebellar ataxias, spinal-bulbar muscular atrophy, Friedreich's ataxia, fragile X syndrome, and polyalanine disorders [7]. Most tandem repeat expansion disorders show dominant inheritance, with disease mechanisms varying from expansion of a peptide repeat and subsequent disruption

of protein function or stability, to aberrant regulation of gene expression [8].

STR expansion diseases typically show genetic anticipation, characterized by greater severity and earlier age of onset as the tandem repeat expands through the generations [9]. In many STR diseases, the probability that a given individual is affected increases with the repeat length. In some cases severity also depends on the gender of the parent who transmitted the repeat expansion [10]. The number and position of imperfect repeat units also influences the stability of the allele through generations [9]. Together, these features can be used to identify patients with a disease of unknown genetic basis that might be caused by an STR expansion.

Historically, STRs have been genotyped using polymerase chain reaction (PCR) and gel electrophoresis. In such cases, PCR is performed using primers complementary to unique sequences flanking the STR. The PCR product is then run on a capillary electrophoresis gel to determine its size. Although this method has been scaled to handle dozens of samples, it is still labor-intensive and costly. Each new STR locus to be genotyped requires the design and testing of a new set of PCR primers, along with control samples.

A number of diseases are known to be caused by any one of multiple variants, including STR expansions, single nucleotide variants (SNVs), or short indels. For

* Correspondence: alicia.oshlack@mcri.edu.au

¹Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, VIC, Australia

²School of Biosciences, The University of Melbourne, Parkville, VIC, Australia
Full list of author information is available at the end of the article



© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

example, there are > 10 STR loci in as many genes that are known to cause ataxia [11], as well as SNVs and indels in dozens of genes [12]. For such diseases, this can mean hundreds of dollars spent per STR locus, plus additional costs for SNV and short indel testing. For such conditions, there is a clear need for a single genomic test that can detect all relevant disease variants including SNVs, indels, and STRs.

The ability to genotype STRs directly from next-generation sequencing (NGS) data has the potential to reduce both the time and cost to reaching diagnosis and to discover new causal STR loci. It is becoming increasingly common to sequence the genomes or exomes of patients with undiagnosed genetic disorders. Currently, the analysis of these data is focused on SNVs and short indels, and while NGS has identified hundreds of new disease-causing genes, to our knowledge no new pathogenic STR expansions have been discovered. STRs are generally only investigated in an ad-hoc manner at known loci if they are a common cause of the clinical phenotype. The ability to screen for STR expansions in NGS data gives the potential to perform disease variant discovery in those patients for which no known pathogenic variants are found.

The vast majority of current STR genotyping tools for short-read sequencing data (most notably LobSTR [13], HipSTR [14], and RepeatSeq [15]) are designed to look at normal population variation by looking for insertions and deletions within reads that completely span the STR. These tools are limited to genotyping alleles that are less than the read length and require sufficient unique flanking sequence to allow them to be mapped correctly. However, for most STR loci causing Mendelian disease in humans, pathogenic alleles typically exceed 100 bp, with pathogenic alleles at some loci in the range of 1000–10,000 bp [16], far exceeding the size cut-off for detection using these algorithms.

One STR genotyper, STRViper [17], which is also designed to detect population variation, has the potential to detect alleles exceeding the read length by looking for shifts in the distribution of insert size from paired reads. This method requires that the insert size distribution has a relatively low standard deviation (SD) and is limited to repeats smaller than the insert size with enough flanking sequence to map both pairs to the reference genome. The mean insert size can be as low as 300–400 bp, meaning that for many large pathogenic expansions, there may be very few or no spanning read-pairs. Such methods would therefore have limited utility for the detection of allele sizes expected for pathogenic expansions. Another tool, ExpansionHunter [18], uses read-pair information and recovery of mis-mapped reads to estimate the length of STRs. For known pathogenic sites, ExpansionHunter can be used to determine if the

length of the STR is in the pathogenic range. This tool was originally developed for the FTDALS1 repeat and only works on a specific set of pre-defined loci and is therefore not a genome-wide method. Similarly, exSTRa [19] detects expansions in a set of 21 pre-defined loci and requires a set of matched control samples in order to define the statistical probability of an expansion. In contrast to ExpansionHunter, the exSTRa method does not attempt to estimate allele lengths.

While long-read sequencing technologies can potentially sequence through larger repeat loci [20], they are currently far too expensive for clinical use. High error rates and low throughput also make these technologies less suitable for genotyping SNVs and short indels and are thus a poor alternative to short-read sequencing in a clinical setting. Clearly, there is still a great need to be able to detect STR expansions from short read data.

Here we present STRetch, a new method to detect rare expansions at every STR locus in the genome and estimate their approximate size directly from short-read sequencing. We show that STRetch can detect pathogenic STR expansions in short-read PCR-free whole-genome sequencing (WGS) data and can detect expansions at STR loci not known to be pathogenic. We also demonstrate the application of STRetch to solve cases of patients with undiagnosed disease, in which STR expansions are a likely cause.

STRetch is open source software, available from github.com/Oshlack/STRetch.

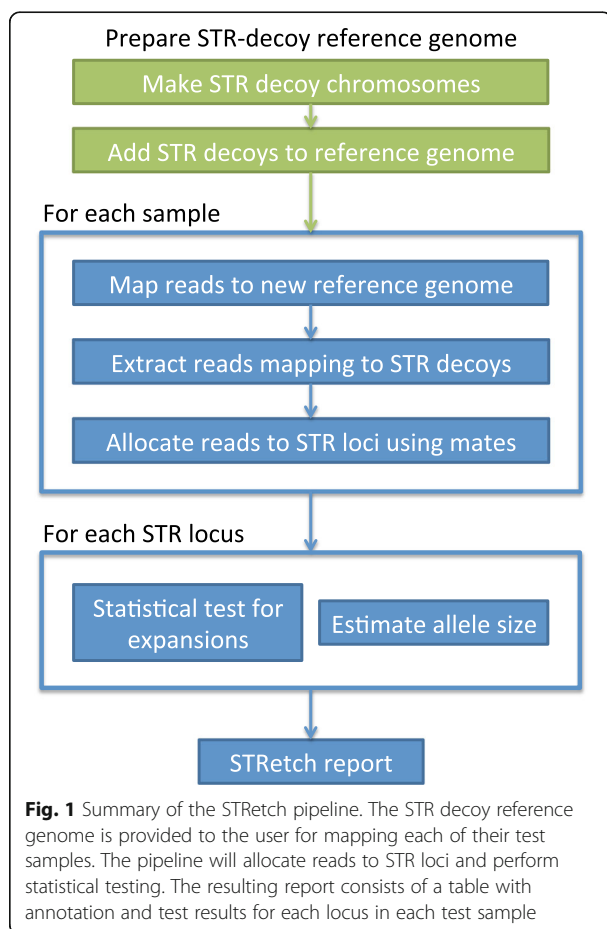
Results

The STRetch method

The STRetch method has been designed to identify expanded STRs from short-read sequencing data and give approximate sizes for these alleles. Briefly, the idea behind STRetch is to first construct a set of additional sequences comprising all possible STR repeat units in the range of 1–6 bp. These are then added to the reference genome as “STR decoy chromosomes.” By mapping to this modified genome, STRetch identifies reads that originated from large STR expansions, containing mostly STR sequence, that now preferentially map to the STR decoys. These reads are then allocated back to the genomic STR locus using read-pair information and the locus is assessed for an expansion using a statistical test based on coverage of the STR. A summary of the STRetch method is presented in Fig. 1, with further specifics detailed below.

STR decoy chromosomes: generating an STR-aware reference genome

A key feature of STRetch is the generation of STR decoy chromosomes to produce a custom STR-aware reference genome. Most aligners have difficulty accurately mapping reads containing long STRs. For example, although



the BWA-MEM algorithm has superior performance for mapping reads containing STRs [21], reads containing long STRs sometimes map to other STR loci with the same repeat unit or completely fail to map [18]. The systematic mis-mapping of STR reads is unsurprising considering that BWA-MEM is optimized to find the longest exact match [22]. For a read made up primarily of STR sequence, the best match is likely to be the longest STR locus in the reference genome with the same repeat unit.

STRetch takes the issue of systematically mis-mapped or unmapped STR reads and uses this as a way to identify reads that contain long STR sequence. To achieve this, we introduce the concept of STR decoy chromosomes. These are sequences that consist of 2000 bp of pure STR repeat units that can be added to any reference genome as additional chromosomes. STR decoy chromosomes for all possible STR repeat units in the range of 1–6 bp are generated and filtered for redundancy, resulting in 501 new chromosomes that are added to the reference genome (STRetch provides hg19 with STR decoys, see “Methods”). While reads with STR lengths similar to the allele length in the reference

genome will map to their original locus, reads containing large STR expansions will preferentially align to the STR decoy chromosomes. These reads are then further examined for evidence of a pathogenic expansion.

Mapping to STR decoys to identify reads containing STRs

Once the new STR decoy reference genome is created, the first step maps reads against the new reference genome using BWA-MEM. If the data have already been mapped, STRetch can optionally extract and re-map a subset of reads likely to contain STRs. Extracted reads are those that aligned to known STR loci (defined using Tandem Repeats Finder (TRF) [23]), as well as any unmapped reads (see “Methods”). Any reads mapping to the STR decoy chromosomes are inferred to have originated from an STR. Typically ~0.01% of reads map to the STR decoy chromosomes in a PCR-free whole genome.

Determining the origin of STR reads

Next, the reads that map to the STR decoys are assigned to genomic STR positions. STRetch uses the mapping position of the read at the other end of the DNA fragment (the mate read) to infer from which STR locus each read originates. Known STR loci are obtained from a TRF annotation of the reference genome. For a given read, if the mate maps within 500 bp of a known STR locus with the same repeat unit, then the read is assigned to that locus (or the closest matching locus if multiple loci are present). Only 0.93% of STR loci are within 500 bp of another STR locus with the same repeat unit (Additional file 1: Figure S1). This distance accounts for the fragment length of the majority of reads (Additional file 1: Figure S2).

After all possible reads are assigned, there may be a difference between the number of reads mapping to a given STR decoy chromosome and the number of reads assigned to all STR loci with that same repeat unit. Unassigned STR reads can occur for a variety of reasons; for example, if their mate also maps to the STR decoy chromosome, if their mate is unmapped, or if their mate does not map in close proximity to a known locus. The number of unassigned reads will increase in samples with very large STR expansions because more read-pairs will originate purely from the STR. This may result in STRetch underestimating the size of very large alleles; however, such loci will still be reported as significant as they are still assigned substantially more reads compared with control samples.

Detecting outlier STR loci

STRetch next uses a statistical test to identify loci where an individual has an unusually large STR. Specifically, STRetch compares the number of STR decoy reads

assigned to each locus for a test sample with STR reads from a set of control samples. At each locus the read counts are normalized by dividing by the average coverage of the sample. The set of control samples provide a median and variance of counts for each locus. A statistically robust z-score (“outlier score”) is then used to test if the log-normalized number of reads in the test sample is an outlier compared to controls (see “Methods”). The final result is a multiple-testing-adjusted p value describing the significance of an expansion at each locus relative to control samples. Every locus for which reads have been assigned is given a p value.

A variety of control samples can be used in the statistical testing. First, STRetch can be run on a set of controls and the median and variance of the coverage parameters for each locus estimated. These control parameters are then used in testing for significant expansions in disease samples. This approach is ideal for researchers who have access to large sets of sequenced controls. Second, a set of samples that are all being tested can be used as controls for each other in a similar way to the above. The assumption here is that only a small proportion of samples (< 50%) will have the same expanded locus (we refer to this approach as internal controls). Third, STRetch supplies median and variance parameters estimated from a reference set of PCR-free whole genomes (see “STRetch reveals STR expansions in 97 whole genomes”). This third approach is useful for researchers who only have a limited number of samples to test. The advantage of the first and second options is that the sequencing is usually run at the same center with the same library preparation protocols. However, the datasets may be smaller and therefore provide less robust statistical measures. In most cases, option three (using the supplied parameters from PCR-free whole genomes) is the easiest and most accurate when using relatively small datasets (see Table 1).

Estimating the size of STR alleles

When scanning the genome for sites of significant expansions, the statistical test is the most important screen. However, for known disease loci the literature has focused on the length of the variant that is associated with pathogenicity. Therefore, STRetch also makes estimates of allele length. STRetch works on the assumption that, for a given locus, the number of reads containing the STR repeat unit is proportional to the length of the repeat in the genome being sequenced. This is because increasing the length of the STR allele increases the likelihood that reads from that locus will be sampled. Hence, STRetch estimates the size of any detected expansion using the normalized read counts allocated to that STR locus. Using simulation, we indeed found that the allocated read counts are linearly related

to the length. Specifically, we simulated reads from 100 individuals with the genotype $16 \times \text{CAG}/N \times \text{CAG}$ at the SCA8 locus, where N was randomly selected in the range of 0–500 (see “Methods”). Our simulated data exhibit a linear relationship between allele size and the number of reads mapping to the STR decoy chromosome (Additional file 1: Figure S3). We use the slope and offsets from these simulated expansions in estimating the allele size from the normalized coverage at a locus.

Output files

On completion, STRetch generates a tab-delimited output file for each sample that contains all STR loci for which STR decoy reads were detected. Further information includes p values for statistical significance of an expansion, details of the STR locus (position, repeat unit, size in reference), robust outlier z-score, locus read count, and the allele length estimate. By default, this file is sorted such that the most significant expansions are ranked at the top.

STRetch is able to recover true pathogenic expansions

In order to test STRetch, we generated PCR-free WGS on ten individuals: nine with known pathogenic STR expansions and one unaffected family member. Samples were sequenced to a mean coverage of $41.74 \times$ (range $38.35\text{--}49.57 \times$), then processed using the Broad GATK pipeline (mapped to hg38 with BWA-MEM, then processed using the GATK best practices).

For analysis with STRetch, we first extracted reads overlapping all known STR loci annotated by TRF (see “Methods”) and then processed these reads through the STRetch pipeline, using the hg19 reference genome. The STRetch statistics were calculated twice; first using only these ten samples as controls for each other (“internal control”) and then using the 97 WGS samples described below as controls (“reference control”). We also ran LobSTR/HipSTR to estimate the size of the short allele in each case. On average, STRetch reported 18 significant STR expansions per sample (range 4–33).

For six of the ten samples we had information about the disease and the estimated allele size by PCR. For the other four samples (Samples 7, 8, 9, and 10), we were initially completely blinded to all patient information, including phenotype. Disease and allele size estimates were only revealed to us after we had correctly identified the causal STR expansion in each case. Table 1 summarizes the results of this analysis, while Fig. 2 and Additional file 1: Table S1 show allele size estimates for these samples using STRetch, LobSTR, HipSTR, and ExpansionHunter.

For the six samples with known information, STRetch correctly identified three true-positive expansions.

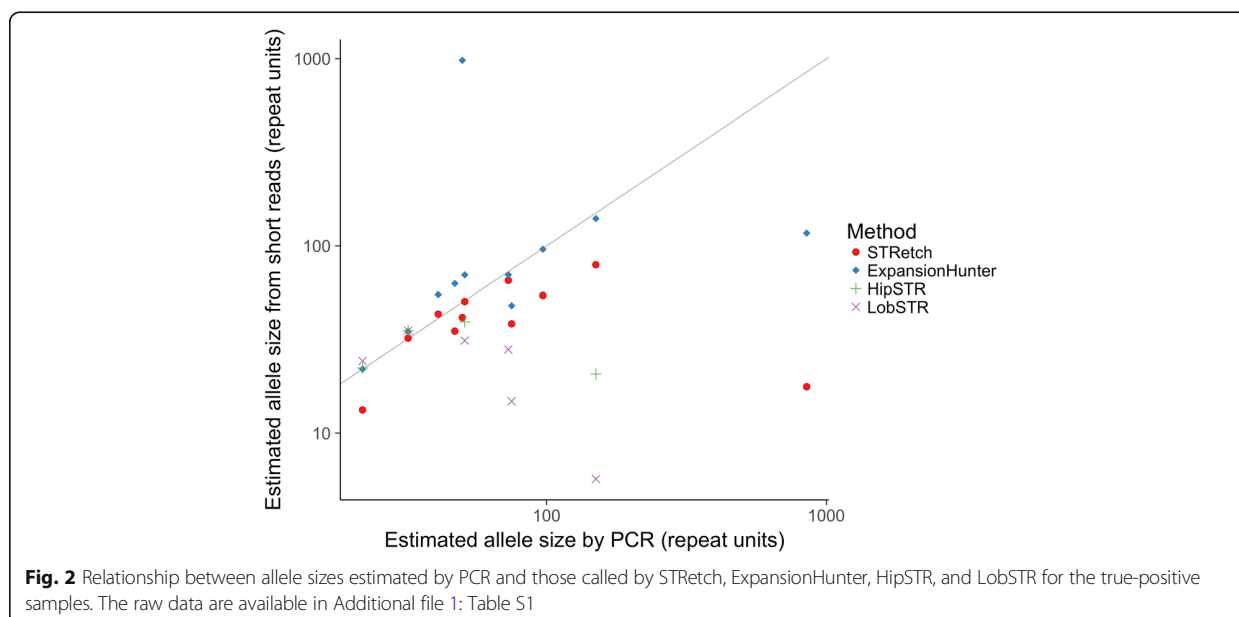
Table 1 Summary of the ten individuals with known STR alleles

Sample	Disease	Gene	Repeat unit	Type	Position	Allele ref	Pathogenic range	Allele PCR	Allele STRetch	Rank (internal control)	p value (internal control)	Rank (reference control)	p value (reference control)
1	SCA1	ATXN1	CAG	Coding	chr6:16327865-16327955	30.3	39-70	51	50.5	1	4.71E-25	1	2.46E-14
2	Unaffected relative of Sample 1	ATXN1	CAG	Coding	chr6:16327865-16327955	30.3	39-70	29/32	32.2	-	-	-	-
3	SCA3	ATXN3	CAG	Coding	chr14:92537355-92537396	14	60-87 (≥52)	73	65.4	793	0.22	3	4.15E-09
4	SCA6	CACNA1A	CAG	Coding	chr19:13318673-13318712	13.3	20-33 (≥19)	22	No call-0 STR reads in all samples	-	-	-	-
5	SBMA	AR	CAG	Coding	chrX:66765159-66765261	33.3	≥38 (≥36)	41	43.3	9	6.90E-09	11	5.19E-05
6	SBMA	AR	CAG	Coding	chrX:66765159-66765261	33.3	≥38 (≥36)	47	35.1 (0 STR reads)	-	-	-	-
7	FTDALS1	C9orf72	GGGGCC	Intronic	chr9:27573482-27573544	10.8	>60 (>30)	>50	41.5	959	0.6	5	1.20E-06
8	DM2	ZNF9	CCTG	Intronic	chr3:128891419-128891502	20.8	75-11,000	>75	38.4 ^a	1	3.84E-23	1	4.05E-16
9	DM1	DMPK	CAG	Non-coding	chr19:46273462-46273524	20.7	>50	>150	79.3	1	3.24E-48	1	1.13E-32
10	FRDA	FXN	GAA	Intronic	chr9:71652203-71652205	6	66-1300 (≥44)	~850	17.7 ^b	2 ^b	1.83e-08 ^b	NA	NA

Allele refers to the number of repeat units. Rank refers to the position of the locus when all the STR loci are ordered based on p value. Pathogenic range is as reported in GeneReviews [27]. Where there is dispute on the pathogenic range or incomplete penetrance, the alternate range is given in parentheses

^aThe DM2 locus is a complex repeat: (TG)n(TCTG)n(CCTG)n, where only expansion of the CCTG repeat causes DM2. However, all repeat units are polymorphic. STRetch estimates CCTG expansions directly, while the PCR measures the entire complex locus

^bResults after manual addition of the FRDA STR to the reference data



Furthermore, STRetch correctly failed to detect an expansion in a pathogenic locus in the true negative (Sample 2). STRetch failed to identify the causal locus in two cases (Samples 4 and 6). For Sample 4, the PCR allele length is only 26 bp larger than the reference, making the entire allele 66 bp, which is well within the read length of 150 bp. STRetch only detects STR expansions that are sufficiently large so that the repeat maps to the STR decoys instead of the reference locus. Indeed, for Sample 4, we see three reads mapped to the genomic locus with a 27-bp insertion and no reads from this locus mapping to the STR decoy. This allele can be detected by tools that look for indels within the read, and indeed both LobSTR and HipSTR are able to correctly call this expansion (Additional file 1: Table S1). Sample 6 was found to have lower coverage over the STR region compared with other samples. However, some evidence of the repeat was observed, such as reads ending in the STR with more repeat units than the reference. As such, this expansion may be detected in future iterations of the software.

For the blinded samples (Samples 7, 8, and 9), we were able to correctly determine the causal STR locus simply by ranking variants by their p values and then looking for any known pathogenic STR loci with a significant p value (Table 1).

For Sample 10, we were initially unable to identify a significant expansion in a known pathogenic gene. After the variant was revealed to be a large GAA expansion at the FRDA locus, we investigated further and discovered that although the reference genome has 6×GAA at this position, the STR is missing from the TRF genome

annotation. TRF failed to annotate this STR due to its relatively small size. After manually adding this locus to the genome annotation and rerunning the analysis on the ten samples in Table 1, STRetch was able to correctly detect a significant expansion at this locus. Note that this locus was not annotated in the control samples so only an “internal control” analysis was done.

Most of the true positives that were detected had significant expansions when using both the 97 reference controls and the ten internal controls. However, there were two samples (Samples 3 and 7) that were only significant when using the much larger set of reference controls. Therefore, for this size dataset, the use of reference controls generally provides more power.

We compared the PCR-sized expansion in these ten samples with the STRetch length estimates and found that STRetch has about a ~20% standard error. We also found that for very long alleles, STRetch substantially underestimates the allele size. One likely explanation is that the current implementation of STRetch is limited by the insert size of the sequencing data. Some alleles will be so large (e.g. DM1 in Sample 9 is >450 bp) that there will be read-pairs where both are completely contained within the STR. In such cases, both pairs will map to the STR decoy and will not be assigned to an STR locus, leading to underestimation of the read length (Additional file 1: Figure S4). In comparison, ExpansionHunter more closely estimated alleles in most cases, with a tendency to overestimate allele sizes (Additional file 1: Table S1). As expected, both LobSTR and HipSTR dramatically underestimated large alleles or did not make a call in many cases.

STRViper was also run on the true-positive samples and reported no significant expansions across all STR

loci annotated with TRF. This is likely due to the large SD of insert size for these samples (130–150 bp), which is typical for the current standard Illumina protocol. STRViper requires smaller insert sizes of approximately 30 bp or less.

We next assessed the sensitivity, specificity, and false discovery rate (FDR) of STRetch on these samples. We considered STRetch calls in 18 samples for which we had DNA available in our lab (the ten test samples described above and eight samples from the 97 controls described below) on a set of 22 pathogenic loci (Additional file 1: Table S2). This gave a test set of 396 loci. We assumed that any significant STRetch calls ($p < 0.05$ after multiple testing correction), beyond the true positive loci described above, are false positives and the two loci that STRetch failed to detect above are false negatives. Any loci where STRetch does not make a significant call are assumed to be true negatives. Of these true negatives, 64 were confirmed by PCR from standard diagnostic testing. Using these assumptions, our measured sensitivity was 0.778, specificity was 0.974, and the FDR was 0.025. We also ran ExpansionHunter on these samples and found that three of the STRetch calls we had assumed to be false positives were also supported as expanded by ExpansionHunter, although not in the pathogenic range. Three of these (the HD expansion and the two SCA3 expansions) were also confirmed as expanded in the non-pathogenic range by diagnostic PCR. Using these updated values, sensitivity was 0.833, specificity was 0.974, and FDR was 0.018. Six of our ten false-positive calls were in the FTDALS locus and another two in the SCA3 locus (although these were expanded, just not pathogenic). This indicates that some loci may be harder to correctly identify than others. Indeed, the FTDALS1 locus is known to be difficult to analyze due to homology with other loci [18]. Overall, STRetch achieves FDRs well below our nominal value of 0.05.

STRetch reveals STR expansions in 97 whole genomes

We performed PCR-free WGS on a set of 97 individuals, most of whom were being investigated for the cause of their Mendelian disease or were immediate family members of such patients. Many of these cases have inherited neuromuscular disorders. Approximately half of these cases have been previously solved, with the causal variant identified as a SNV or indel. The remaining cases are still unsolved. The set also includes four individuals with ataxia. In addition to patient samples and relatives, there are seven unaffected samples including NA12878. Given its enrichment for individuals with Mendelian disease, this control set may contain pathogenic STR expansions. However, we did not expect a large proportion of samples to have the same STR expansion so our assumptions for the statistical tests were highly unlikely to be violated.

All samples had previously been mapped with BWA-MEM against hg19. We used STRetch to extract reads from all annotated STR loci, as well as unmapped reads, and re-mapped these against the hg19 STR decoy genome (see “Methods”). We then proceeded with the rest of the STRetch pipeline. The median and SD were recorded for each locus across all individuals for use as a control set for subsequent analyses. This analysis showed that homopolymer loci are the most variable between individuals, showing dramatically higher SDs, followed by STRs with 5, 6, 3, 4, and then 2 bp repeat units 7 (Additional file 1: Figure S5).

To assess the frequency of expansions at known pathogenic STR loci, we filtered the STRetch results to those significant expansions intersecting with a set of 22 known pathogenic STR loci (Additional file 1: Table S3). We observed 29 significant STR expansions in seven pathogenic loci (summarized in Table 2). Although these are significantly expanded compared to the rest of the control set, their pathogenicity is uncertain as the allele size estimates are only approximate and are often well below the defined pathogenic range.

Nonetheless, a number of the STR expansions are potential candidates for follow-up if the sequenced individuals have a relevant phenotype. STRetch detected a large SCA8 expansion in two individuals; one of whom is an ataxia patient (see below). We also detected a DM2 expansion in another individual. All three variants were highly significantly expanded compared to the other control samples ($p = 4.2e-24$, $8.2e-23$ and $1.5e-10$, respectively) and each was ranked as the most significant for that individual. We have referred these variants back to the originating laboratories to determine whether the variants fit the phenotype and can be validated.

STRetch also identified STR expansions in a number of other pathogenic loci in these samples; however, many of these are unlikely to be sufficiently large to cause disease. Two SBMA expansions were on the limit of detection and significance (ranked 109 and 476, $p = 0.001$ and 0.04 , respectively). We detected SCA36 and FXTAS expansions, which likely reflect sub-clinical variation at

Table 2 Summary of significant expansions in STR disease loci in 97 WGS samples

Disease	Gene	Number of individuals
SCA8	ATXN8/ATXN8OS	2
DM2	ZNF9	1
SBMA	AR	2
SCA36	NOP56	1
FXTAS	FMR1	1
SCA3/MJD	ATXN3	11
FTDALS1	C9orf72	11

these loci, with size estimates of 13.5×AGGCC and 34.6×CCG, respectively. We detected a surprising number of SCA3/MJD_ATXN3 expansions: 11 samples with estimated allele sizes in the range of 30.5–74.3×AGC. As $\geq 60\times$ AGC is considered pathogenic, with $\geq 52\times$ likely showing incomplete penetrance, many of these may be asymptomatic. However, we have observed STRetch to underestimate allele sizes at this locus so these variants could be larger than predicted. We also detected 11 FTDALS1 expansions in the range of 20.9–32.9×CCCCGG, all within the unaffected size range for this locus. Generally affected individuals have > 60 repeats. Also, smaller allele sizes are associated with later age of onset, with symptoms appearing as late as 80 years. Consequently, these results may indicate variation within the normal range, individuals with pre-mutations, individuals who have not yet shown symptoms, or false positives. ExpansionHunter confirmed all the SCA8, DM2, and FXTAS expansions, as well as one SBMA, one SCA3, and three of the FTDALS1 expansions. It failed to genotype the SCA36 locus.

These likely non-pathogenic variants at known pathogenic loci highlight the ability of STRetch to explore population variation and to better determine the true non-pathogenic range of these known pathogenic loci.

Genetic diagnosis and validation of an ataxia patient

As noted above, four ataxia patients were included in our 97 WGS cohort. These patients had previously been tested with for most known ataxia STR expansions, including SCA1, SCA2, SCA3, SCA6, SCA12, SCA17, SCA38, and DRPLA (see Controls 1–4 in Additional file 1: Table S2). The whole genome data had also been examined for causative SNVs and indels in candidate ataxia genes using the GATK Best Practices recommendations [24] and copy number variations using BreakDancer [25] and Genome STRiP [26] without success in diagnosis.

In one of these patients (Control 1), STRetch identified the highly significant expansion of SCA8 (noted above), a known but rare disease locus in ATXN8 with an outlier z-score of 11.11 ($p = 3.55e-24$) and an estimated allele size of 54.4×.

As a result, this SCA8 expansion was validated using a PCR assay, which confirmed a pathogenic CAG expansion with an allele size of 97×. In addition, an affected sibling, not sequenced using WGS, was tested for an expansion at the same locus and was similarly determined to have a pathogenic allele of $\sim 96\times$. The likely pathogenic range for this STR is 80–250 repeat units, with uncertain pathogenicity in the range of 50–70 repeat units. However, this locus shows incomplete penetrance, with cases of unaffected individuals observed at all allele sizes [27]. It is noteworthy that STRetch directly estimates the

size of the STR, while PCR assays at this locus also include the size of the adjacent non-pathogenic but highly polymorphic CTA repeat. As such, this estimate from STRetch should be interpreted as 65× when comparing to the PCR result.

We also ran LobSTR and HipSTR on the WGS data of this patient. At the expanded pathogenic locus, LobSTR called a homozygous 6×TGC insertion, while HipSTR called a homozygous indel from deletion of one TAC repeat unit upstream of SCA8 and an insertion of seven TGC repeat units, for a net total six repeat unit insertion. Both tools report a reference allele size of +15×, so the total allele size is 21× in both cases. PCR analysis of the proband and sibling indicated that both expansions were heterozygous with a short allele size of $\sim 29\times$. However, these sizes may not be directly comparable due to potential variation in the definition of the reference locus size.

STRetch can detect expansions at novel loci

To our knowledge, STRetch is the only method currently available that can use short-read sequencing to screen the entire genome for rare STR expansions. In order to demonstrate that STRetch can indeed detect expansions at loci not previously associated with disease, we performed validation using orthogonal technologies: PCR and Sanger sequencing and PacBio long-read sequencing.

First, we used PCR to estimate the size of an STR that STRetch called highly significant in Sample 5. The locus is a 5-base AACT repeat in an intron of the *MTHFD2* gene (chr2:74430970–74,431,055). This was the most significantly expanded locus in this sample with a p value of 4.81E-21 and a predicted size of 37 repeat units. We designed a PCR assay to genotype the size of the allele in this sample as well as five samples that were not predicted to contain this expansion by STRetch (Samples 1, 2, 6, 8, and 9 from above, Additional file 1: Figure S6). A standard control sample of unknown genotype (CEPH individual 1347–02) was also tested. Sanger sequencing yielded an allele size of > 59 repeat units in Sample 5 (the sequencing quality dropped off beyond this size), significantly larger than the alleles from the control samples ($p = 2.71E-03$, Additional file 1: Figure S7), as predicted by STRetch. By sizing the PCR product on a gel, we estimate the larger allele in this sample to be 62 repeat units. We configured ExpansionHunter to estimate the size of this locus resulting in a genotype of 47/64, much larger than the length of 17.2 repeats in the genome. While this STR is unlikely to be pathogenic, this result highlights the potential to use STRetch to screen for novel expansions that may be related to disease.

To further demonstrate the utility of STRetch to identify novel loci as expanded (relative to controls), we

compared our analysis of short-read data with a previously published genome analysis which utilized long-read PacBio data [28]. Specifically, we ran STRetch on Illumina short-read data from an artificial diploid sample created by combining the two haploid genomes, CH1 and CH13 (two replicates), and called significantly expanded loci compared with our 97 reference controls. We then compared all significant STRetch hits with variants called from long-read PacBio sequencing of the two haploid samples. Of the 18 significant STRetch calls (over two replicates) we were able to confirm 14 from the PacBio data as expanded (Additional file 1: Table S4). Of the four non-validated calls, two were in the non-standard chromosome chrUn_gl000220, which was not present in the PacBio variant data. The remaining two were both homopolymer A STRs. Homopolymers were excluded from the PacBio variant call set, preventing validation of these loci. While none of the loci represent likely pathogenic expansions, this analysis demonstrates that truly significant expanded variants can be found in novel loci using STRetch with a low FDR.

Discussion

STRetch is currently the only genome-wide method to scan for STR expansions using short-read WGS data. We have demonstrated the ability of STRetch to detect known pathogenic STR expansions in short-read PCR-free WGS data. STRetch correctly detected the pathogenic expansion in most of our ten test samples; seven true positives and one true negative were correctly identified; two expansions were missed. The size of one of the missed expansions was below the detection threshold for STRetch. The method performed well on these samples, with a sensitivity of 0.778, specificity of 0.974, and a FDR of 0.025. We applied STRetch to 97 PCR-free WGS samples to detect both potentially pathogenic STR expansions and expansions of moderate length in STR disease loci, where the allele size is likely below the threshold for disease. Importantly, this set of analyzed genomes can act as a control set and statistical parameters for the STR loci are provided to use in testing for expansions with STRetch. Within this cohort, STRetch revealed a previously undetected pathogenic STR expansion in SCA8 that was validated by PCR in the proband and an affected sibling.

As expected, we found that STR genotypers such as LobSTR and HipSTR, which are designed to genotype short STR variation, were unable to detect large pathogenic variants in WGS data. These tools instead called a homozygous genotype, corresponding to the size of the non-expanded allele, called a heterozygote with slight variation in the small allele, or failed to make a call. Using these tools in conjunction with STRetch allows the estimation of

the short allele in cases where STRetch has detected an expansion, allowing us to obtain a more complete picture of the genotype from short-read sequencing data.

ExpansionHunter performed relatively well when estimating allele sizes, although tended to overestimate, where STRetch tended to underestimate. However, a key limitation of ExpansionHunter is that it does not use a statistical basis for detecting significantly expanded loci and is not currently configured to estimate lengths across the genome; each locus of interest must be defined in a separate configuration file. Therefore, this cannot be used as a genome-wide scan for novel loci.

The main limitation of STRetch is its tendency to underestimate the allele size of STR expansions, especially for variants larger than the insert size. However, this limitation is mostly relevant for known pathogenic loci where there is already an established relationship between the length of the allele and disease characteristics. In genome-wide scans for rare expansions, we believe it is more important to use a statistical test that indicates the probability that an expansion exceeds the normal population variation. Estimating allele length by itself does not provide this information. Another limitation of STRetch is that it has not been tested on PCR+ or targeted sequencing data.

We have demonstrated the potential application of STRetch to discover a novel significant expansion at an intronic STR locus in *MTHFD2* and validated it by PCR and Sanger sequencing. In addition, further expansions detected by STRetch were also shown to be expanded using long-read sequencing data.

Conclusions

Here we have introduced STRetch, a method to test for rare STR expansions from WGS data. We have shown that STRetch can detect pathogenic STR expansions relevant to Mendelian disease with a low FDR. Although the emphasis has been on STRs known to cause Mendelian disease, a key advantage of STRetch over other methods is its genome-wide approach. STRetch performs statistical tests for expansions at every STR locus annotated in the reference genome and so has the potential to be used not only for diagnostics, but also in research to discover new disease-associated STR expansions. We hope the application of STRetch to WGS of patient cohorts will enable new discoveries of disease-causing STR expansions.

Methods

The STRetch pipeline

The STRetch pipeline is implemented using the Bpipe [29] pipeline framework (v0.9.9.3). This allows for a pipeline combining standard bioinformatics tools with novel scripts and is compatible with many

high-performance computing environments, allowing large-scale parallelization over multiple samples (see Additional file 1: Supplementary Methods).

To summarize the pipeline and components:

Reads are mapped to the reference genome with STR decoy chromosomes using BWA-MEM [22] (v0.7.12) and SAMtools [30] (v1.3.1). STRetch then counts the number of reads mapping to each STR decoy chromosome using bedtools [31] (v2.26.0). Reads mapping to the STR decoy chromosomes are allocated to an STR locus using paired information (Python v3.5.2 script: `identify_locus.py`). Median coverage over the whole genome or exome target region is calculated using `goleft covmed` [32] (v0.1.8), which is later used to normalize the counts. STRetch then predicts the size of the expansion using the number of reads allocated to the locus (Python script: `estimateSTR.py`).

The STRetch pipeline is freely available under an MIT license from github.com/Oshlack/STRetch.

Generating STR decoy chromosomes

To produce STR decoy chromosomes, STRetch generates a set of all possible STR repeat units in the range of 1–6 bp. These are then grouped by those repeat units that are equivalent as a circular permutation of each other or the reverse complement. For each group, the first repeat unit lexicographically was taken to represent that group, for example, CAG = AGC = GCA = CTG = TGC = GCT; the group is represented by AGC. STRetch filters out repeat units that could be represented by multiples of a shorter repeat unit. For example, ATAT would be filtered out as it is already represented by AT. This resulted in 501 unique repeat units. STRetch then uses an “STR decoy chromosome” for each repeat unit, which consists of the repeat unit repeated in tandem for 2000 bp (script: `decoy_STR.py`). This length was chosen to well exceed the insert size, but is configurable. The additional chromosomes can be added to any reference genome (hg19 with STR decoy chromosomes was used for all analyses).

Extracting likely STR read-pairs from aligned BAMs

In the case where reads have already been mapped to a reference genome, STRetch provides the option of extracting likely STR read-pairs from the BAM file for analysis, rather than remapping all reads in the sample.

In this case, STRetch defines a region where STR reads are likely to align by taking the TRF [23] annotation of the reference genome and expanding the region to include 800 bp of flanking sequence on each side. Reads aligned to this region, and all unmapped reads, are extracted using SAMtools `view`. These are sorted to place together read-pairs using SAMtools `collate` and then are extracted in `fastq` format using bedtools `bamtofastq`. Unpaired reads are discarded.

Aligning and allocating reads to STR loci

STRetch uses BWA-MEM to align reads to the custom reference genome. Any read mapping to the STR decoy chromosomes is presumed to have originated from an STR locus.

To determine from which STR locus the reads originated, the mates of the reads mapping to a given STR decoy chromosome are collected. If the mate maps within 500 bp of a known STR locus with the same repeat unit, it is assigned to that locus. If multiple loci fall in this range, the read is assigned to the closest locus. This distance was chosen because the mean insert size of our WGS data was 372–415 bp (Additional file 1: Figure S2), so we expect the mate to map within 500 bp or less of the STR locus. We found that increasing this value did not increase the number of reads allocated to true-positive loci. Another consideration when setting this parameter is the potential to misallocate reads from neighboring STR loci. To combat this, reads are only allocated to a locus with a matching repeat unit. Although rare, there are instances of two STR loci with the same repeat unit occurring close together. In hg19, 0.93% of STR loci are within 500 bp of another STR locus with the same repeat unit, 1.34% within 1000 bp, and 2.04% within 2000 bp (see Additional file 1: Figure S1). We used 500 bp as the smallest value of this parameter that allows detection of all relevant reads while minimizing inclusion of inappropriate reads. This value can be configured in the pipeline if required.

To correct for library size (total number of aligned reads), the counts for each STR locus are normalized against the median coverage for that sample. Counts are normalized to a median coverage of $100 \times \log_2(100 \times (\text{raw counts} + 1) / \text{median sample coverage})$. \log_2 normalized counts are used in subsequent statistical analyses.

Detecting outliers

To detect individuals with unusually large STRs, STRetch calculates an “outlier score” for each individual at each locus. The outlier score is a z-score calculated using robust estimates, with a positive score indicating the STR is larger than the median. Robust estimates are used to reduce the impact of potential expansions present in control samples, which can be particularly important when comparing within a small cohort of samples.

A robust z-score and p value is calculated for each locus, l , using the log-normalized counts. First, the median and variance across all samples for locus, l , is estimated using Huber’s M-estimator [33, 34]. This calculation can be performed over all samples in a batch or a set of control samples (estimates from the set of 97 PCR-free whole genomes are provided with STRetch). We test the null hypothesis that the log counts, y_{ib} at

locus l for sample i is equal to the median log counts at locus l for the control samples. The alternative hypothesis is that the median log counts for locus l are greater for sample i compared to the control samples. Hence for each sample i and locus l we obtain a robust z-score

$$z_{il} = \frac{y_{il} - M}{\sigma_M},$$

where M is the median and σ_M is the square root of the M-estimator of the variance. One-sided p values are then obtained from the standard normal distribution and adjusted for multiple testing across the loci using the Benjamini–Hochberg method [35]. A locus is called significant if the adjusted p value is < 0.05 .

Estimating allele sizes

We reasoned that the size of an expanded allele would be proportional to the number of reads containing STR sequence and hence the number of reads allocated to the STR locus. In order to estimate allele sizes we performed simulations of a single locus at a range of allele sizes. Specifically, reads were simulated from the SCA8 locus in *ATXN8*. One allele was held constant at 16xAGC and then we simulated repeat lengths in the other allele in the range of 0–500 repeat units (selected at random from a uniform distribution). Alternate versions of the hg19 reference genome with these alleles were produced using GATK v3.6 FastaAlternateReferenceMaker. Reads were simulated from 10,000 bp either side of the SCA8 locus using ART MountRainier-2016-06-05 [36]. Reads were 150-bp paired-end, with insert sizes sampled from a normal distribution (mean 500 bp, SD 50 bp) and 30× coverage (proportional coverage sampled from each haplotype). The Illumina error profile was used. Simulation code is available at github.com/hdashnow/STR-pipelines.

A plot of the number of reads mapping to the AGC decoy chromosome against the number of AGC repeat units inserted into the *ATXN8* locus shows a clear linear relationship between these two variables (Additional file 1: Figure S3).

We can use this information from the simulated data to provide a point estimate of the allele size of any new sample we analyze with STRetch in the following manner. We fit a linear regression between the number of reads mapping to the STR decoy and the size of the allele from the simulated data (both \log_2 transformed), in order to obtain estimates of the intercept and slope parameters, β_0 and β ,

$$y = \beta_0 + \beta x + \varepsilon, \varepsilon \sim N(0, \sigma^2).$$

Here y is $\log_2(\text{coverage})$ and x is $\log_2(\text{allele size})$. Given a new data point from a real sample, the $\log_2(\text{coverage})$

for an STR locus of interest, the point estimate of the allele size is thus

$$\log_2(\widehat{\text{allele size}}) = \frac{\log_2(\text{coverage}) - \hat{\beta}_0}{\hat{\beta}}.$$

where allele size is the number of base pairs inserted relative to the reference and coverage is the normalized number of reads allocated to the locus.

Reference data

Reference genome: ucsc.hg19.fasta, with STR decoy chromosomes added as described above.

STR positions in genome annotated bed file: hg19.simpleRepeat.txt.gz. Source: <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/>.

Known STR loci are obtained by performing a TRF [23] annotation of the reference genome. Pre-computed annotations of many genomes are available from the UCSC Table Browser (<https://genome.ucsc.edu/cgi-bin/hgTables>) [37]. TRF annotations are converted to bed files annotated with two additional columns: the repeat unit/motif and the number of repeat units in the reference.

Running other STR genotypers

To estimate the size of the shorter allele, LobSTR and HipSTR were run on BAM files containing the locus of interest and 1000 bp of flanking sequence on either side. We used LobSTR version 4.0.6 with default settings and the LobSTR reference genome and annotation hg19_v3.0.2. HipSTR version 0.4 was used --min-reads 2 and otherwise default settings, with the provided hg19 reference genome and annotation. In some cases, the tools make a different call as to the reference allele in hg19. To make variant calls comparable to STRetch calls, we converted genotypes to numbers of repeat units inserted relative to the reference defined by that tool, then applied that to the reference allele given by STRetch. For example, if the STRetch reference allele is 20 repeat units, while in HipSTR it is 10 repeat units, a HipSTR genotype of 10/15 would be reported as 20/25. All imperfect repeat units or other variation were ignored and only the total size of alleles taken.

ExpansionHunter version 2.5.5 was run using the 17 provided STR loci specifications, as well as manually defined loci for STRs in *ATXN8*, *MTHFD2*, *NOP56*, and *ZNF9* (see Additional file 1: Supplementary Methods). The “original” version of STRViper was downloaded from <http://bioinf.scmb.uq.edu.au/STRViper> and run using default settings.

Validation of novel STR expansions

To validate the novel STR expansion in an intron of *MTHFD2* (as predicted by STRetch), PCR was conducted

using GoTaq G2 colorless master mix (Promega, USA) with 0.5- μ M primers (see Additional file 1: Table S5) and 10 ng of template DNA per reaction. Cycling was as follows: 95 °C for 2 min, 40 cycles of 95 °C for 15 s, 68 °C for 15 s, and 72 °C for 30 s, and a final extension for 5 min at 72 °C. Samples were analyzed on a 2% agarose gel stained with ethidium bromide (0.5 μ g/mL). Product sizes and STR length were estimated relative to a 100-bp DNA ladder by generating a standard curve (NEB, USA). For sequencing, individual alleles were separated by band-stab PCR [38] (Additional file 1: Table S5), purified using a QIAquick PCR purification kit (QIAGEN, USA), and sequenced using the Sanger method. The samples tested were Samples 1, 2, 5, 6, 8, and 9 from the true-positive samples described above and one standard control sample 1347–02 (CEPH).

To validate STRetch results with long-read data, STRetch was run on Illumina short-read data from an artificial diploid sample created by combining the two haploid genomes, CH1 and CH13. The two replicates were obtained from SRA: <https://www.ncbi.nlm.nih.gov/sra>, accession numbers ERX1413365 and ERX1413368 [39]. We compared the STRetch results to publicly available PacBio variants for these same samples [28]. The authors generated these variants using their SMRT-SV software [27] against the hg38 reference genome. Long structural variants (SVs) > 50 bp were obtained from dbVar; <https://www.ncbi.nlm.nih.gov/dbvar> accession number nstd137 [40] and short SVs from http://eichlerlab.github.io/pacbio_variant_caller/ [41]. A STRetch call was considered validated if the PacBio data contained a substantial expansion relative to the reference genome at the same position with the same repeat unit.

Additional file

Additional file 1: Supplementary tables, figures and methods STRetch_additional_file_1. (PDF 1095 kb)

Funding

HD is supported by an Australian Government Research Training Program (RTP) Scholarship, a Murdoch Children's Research Institute Top Up Scholarship, and an Australian Genomics Health Alliance PhD Award, funded by NHMRC (GNT1113531). NGL is supported by NHMRC Fellowship APP1117510, NHMRC Project Grant APP1080587, and EU Collaborative Grant APP1055295. AO is funded by a National Health and Medical Research Council, Career Development Fellowship GNT1126157. This work was supported in part by the National Human Genome Research Institute, the National Eye Institute, and the National Heart, Lung and Blood Institute grant UM1 HG008900.

Availability of data and materials

The STRetch software is available from <https://github.com/Oshlack/STRetch> [42]. Control sample summary statistics are available from Figshare (<https://doi.org/10.6084/m9.figshare.6830282>) [43]. PCR-free WGS of the ten test samples is available from the Sequence Read Archive (<https://www.ncbi.nlm.nih.gov/sra>), accession SRP148723 (individual sample accessions SRX4114164–SRX4114173). Public short-read data used for validation was obtained from SRA, accession numbers ERX1413365 and ERX1413368 [39]. Corresponding

validated long SVs were obtained from dbVar (<https://www.ncbi.nlm.nih.gov/dbvar>) accession number nstd137 [40] and short SVs from http://eichlerlab.github.io/pacbio_variant_caller/ [41].

Authors' contributions

HD and AO conceived the method and drafted the manuscript. HD implemented the algorithm and performed data analysis. BP and SS contributed to algorithm development. AH performed data analysis. HD, AH, SS, and AL performed software testing. MD, NGL, and JSC performed experimental validation. NGL and PL contributed patient samples. ML and DGM provided reference samples. DGM provided sequencing and computational resources. ML, BP, AH, AL, MD, JSC, NGL, and DGM contributed to the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

This study was approved by the University of Western Australia Human Research Ethics Committee (IRB approval number RA/4/20/1008). All individuals have given written informed consent for publication. The experimental methods in this study comply with the Helsinki Declaration.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, VIC, Australia. ²School of Biosciences, The University of Melbourne, Parkville, VIC, Australia. ³Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. ⁴Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁵Florey Institute of Neuroscience and Mental Health, University of Melbourne, Parkville, VIC, Australia. ⁶Department of Diagnostic Genomics, PathWest Laboratory Medicine, QEII Medical Centre, Nedlands, WA, Australia. ⁷Neurogenetic Unit, Royal Perth Hospital, Perth, WA, Australia. ⁸Harry Perkins Institute of Medical Research, Centre for Medical Research, University of Western Australia, Nedlands, WA, Australia.

Received: 18 December 2017 Accepted: 7 August 2018

Published online: 21 August 2018

References

1. The 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467:1061–73.
2. Gemayel R, Vincens MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet*. 2010;44:445–77.
3. Orr HT, Zoghbi HY. Trinucleotide repeat disorders. *Annu Rev Neurosci*. 2007; 30:575–621.
4. Hamada H, Seidman M, Howard BH, Gorman CM. Enhanced gene expression by the poly (dT-dG). poly (dC-dA) sequence. *Mol Cell Biol*. 1984;4:2622–30.
5. Li YY-CC, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol*. 2002;11:2453–65.
6. Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet*. 2005;6:743–55.
7. van Eyk CL, Richards RI. Dynamic mutations. Tandem repeat polymorphisms. New York: Springer; 2012. p. 55–77.
8. Hannan AJ, editor. Tandem repeat polymorphisms: genetic plasticity, neural diversity and disease. Austin/New York: Landes Bioscience/Springer Science +Business Media; 2012.
9. Mirkin SM. Expandable DNA repeats and human disease. *Nature*. 2007;447: 932–40.
10. Sherman SL, Jacobs PA, Morton NE, Froster-Iskenius U, Howard-Peebles PN, Nielsen KB, et al. Further segregation analysis of the fragile X syndrome with special reference to transmitting males. *Hum Genet*. 1985;69:289–99.

11. Margolis RL. The spinocerebellar ataxias: order emerges from chaos. *Curr Neurol Neurosci Rep.* 2002;2:447–56.
12. Fogel BL, Lee H, Deignan JL, Strom P, Kantarci S, Wang X, et al. Exome sequencing in the clinical diagnosis of sporadic or familial cerebellar ataxia. *JAMA Neurol.* 2014;71:1237–46.
13. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.* 2012;22:1154–62.
14. Willems T, Zielinski D, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods.* 2017;14:590–2.
15. Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res.* 2013;41:e32.
16. Ashley EA. Towards precision medicine. *Nat Rev Genet.* 2016;17:507–22.
17. Cao MD, Tasker E, Willadsen K, Imelfort M, Vishwanathan S, Sureshkumar S, et al. Inferring short tandem repeat variation from paired-end short reads. *Nucleic Acids Res.* 2014;42:e16.
18. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Kingsbury Z, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *bioRxiv.* 2016; <http://biorxiv.org/content/early/2016/12/19/093831.abstract>
19. Tankard RM, Delatycki MB, Lockhart PJ, Bahlo M. Detecting known repeat expansions with standard protocol next generation sequencing, towards developing a single screening test for neurological repeat expansion disorders. *bioRxiv.* 2017; <http://biorxiv.org/content/early/2017/06/30/157792.abstract>
20. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiaz F, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature.* 2015;517:608–11.
21. Gymrek M. A genomic view of short tandem repeats. *Curr Opin Genet Dev.* 2017;44:9–16.
22. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr arXiv.* 2013;3. <http://arxiv.org/abs/1303.3997>
23. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 1999;27:573.
24. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics.* 2013;43:11.10.1–33.
25. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods.* 2009;6:677–81.
26. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large multiallelic copy number variations in humans. *Nat Genet.* 2015; 47:296–303.
27. Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Stephens K, et al. *GeneReviews.* Seattle, WA: University of Washington; 2018.
28. Huddleston J, Chaisson MJ, Meltz Steinberg K, Warren W, Hoekzema K, Gordon DS, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* 2017;27:677–85.
29. Sadedin SP, Pope B, Oshlack A. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics.* 2012;28:1525–6.
30. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25: 2078–9.
31. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–2.
32. Pedersen B. *goleft.* 2016. github.com/brentp/goleft
33. Ripley BD. *Modern applied statistics with S.* New York: Springer; 2002.
34. Huber PJ. *Robust statistics.* New York: Wiley; 1981.
35. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B JSTOR.* 1995; 57:289–300.
36. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics.* 2012;28:593–4.
37. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC table browser data retrieval tool. *Nucleic Acids Res.* 2004;32: D493–6.
38. Bjourson AJ, Cooper JE. Band-stab PCR: a simple technique for the purification of individual PCR products. *Nucleic Acids Res.* 1992;20:4675.
39. Eichler E, Surti U. Evaluating variant calling accuracy with CHM1 and CHM13 haploid data. Accession PRJEB13208. *BioProject.* 2016. <https://www.ncbi.nlm.nih.gov/bioproject/316945>.
40. Huddleston J. Structural variant call data. Accession nstd137. *dbVar;* 2016. <https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd137/>.
41. Huddleston J, Chaisson M, Steinberg K, Warren W, Hoekzema K, Gordon D, et al. SMRT-SV: Structural variant and indel caller for PacBio reads. Small indel data. 2017. http://eichlerlab.github.io/pacbio_variant_caller/.
42. Dashnow H, Sadedin S, Halman A. Oshlack/STRetch: STRetch v0.2.0. 2018. <https://doi.org/10.5281/zenodo.1313915>. Accessed 18 Jul 2018.
43. Dashnow H. STRetch summary statistics from 97 PCR-free whole genomes dataset. 2018. <https://doi.org/10.6084/m9.figshare.6830282>.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)



Chapter 4 Frequency and variability of STR expansions

Acknowledgement

I would like to thank my collaborators for all their contributions to this chapter. In particular, I would like to thank Alicia Oshlack, for her advice and guidance. Belinda Phipson advised on statistical analyses. Simon Sadedin assisted with running STRetch in the cloud environment, as well as developing the Bazam tool, which greatly reduced the runtime required for this analysis. Daniel G. MacArthur provided sequencing data and computational resources.

Supplementary materials to this chapter can be found in Appendix C.

4.1 Introduction

Population allele frequency databases such as ExAC/gnomAD [37,47], 1000 Genomes [6] and the Exome Variant Server [46] are powerful ways to filter out common single nucleotide variants (SNVs) and short indels in the context of Mendelian disease genomics, and have become ingrained in both research and clinical practice. These databases work on the assumption that if a variant is found frequently in healthy individuals it is unlikely to cause a severe Mendelian disease, because such variants would be selected against. This data can also be used to prioritise genes that have lower than expected rates of variation. For example, pLI and observed/expected ratio scores, reported in ExAC and gnomAD respectively, can be used to identify genes with an unexpectedly low number of loss of function variants compared to that expected by chance [37,47,151]. Variant frequencies and gene scores can be used to prioritise variants in clinical and research populations, as evidence that a given variant may be disease-causing. There has been one study to date, considering small allele variation at STR loci across the genome [152]. This study used LobSTR to call STR alleles within the read length on a set of 300 individuals. They found evidence for constraint of STRs in coding regions, especially genes known to cause autosomal dominant developmental disorders, genes with $pLI \geq 0.9$ (loss of function intolerant genes), and highly expressed brain genes. Another study assayed 31 known tandem repeat disease loci across 12,632 individuals and identified 192 individuals with pathogenic-sized STR expansions [96]. While these studies have begun to reveal population variation at STR loci, there has not yet been any reported large study of expansions greater than the read length at all annotated STR loci across the genome. Therefore, when faced with genome-wide STR expansion results for a patient without an expansion at a known pathogenic locus, we have little evidence to rank the hundreds or thousands of expansions found

across a genome for further investigation. There is a clear need for data-driven strategies to prioritise potentially pathogenic STR loci.

Our goal was to create “gnomAD for STR expansions”, by running our method STRetch on whole genome sequencing data from thousands of individuals to estimate the rate of STR expansions across the genome. However, obtaining access to the thousands of samples and vast computational resources required for such an analysis proved beyond the scope of this thesis. In particular, much of the data we attempted to access was siloed within specific institutions or groups, or was clinical data with limited approval for reuse or publication. Therefore, we ran STRetch on the largest set of PCR-free whole genome we were able to access: 362 samples sequenced by the MacArthur lab. This is similar in sample size to the previous publication reporting natural STR variation for STR lengths smaller than the read length [152]. In this chapter we describe the analysis of this data, working towards understanding the rates of STR expansions at a population scale. We detected STR expansions greater than the read length. While the previous analysis considered a similar set of STR loci, they detected only alleles within the read length, therefore there is little to no overlap of the alleles detected between the previous study and those investigated in this chapter [152].

Although this analysis was inspired by ExAC and gnomAD, the analysis of STRs differs from these projects. ExAC/gnomAD considers SNV and short indel variants. These variants have well established variant-calling tools (in particular, GATK HaplotypeCaller for ExAC/gnomAD) that have been tested and validated using various types of orthogonal data [153].

Unfortunately obtaining validation data for STR loci is particularly difficult: amplifying STRs by PCR is problematic and the accuracy of resulting allele calls is questionable due to known biases such as stutter, limitations on the total fragment size that can be amplified causing allelic drop-out, and preferential amplification of the shorter allele [95,96]. This makes it difficult to

assess the quality of validation of any expansion found by short read sequencing. Validation with long-read sequencing is a viable alternative, however there is evidence for disparities in STR allele size estimates between long read sequencing technologies [103].

Another key way in which STRs differ from SNVs and indels is that SNVs and indels are generally limited to a small number of alleles. The majority of SNVs exist in one of two states with a much smaller number of multi-allelic variants found in population studies [37]. In contrast, STR loci typically have many alleles, each a different number of repeat units. While SNVs are considered in terms of presence/absence, we often think of STRs as having presence (variant from the reference) and magnitude (allele size). Previous STR validations (ours included) tend to consider both these attributes. The known difficulties in obtaining an accurate STR allele size estimate often necessitate a more nuanced approach to validation than that used for SNVs/indels.

The field of STR expansion detection from short-read sequencing data is relatively immature in terms of analysis software. To date there are only five tools (including STRetch) that are able to detect large STR expansions in short read data, with most of these published in the past two years. Each tool has different strengths and limitations, discussed in the Introduction and in Chapter 3. For STRetch, an important limitation for this analysis is limited dynamic range. STRetch is only sensitive to moderate to large STR expansions, specifically those that are larger than the reference allele. While STRetch can detect very large expansions, it tends to underestimate the size of alleles great than the insert size of the sequencing data. STRetch is also limited to detecting expansions at specified STR loci.

STRetch works by competitively mapping reads to the reference genome as well as a set of 501 STR decoy chromosomes which represent all possible

combinations of 1-6bp repeat units. STR loci in the reference genome are relatively small (numbers) so that reads arising from an expansion will tend to preferentially map to the decoy chromosomes. STRetch then looks for pairs of reads where one read aligns to one of the STR decoy chromosomes while the other aligns to the reference genome near a known STR locus with the same repeat unit. However, if the sample's allele is a similar size to the reference allele, then both of the reads are more likely to map to the reference genome and so not be counted by STRetch. Therefore, STRetch is more likely to detect STR expansions that are larger than the reference allele. Additionally, although STRetch is able to identify very large expansions, the allele size estimate is approximately bounded by the insert size of the sequencing data. This is because as the sample's allele increases beyond the insert size, there will be a number of reads where both in the pair are completely made up of STR repeat units. These reads will both map to the STR decoy chromosomes and so generally cannot be uniquely attributed to a specific STR locus.

Finally, STRetch is also limited to detecting expansions at specified STR loci, by default all those that are annotated in the reference genome. Although the use of STR-decoy chromosomes allows us to observe reads that do not correspond to known loci, these are currently not allocated to a specific genomic context.

In this chapter we first explore the distribution of STRs annotated in the human genome and comment on how the annotation of these loci might be used to prioritise STRs in a locus-discovery setting. We apply STRetch to the analysis of 362 PCR-free whole genomes including a subset of 125 unrelated individuals with no known Mendelian disease (described as “unaffected”). By counting reads assigned to the STR decoy chromosomes in these individuals we can identify inconsistencies between STRs annotated in the reference genome and evidence of STR expansions in these samples. We explore the patterns and frequency of STR expansions in this group of individuals,

including those in known pathogenic STR loci, and use this to inform the way in which we prioritise variants as potentially pathogenic. In addition to exploring population variation of STR expansions, this chapter extends the ideas and analyses in the STRetch paper [104] (Chapter 3), and looks at the evidence for and against false positives in STRetch results. We also consider the choice of control samples when running STRetch and how this impacts the number of significant STRetch results, and comment on control choice in research and clinical settings.

4.2 STR loci in the human reference genome

We first assessed the distribution of all STR loci in the human reference genome. To define the set of all STRs in the human hg19 reference genome, we used the Tandem Repeats Finder annotation downloaded from UCSC and limited it to loci with 1-6bp repeat units. The resulting set of STRs consisted of 308,585 loci, 304,717 of which were on the “canonical chromosomes” (chromosomes 1-22, X and Y). All other STRs were excluded from further analysis as those on non-canonical chromosomes (e.g. alternate haplotypes and contigs that are not assigned to a specific position) are not included in our other annotation data. The STRs were distributed amongst the chromosomes roughly as would be expected based on the relative sizes of the chromosomes (Figure 4.1A). We found that 2bp repeat units were most frequent, followed by 1 and 4 bp, 5, then 3 and 6 bp repeat units (Figure 4.1B), and the most common individual repeat unit was A/T homopolymers, which collectively made up 22.71% of STRs in the genome. There are 501 possible permutations of repeat units in the range 1-6bp (see Chapter 3 for more details), however only 382 different repeat units are annotated in the human genome.

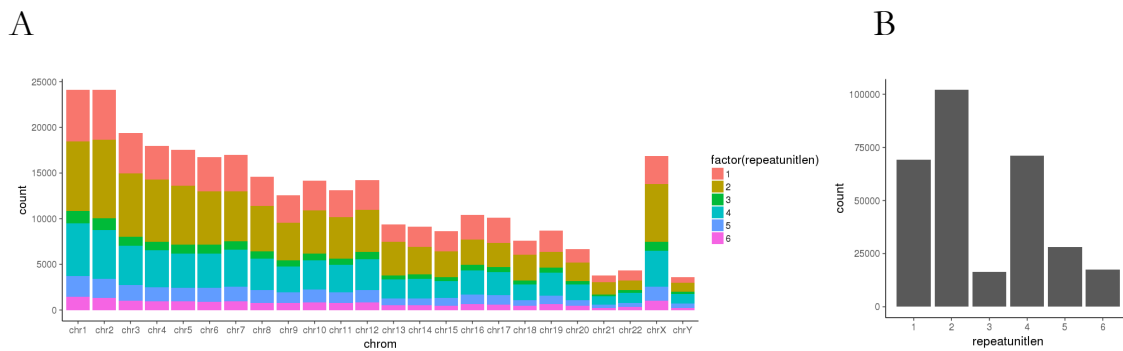


Figure 4.1: STR loci counts from the Tandem Repeats Finder annotation of hg19.

A) Number of STR loci annotated on each canonical chromosome, coloured by repeat unit length. B) Total number of STR loci in the genome with each repeat unit length.

4.2.1 Annotating STR loci

STR loci were annotated with the genomic feature that they overlap using STRetch [annotateSTR.py](#). Genomic features were defined by Gencode Release 19 (GRCh37.p13) Comprehensive gene annotation GFF3 obtained from https://www.gencodegenes.org/human/release_19.html, which includes 57,781 gene names. Introns were added using genomertools version 1.5.10 [154]. If a single STR overlapped multiple genomic features then only one was annotated with the first feature in order of precedence: CDS, exon, intron, and gene. Non-CDS exons are predominately made up of 5' and 3' UTRs, while the 'gene' feature was a misc./not otherwise specified category with few features, and was removed from most subsequent plots. An STR was categorised as intergenic if it did not overlap with any gene features. For STRs that overlapped a gene, we indicated if that gene was present in OMIM. STRs were also annotated with their distance to the closest transcription start site (negative for upstream, positive for downstream, 0 for overlapping). We annotated 23 known pathogenic STR loci, as described in Chapter 3 (see also Appendix B Table S3).

The vast majority of STRs were in introns or intergenic regions. Only 623 (0.20

%) STRs overlapped coding regions, with another 4730 (1.55%) overlapping other exons. A further 112161 (36.81%) STRs overlapped with an OMIM gene. The majority of these were intronic, however 527 STRs overlapped an OMIM CDS and 2967 overlapped an OMIM exon. For context, Gencode 19 annotates 1.15% of the genome as coding sequence, and 3.94% as any exons including CDS (Appendix Table 6.9). About half of STRs were in introns, consistent with introns making up almost half the hg19 reference genome. Three and six bp repeat units were enriched in coding regions and to some extent exons (Figure 4.2).

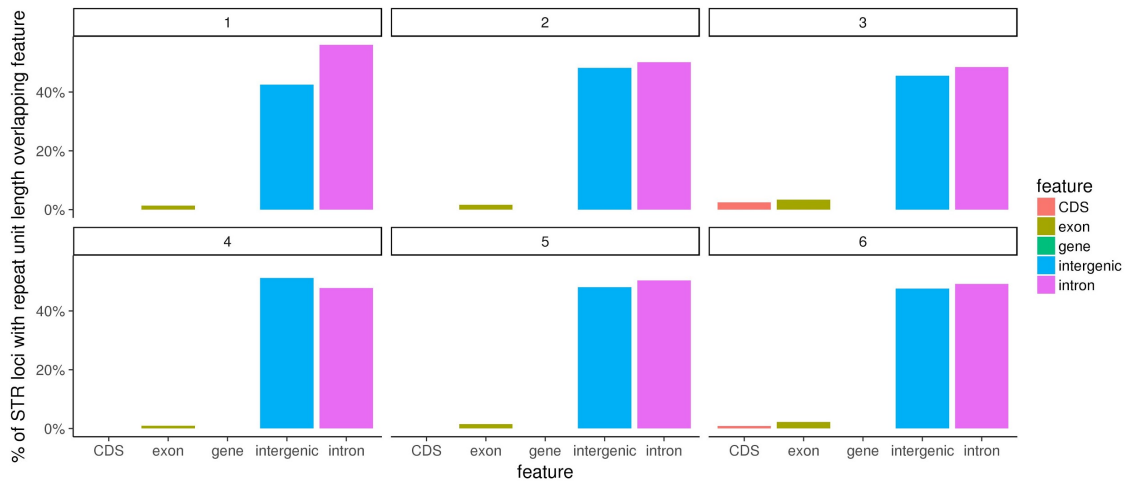


Figure 4.2: Percentage of the STR loci in each repeat unit category that overlap a given genomic feature.

We found 20,140 (34.86%) genes with at least one STR overlapping any of their features (including non-coding exons and introns), 537 (0.93%) genes with at least one STR overlapping one of their coding regions, and 3975 (6.88%) genes with at least one STR overlapping a non-coding exon. Of the OMIM genes, 11297 had at least one STR overlapping any of their features, and 450 had at least one STR overlapping one of their coding regions. Finally, 8624 (14.93%) genes had an STR within 3000 bp upstream of their transcription start site (TSS), which therefore may be involved in promoter function. Of these, 7195 (12.45%) STRs lay within 2000 bp, and 4929 (8.53%) within 1000 bp of the TSS.

4.3 Counting reads assigned to the STR decoy chromosomes

We ran STRetch on 362 PCR-free whole genomes, comprised of individuals being investigated for the cause of their Mendelian disease, or immediate family members of such patients, including:

- A set of 219 individuals sequenced as part of the Rare Genomes Project (RGP, Centre for Mendelian Genomics <https://cmg.broadinstitute.org/>), including 84 affected individuals as well as many of their parents and some unaffected siblings.
- The 143 “other” samples, predominantly Mendelian disease cases sequenced by the MacArthur group, as well as some family members. These included the majority of the 97 genomes from the STRetch paper, who were used as the control set for the original publication.

STRetch v0.3.3 was run with default settings on Google Cloud infrastructure and using hg19. Since the version described in paper in Chapter 3, STRetch was updated to use Bazam to extract reads aligning to STR loci for re-mapping to the STR decoy chromosomes. With the addition of Bazam, STRetch runtime was reduced by approximately 60% with no loss of sensitivity (validated in the Bazam paper [155]).

STRetch works by first aligning reads to a reference genome with additional STR decoy chromosomes. These 501 chromosomes represent all possible combinations of 1-6bp repeat units. Next STRetch looks for pairs of reads where one read is aligned to an STR decoy chromosome and the other is aligned to the reference genome. It then attempts to assign this count to the closest annotated STR locus with a repeat unit that matches the STR decoy chromosome. However, a read can't be assigned if there isn't a matching STR locus annotated nearby. We therefore counted the number of reads aligned to each of the decoy chromosomes, which allows measurement of the frequency

of STR repeat units in the samples even if they are not annotated in the reference genome. This is important as a large number of reads aligned to an STR decoy chromosome could indicate either a large number of moderately expanded STR loci, or a smaller number of hugely expanded STRs. Where a read pair can be used to anchor these to the reference genome, we can potentially use this to detect unannotated or *de novo* STR expansions. However, if both reads in a pair align the STR decoy chromosomes unique allocation is generally not possible (unless there is only one STR expansion with that repeat unit).

When all samples were considered together, we found the largest number of reads aligned to the C/G homopolymer STR decoy chromosome (72.7% of STR reads), followed by the homopolymer A/T decoy (16.9% of STR reads, Table 4.1). This was surprising and contrasted with our expectation based on STR annotations of the reference genome, where A/T homopolymers made up 22.71% of STRs, while C/G homopolymers which made up only 0.01% of STR loci. The high frequency of A/T homopolymers has been proposed to reflect an evolutionary history of re-integrations of poly-A tails into the genome [156]. The third most frequently aligned decoy chromosome was AACCT (4.86% of STR reads). This is the telomere repeat unit [157], which is rare in the reference genome (170 loci, 0.06% of STRs) so we expect to see this frequently on the STR decoys. These results were generally consistent when the 59 most highly variable samples were removed: C/G homopolymers still represented the most common STR read (~79% of STR reads across all samples), the telomere sequence moved up to 2nd most common (~7% share of STR reads), and A/T homopolymers dropped to the 3rd most common with ~4% of STR reads (Appendix Table 6.10).

Table 4.1: The top ten STR decoy chromosomes ranked by the number of reads aligned to them across all 362 samples.

Repeat unit	Mean decoy counts per sample	Mean decoy counts / sequencing depth	Percent of decoy reads
C	137927	4043	72.7
A	32061	954	16.9
AACCCT	9212	265	4.86
AATGG	2704	78.3	1.43
ACTCC	1822	52.6	0.961
ACC	587	16.9	0.309
ACACT	520	15.1	0.274
AT	463	13.5	0.244
AGGG	318	9.18	0.168
AG	297	8.58	0.157

The mismatch between the number of annotated STRs and the number of STR decoy reads may indicate loci missing from the reference genome and/or *de novo* STR expansions. It has been shown that the human reference genome is still incomplete, especially highly repetitive heterochromatic and extreme GC regions, due to limitations in sequencing and genome assembly.

Furthermore, the human reference genome underestimates the lengths of many STR loci and may not represent the most common allele for many other loci [158]. It is therefore possible that these excess STR decoy reads arise from

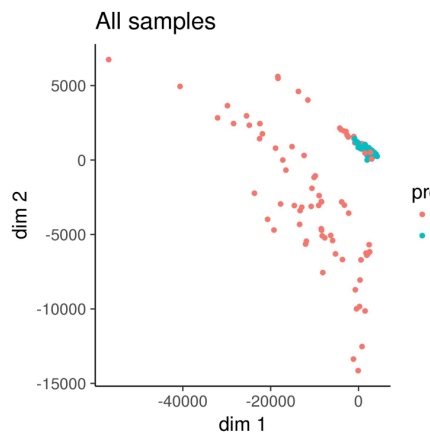
common sequences that are not represented in the reference genome. Alternatively, they may also be caused by some technical artefact of the sequencing process or contamination.

We used the counts of reads assigned to each STR decoy chromosome to generate an MDS plot of all samples. When coloured by project, we saw evidence of batch effects in the data; samples from the RGP clustered closely with each other, and with some other samples, while many of the samples from other projects appeared more dispersed (Figure 4.3A). These batch effects were not explained by median sequencing depth which was diverse across projects (Appendix Figure 6.2).

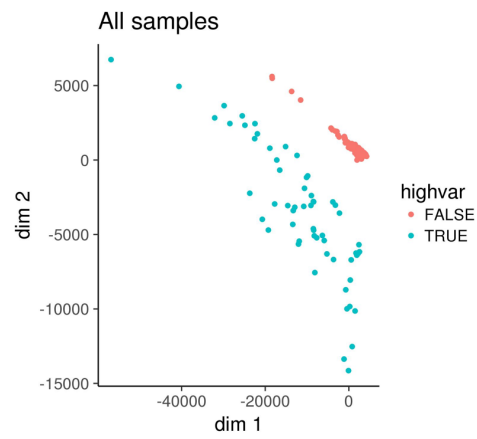
To explore batch effects further, we divided samples into high and low variability based on “highvar” samples having greater than 10,000 STR loci with any reads assigned (this will be explored further in the following section). We saw clear separation of high and low variability samples on the MDS plot (Figure 4.3B). Inspection of STR repeat units by rank of variance across all samples (the same metric used to rank genes in the limma plotMDS function when gene selection is set to "common") showed that homopolymers contributed the most variation, followed by the telomere sequence (Figure 4.3C). The top ten repeat units used in the MDS closely matches those seen have the most reads aligned to decoy chromosomes (Table 4.1), as would be expected for Poisson count data. We removed the 59 highly variable samples, repeated the MDS plot, and found substantially lower separation between the RGP and other project samples (Figure 4.3D). We removed all homopolymers and performed MDS on all 362 samples (Figure 4.3E). The second dimension was dominated by a single sample, however plotting the first and third dimensions showed relatively minor separation of the RGP and other project samples (6F). This indicates that much of the variation initially noted in the highly variable samples was present in the homopolymers. This further

pointed towards a technical bias in the data, as homopolymers are known to be particularly difficult to amplify and sequence accurately [20].

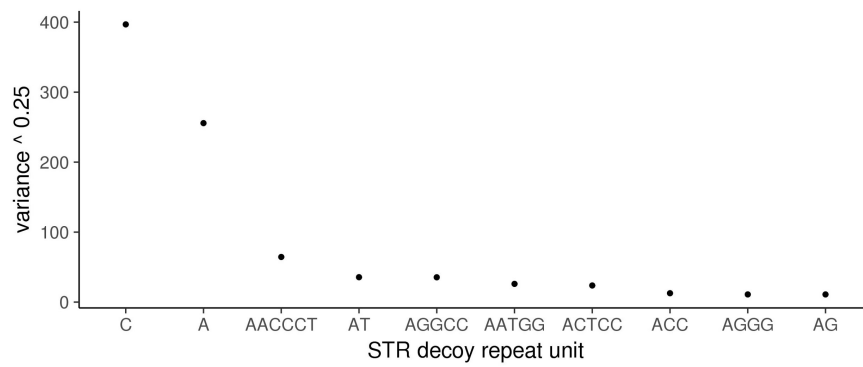
A



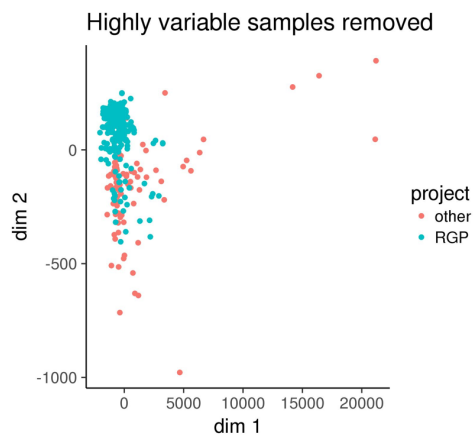
B



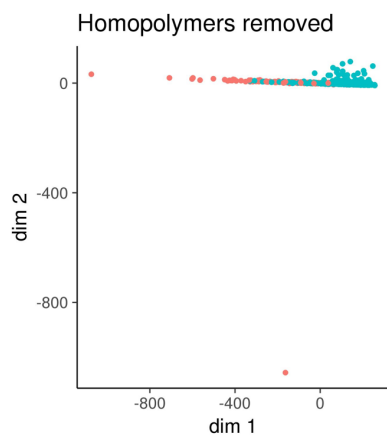
C



D



E



F

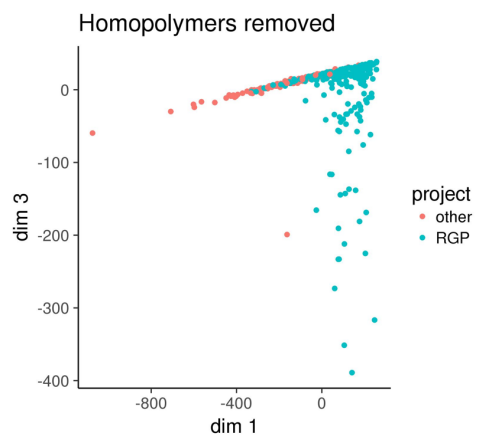


Figure 4.3: MDS plots of STR decoy counts.

A) MDS of STR decoy counts from all samples, coloured by project. B) MDS of STR decoy counts from all samples, coloured by variability category. C) Top 10 most variable decoy chromosomes. D) MDS of STR decoy counts from the 303 lower variability samples, coloured by project. E) MDS of STR decoy counts from all samples with homopolymer STR decoys removed. F) Dimension 1 and 3 of MDS in plot E.

4.4 Frequency and patterns of STR expansion

We examined the STRetch outputs on all samples by looking at the number of STR decoy reads assigned to each genomic STR locus and the p values calculated from comparison of normalised read counts with controls. More specifically, STRetch works by looking for pairs of reads with one aligned to an STR decoy chromosome and the other aligned near an annotated STR locus with the same repeat unit as the decoy chromosome. A single read pair assigned to an STR locus contributes one count to that locus. Detecting a number of counts at a single STR locus in one individual suggests that an STR expansion may have occurred. STRetch then normalises these counts by median sequencing coverage across the genome and takes the log₁₀ to calculate the “locuscoverage_log”. These values are then compared against a set of control samples to determine if that individual has significantly more reads compared to controls.

Most STRs show no evidence of a large expansion. In our sample of 362 individuals, 88,285 (28.97%) loci had at least one STR decoy read assigned in any individual. Most loci were only detected in one or a small number of individuals, and similarly most loci were significant in one or a small number of individuals (Figure 4.4). Further, 76,911 (25.17%) loci had at least one individual with a significant expansion at $p < 0.05$ and 73,729 (24.20%) loci had at least one individual with a significant expansion at $p < 0.01$ (using FDR adjusted p-values as described in Chapter 3).

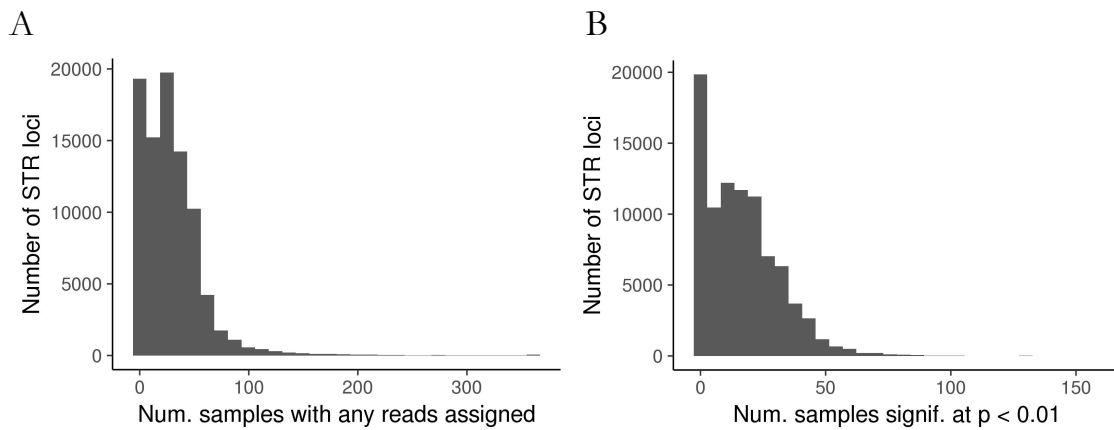


Figure 4.4: STRetch results per locus across all samples.

A) Number of samples that have STR loci with any reads from the decoy chromosomes assigned B) Number of samples with significant STR loci at $p < 0.01$.

Looking per individual, samples had a mean of 1588 loci with any STR decoy reads assigned to them (range 830-51,038). A mean of 5655.4 loci (range 23-49,428, median 129.5) per sample were significant at $p < 0.05$, while a mean of 4024.88 loci (range 14-48,217, median 92.5) per sample were significant at $p < 0.01$.

There were a small number of individuals with a very large number of expanded STRs (Figure 4.5A). There was no relationship between the number of significant STR loci per sample and sequencing depth a (Appendix Figure 6.2). STR calls were distributed disproportionately among samples, with 59 of the STRetch controls samples having unusually high numbers of STR expansions (Figure 4.5B). These were all sequenced as part of the same relatively old batch, so we suspect that the unusually high numbers of STR expansions were due to differences in sample preparation and sequencing relative to our newer samples.

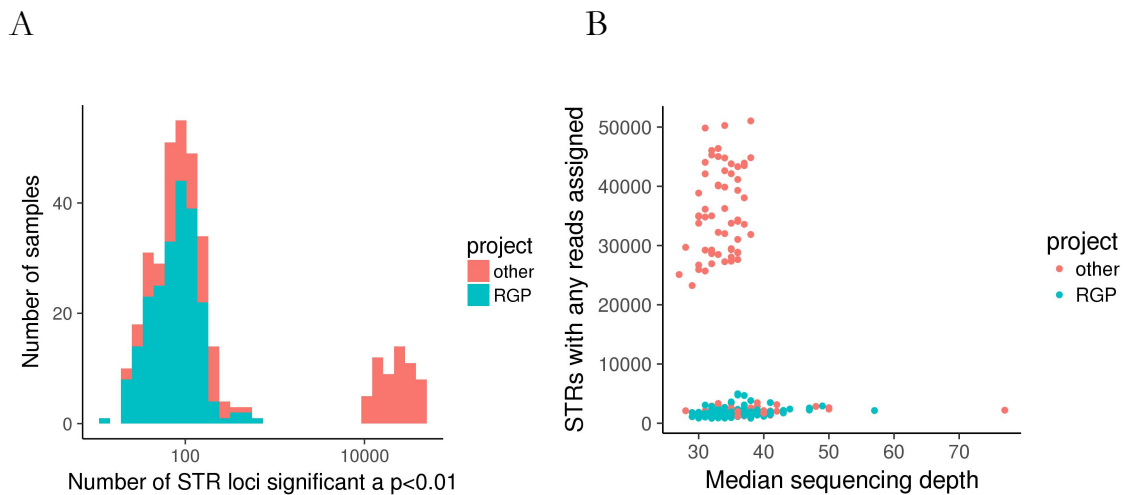


Figure 4.5: STRetch results per sample.

A) Number of significant STR loci per sample at $p < 0.01$. B) Number of STR loci with any reads assigned vs. median sequencing depth across the genome for each sample.

Removing highly variable samples from the analysis substantially reduced the number of variant calls. We removed samples with more than 10,000 STR loci with any reads assigned. In the remaining 303 samples, 56,962 loci (18.69%) had any decoy reads assigned in at least one individual, 13493 loci (4.43%) had at least one individual with a significant expansion at $p < 0.05$ and 9515 loci (3.12%) had at least one individual with a significant expansion at $p < 0.01$. We observed a mean of 1653 loci with at least one read assigned per sample (range 830 - 4967) with a mean of 213.4 loci (35-2155) per sample significant at $p < 0.05$ and 153 loci (24-1106) per sample significant at $p < 0.01$. There still remained a small number of samples with larger numbers of loci with any counts and larger numbers of significant STR loci, but this pattern is not as dramatic as in the full data set (Figure 4.6).

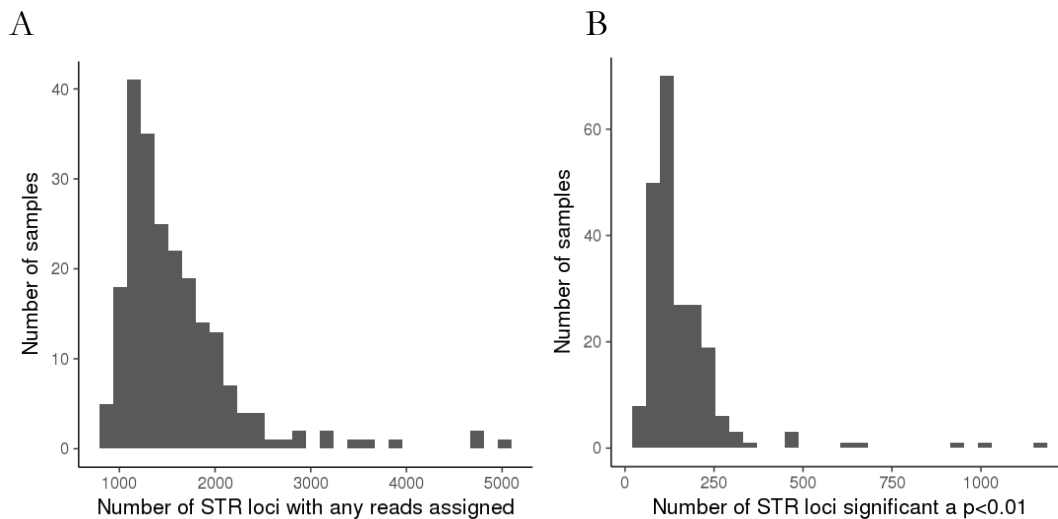


Figure 4.6: STRetch results per locus across the 303 less variables samples.

A) Number of STR loci with any reads assigned B) Number of significant STR loci at $p < 0.01$.

4.5 STR constraint and prioritising potentially pathogenic expansions

As we have seen, when running STRetch on a single individual we will typically find hundreds of significant STR expansions. If the individual does not have a result in one of the known pathogenic STRs (or in one that does not match the phenotype) we must prioritise investigation of the remaining list of STRs. To some extent we can look at attributes of known pathogenic loci to inform prioritisation of new loci to investigate as potential pathogenic variants. However, the number of well-studied pathogenic STRs is limited and the list is likely highly enriched for certain attributes due to ascertainment bias. As shown in Table 1.1 in the Introduction, the majority of pathogenic STRs are coding and have 3 bp repeat units (mostly CAG). This could indicate either that the CAG repeat unit is more likely to be pathogenic, or that it has been focused on for historical reasons. The CAG Huntington disease locus was one of the first discovered and has one of the best understood STR disease mechanisms, which has likely led to researchers focusing on other CAG loci, especially in coding regions. Thus, while known attributes can assist identification of pathogenic loci, they may also limit the search for novel loci.

We therefore explored a data-driven approach to prioritising STR loci for further consideration.

We first considered the variability of STR loci across our full set of individuals. Our assumption was that variants that are highly variable are probably not pathogenic (even if they are significant), because we wouldn't expect that lack of constraint from a pathogenic locus. To investigate this, we calculated Huber's robust estimates of median and variation of the log sequencing-depth-normalised counts of reads assigned to each STR locus across all individuals. We observed a strong relationship between the median and variance, especially at lower values (Figure 4.7A-D), a pattern is typical of count data. STRetch has a limited dynamic range, therefore, many of the loci with zero reads may have smaller alleles that are below our detection threshold. Most of the loci with significant expansions also have low counts, as would be expected if expansions were rare in the population (Figure 4.7A). Most CDS and exonic STRs fall in the lower count and lower variance portions, suggesting that there may be selective pressure against large expansions at these loci (Figure 4.7B). Those known pathogenic STR loci for which we observed counts in these samples show even less variation (Figure 4.7C).

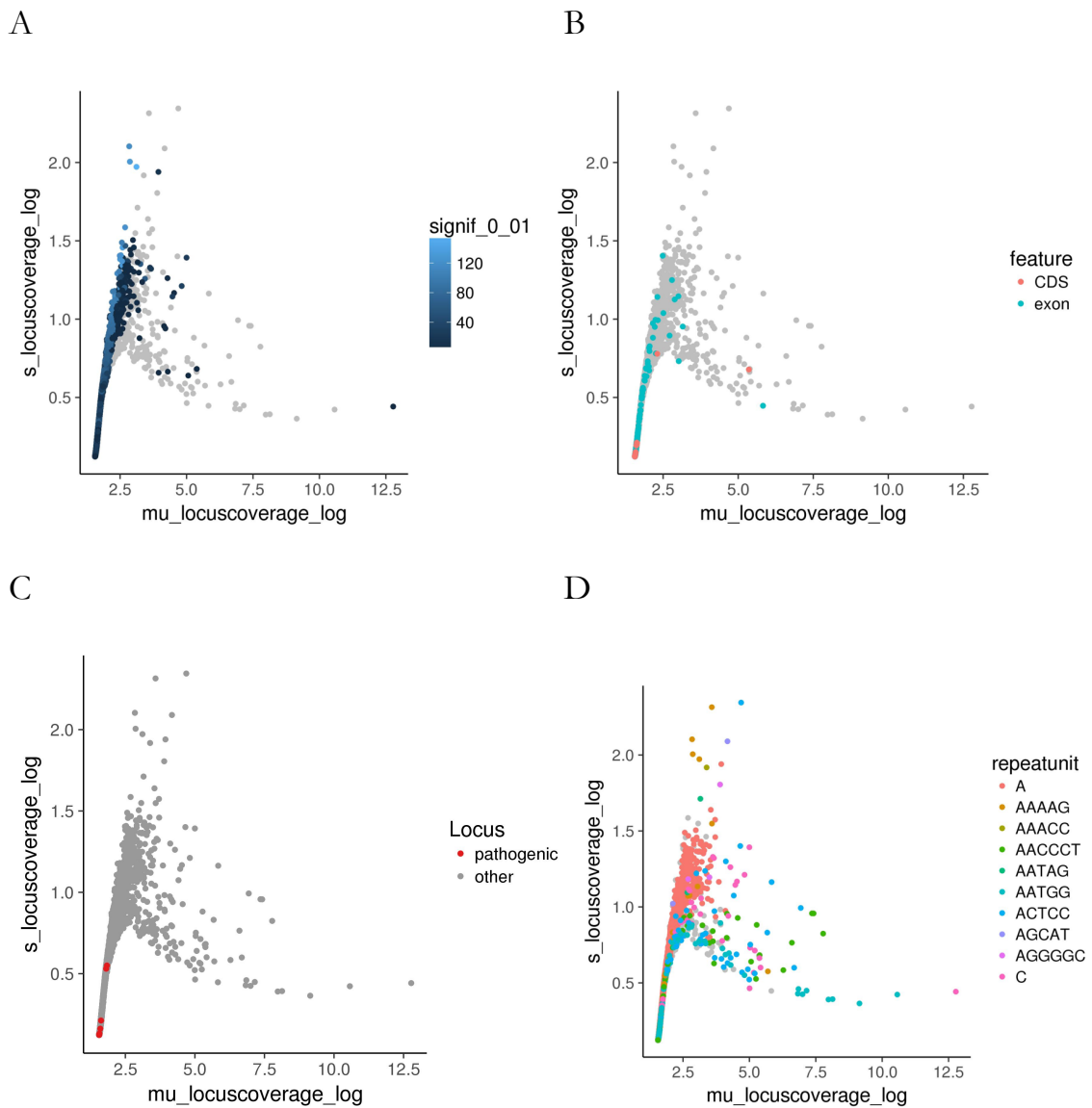


Figure 4.7: Huber's estimates of median and variance across all samples for each STR locus.

A) Coloured by number of samples with significant calls a $p < 0.01$. B) STR loci overlapping CDS and non-coding exons are coloured, with all other loci in grey. C) Known pathogenic STR loci are coloured with all other loci in grey. D) STR loci sharing a repeat unit with the top ten most variable or top ten highest mean loci are coloured, with all other loci in grey.

As discussed earlier, we found homopolymers (A, C) and telomeric repeat units (AACCT) to be the most frequent and variable in terms of the number of reads aligned to the STR decoy chromosomes. When looking at individual STR loci, only the homopolymer A's appear amongst the most variable loci, while homopolymer Cs and telomeric repeats have some of the highest medians (Figure 4.7D). The top ten most variable loci had repeat units

ACTCC, AAAAG, AGCAT, AAAAG, A, AAACC, AGGGGC and AATAG. The repeat units of the top ten loci in terms of assigned counts were: C, AATGG, and AACCCCT. These high coverage loci are particularly long in the reference genome. All are over 100 bp and some are 1,000-5,000 bp. Together these high variance or high coverage repeat units account for almost 80% of loci we observe any variation in, but only 25% of all loci annotated in the genome.

We found evidence for reduced variation across individuals in STRs that overlap coding regions. To further investigate this variation in coding and non-coding regions, we first divided the STR loci into ten groups based on variance estimates. We then calculated the expected number of loci overlapping each feature based on STR loci distributed proportional to the size of the group, and compared this to the observed values. We found that the lower variance deciles were enriched for STR loci that overlapped CDS, and conversely, that the higher variance deciles were markedly depleted for CDS loci (Figure 4.8A). Non-coding exonic STRs showed a much more subtle enrichment at lower deciles, while intronic and intergenic STRs showed approximately expected numbers of loci at all variant deciles. Due to the small number of STRs in coding regions we only expected 18 loci in each decile, and observed zero to four in the higher variance deciles. We combined the CDS and exon categories to boost numbers (Figure 4.8B). We still found a trend for enrichment in the lower variation deciles for the CDS/exon category, however it was more subtle. We calculated observed and expected numbers of STR loci in each group for the remaining loci for which no STR reads were observed in any samples (Table 4.2) and found more STRs overlapping coding regions than expected for these loci. Overall, we saw a tantalising trend towards reduced variation of STR loci in coding and possibly also non-coding exons. However, because of the small numbers involved we cannot make a clear conclusion.

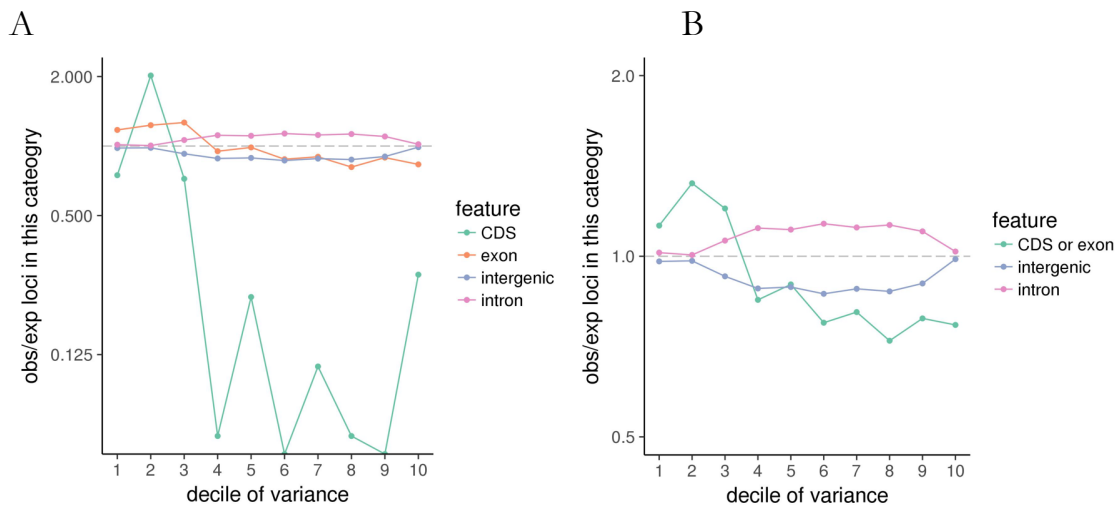


Figure 4.8: The Huber’s variance of the log normalised STR locus counts were divided into ten approximately even deciles (x-axis: decile of variance).

For each decile we calculate the observed number of STR loci overlapping a given feature type and divide this by the expected number of loci in that category, assuming the number of loci overlapping a given feature is in proportion to the number of loci in that decile (y-axis: obs/exp loci in this category). An obs/exp value greater than one indicates enrichment, while less than one suggests depletion. A) All features apart from “gene”. B) CDS and exon features combined.

Table 4.2: STR loci with no STR decoy reads assigned in any sample.

We calculate the observed number of STR loci overlapping a given feature type and divide this by the expected number of loci in that category, assuming the number of loci overlapping a given feature is in proportion to the number of loci in that category.

Feature	Observed	Expected
CDS	547	442
Exon	3367	3355
Intergenic	106068	102604
Intron	106419	109999

Combined, our results provide some evidence for constraint in STRs overlapping coding regions. Together with our finding of reduced variance at

known pathogenic STR loci, this indicates that these annotations may provide a useful way of prioritising STR loci for further consideration. To further examine the relationship between STR variation and their position relative to genomic features, future studies would likely need to increase the sample size as well as the dynamic range. As noted earlier in the chapter, each individual tends to have variation at different STR loci, so increasing the number of samples may help to increase the number of loci in which we observe variation. We also expect that there is a large amount of STR variation that is below the range of allele sizes that STRetch can detect. Further development of STRetch or use of a combination of tools may therefore increase the dynamic range of such an analysis.

4.5.1 STR expansions in set of unaffected individuals

Thus far our analysis has considered the full set of individuals available to us. However, many of these samples have likely Mendelian disease. We therefore selected a subset of 125 samples (67 female, 58 male) that were unaffected and unlikely to be related to each other. These samples consisted of all the parents from the RGP who were listed as “unaffected” and for whom we have metadata (relationships/affected status/disease/HPO terms). We ran STRetch on these samples using the same 125 samples as controls for each other.

We observed similar patterns of variability in the pathogenic loci in the unaffected samples to that seen in the full data set (Figure 4.9A). Even amongst pathogenic STR loci we found diversity in variance estimates. In particular, we found a larger degree of variability in the FTDALS1_C9orf72 and FXTAS_FMR1 loci (Figure 4.9B) than other pathogenic loci, even those in non-coding regions.

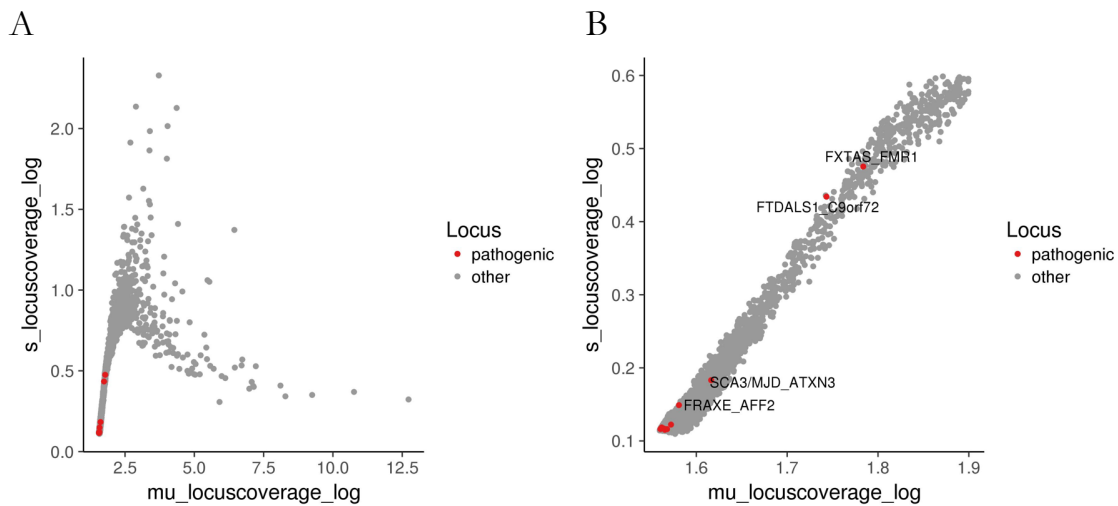


Figure 4.9: Huber's estimates of median and variance across the unaffected samples for each STR locus. Known pathogenic STR loci are coloured with all other loci in grey. A) All loci. B) Zooming in on the low median, low variance loci.

FTDALS1 is an intronic STR, generally reported to have a pathogenic range of 60-2000 GGGGCC repeat units, however some studies have reported pathogenic alleles as small as 30 repeat units while others have reported unaffected individuals with alleles in the pathogenic range [159].

Understanding of FTDALS1 allele sizes is further complicated by an age at disease onset ranging from 30 to 70 years, which means it can be difficult to distinguish affected and unaffected individuals as a current lack of symptoms doesn't preclude onset later in life.

FMR1 is a non-coding UTR STR locus associated with three different conditions at different allele sizes and in different sexes: fragile X syndrome (FXS), fragile X-associated tremor/ataxia syndrome (FXTAS), and FMR1-related primary ovarian insufficiency (POI, also known as premature ovarian failure-1, POF1) [52,160]. FXS is caused by > 200 CGG repeat units (typically thousands). Allele sizes between 55 and 200 repeat units (classed as a premutation) convey increased risk for FXTAS, with higher penetrance in males and increased risk with age. Females with the premutation also have an approximately 20% risk of POI. Like FTDALS1, analysis of FXTAS is

complicated by reduced penetrance and late-onset, making it difficult to demarcate affected and unaffected individuals.

Across all 125 individuals we observed 50 significant (p adj. < 0.01) calls in eight known pathogenic STR loci (Table 4.3). Comparing the number of significant calls at $p < 0.01$ vs 0.05 and looking at the distribution of p values across samples highlighted the impact of using a specific cut-off (Figure 4.10A). Looking at the distribution of p values or using a ranking approach may be appropriate here. Although these individuals were listed as unaffected, they had affected children. We therefore inspected the phenotypes for these families, and found that none of the significant STR expansions matched the phenotypes. For all but two of these significant pathogenic loci, the allele sizes estimated by STRetch were below the pathogenic range (Figure 4.10B-C). The two pathogenic-sized alleles were in the SBMA and SCA8 loci.

Table 4.3: Number of significant STRetch calls within known pathogenic STR loci across all 125 unaffected samples at $p < 0.05$ and $p < 0.01$.

Pathogenic STR locus	N signif. at $p < 0.05$	N signif. at $p < 0.01$
DM2_ZNF9	1	1
FRAXE_AFF2	8	7
FTDALS1_C9orf72	22	20
FXTAS_FMR1	15	15
HD_HTT	1	1
SBMA_AR	1	1
SCA3/MJD_ATXN3	5	4
SCA8_ATXN8/ATXN8 OS	1	1

The SBMA finding may be a true positive. SBMA is an X-linked condition with typical age of onset between 30 and 50 years [161]. The individual with the potentially pathogenic expansion is female, and heterozygous females generally have no or sub-clinical symptoms of SBMA. STRetch did not detect the SBMA expansion in any of her other family members.

It is unclear if the SCA8 finding is a true positive. Although alleles for the SCA8 locus greater than 80 CAG repeat units are considered highly penetrant, alleles ranging from 71 to more than 1300 repeats have been found in both affected and unaffected individuals. This is further complicated by an age of onset of up to 73 years [162]. So, although this individual is currently listed as unaffected, they may show symptoms in the future, or may never show

symptoms even if this expansion is true. This highlights the difficulty in distinguishing false positives from true STR expansions in the sub-pathogenic range for a late onset or low penetrance disease, as well as the potential for discovering secondary findings for late-onset disease when testing STRs.

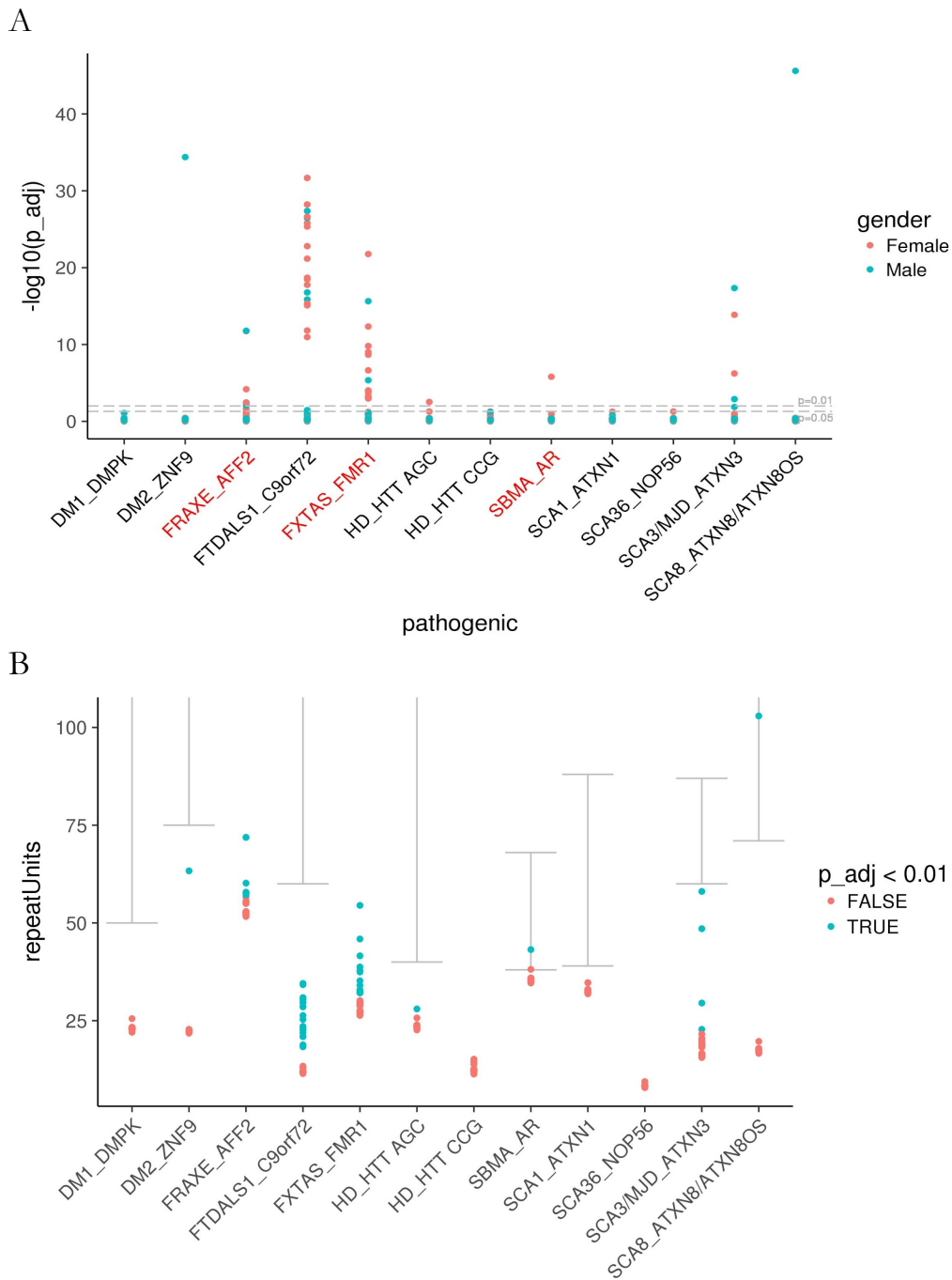


Figure 4.10: STRetch results for all pathogenic loci with results in the 125 unaffected samples.

Note that 23 pathogenic loci were tested, but most had no reads assigned and therefore were not reported. The HD_HTT CCG locus is not classed as pathogenic but it abuts the pathogenic AGC locus and may be a moderator. A) P values for each sample at the pathogenic loci coloured by reported gender. The labels of loci on the X chromosome are coloured in red. B) STRetch size estimates of the total number of repeat units, coloured by significance at $p < 0.01$. Pathogenic allele sizes are indicated by grey bars (many are outside the plotted range).

We found twenty unaffected individuals with significant expansions in FTDALS1_C9orf72, a locus that we have previously reported as a likely false positive [104]. It is noteworthy however that we have previously reported and PCR-confirmed significant non-pathogenic expansions in the SCA3 and HD loci. That is, the loci were expanded relative to other samples, but the allele was below the pathogenic range, suggesting that STRetch is actually working as expected. We were also able to confirm one of the FTDALS1_C9orf72, expansions but not the rest (see

STRetch paper supplementary materials in Appendix B). We also note that the pathogenic loci in which we observe the most expansions also have the highest median counts, indicating that these are larger loci overall. This gives us more power to detect expansions in these loci.

Three of the pathogenic loci we observed in this set of samples were on the X chromosome. While males and females had approximately equal rates of significant expansions in these loci (Figure 4.10A), they have different coverage of the X and Y chromosomes. We therefore suggest that future implementations of STRetch could use matched-sex samples as controls for STR loci on the sex chromosomes.

4.6 The impact of sample choice in STRetch control sets

STRetch uses a cohort of control samples to calculate its p values. For each locus it compares the test sample against this set of controls, asking if the sample has significantly more reads assigned to this locus than the controls. The choice of controls may therefore have a large impact on the results. To investigate the impact of the control choice on STRetch results we ran STRetch on the full dataset of 362 samples (using the same samples as controls), and then on 3 different sample subsets: 125 unaffected parents from the RGP set (“unaffected”), the 59 most highly variable samples (“highvar”), and the remaining 303 less variable samples (“lowvar”). There is no overlap between the “highvar” and “lowvar” group, and all “unaffected” are also in “lowvar”. In each case we used the samples in that set as controls (the default behaviour of STRetch) and also emitted the per locus summary statistics for each set so that they could be used as controls for other samples.

For each sample subset we counted the number of loci in each sample with significant expansions at $p < 0.01$ (Figure 4.11A). Looking at all samples we found a number of outliers with many more significant calls per sample, corresponding to the previously identified “highvar” samples. Running these

highly variable samples separately resulted in a smaller of significant calls per sample, as the control set was smaller and more variable. The “lowvar” set had a slightly larger number of significant calls per sample, and the unaffected set had even more. Looking at this data in another way, for each locus we counted the number of samples with significant expansions at $p < 0.01$, and displayed it as a percentage to allow comparison across sample subsets of varying sizes (Figure 4.11B). This showed that most loci were only significant in one or a very small number of samples, with a long tail of small numbers of loci that are significant in multiple individuals. However, for the highvar samples we found an unexpected peak in this distribution, with a large number of loci significant in many samples. This is also seen in all samples to a lesser extent as the same samples are included.

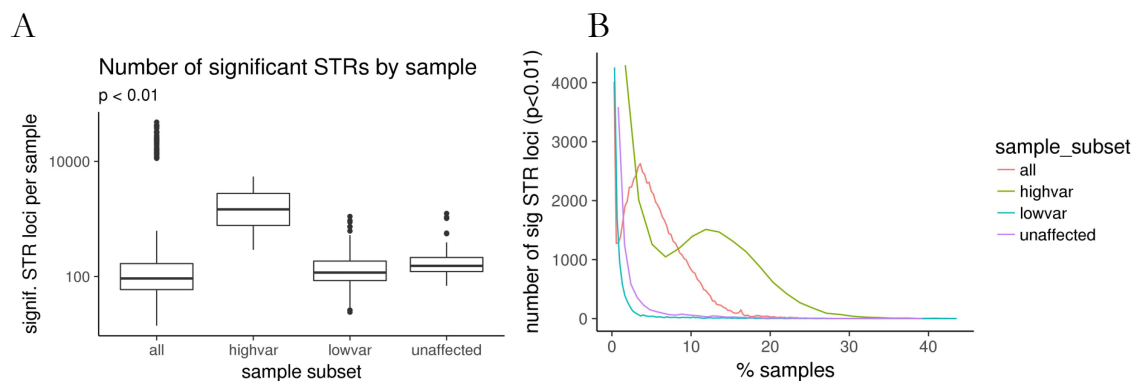


Figure 4.11: Comparing the number of significant ($p < 0.01$) STR loci across the entire set of samples and three different subsets: the previously identified “highvar” samples with more than 10,000 STR loci with any reads, the “lowvar” samples with less than 10,000 STR loci with any reads and the “unaffected” samples.

In each case p values are calculated using all samples in the same set as controls. A) Number of significant STR loci per sample. B) Percent of samples with a given number of significant STR loci. Note that there are also a large number of loci that are significant in none of the samples (0 on the x axis). These are not shown.

To further investigate the impact of control choice we ran STRetch on the same set of 59 highly variable samples using the four different control sets. STRetch called an order of magnitude fewer significant STRs per sample using the “highvar” control set compared with the other three sets of controls

(Figure 4.12A). Even though the “all” control contained all the highvar samples, the use of a robust variance estimate dampened the impact of these samples (which made up a relatively small proportion of the “all” set), resulting in high numbers of significant calls. Looking at the number of loci with significant calls across multiple samples we found that for the “all”, “lowvar” and “unaffected” controls some loci were significant in 100% of samples, and many more are significant in more than 50% of samples, while the “highvar” control set controls this number much better (Figure 4.12B). Since it’s very unlikely that these highvar samples are enriched for true STR expansions this “hump” in the distribution likely represented false positives, however it’s useful to note that the use of similarly variable controls to the samples helped to control the number of these likely spurious significant calls.

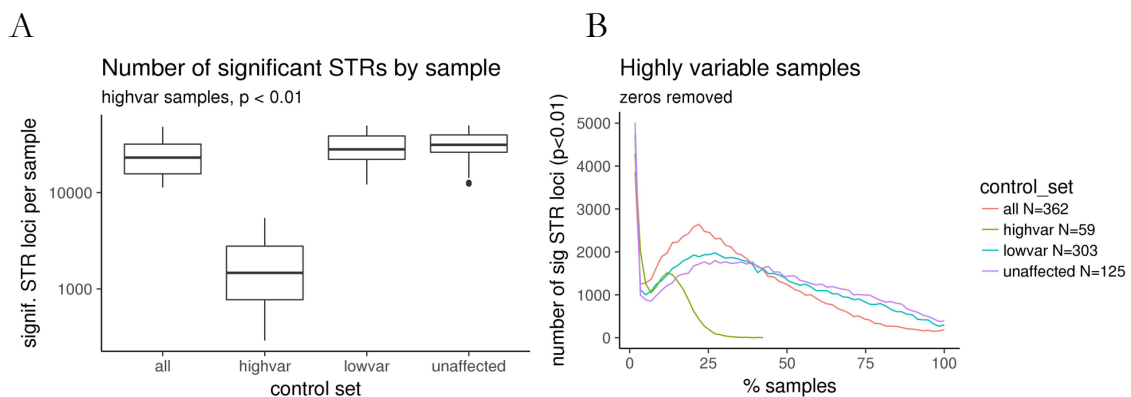


Figure 4.12: Comparing the number of significant STR loci ($p < 0.01$) called in the “highvar”, highly variable samples with more than 10,000 STR loci, using four different control sets: “all” samples, “highvar”, “lowvar” (less than 10,000 STR loci with any reads) and the “unaffected” samples.

A) Number of significant STR loci per sample. B) Percentage of samples with a given number of significant STR loci. Loci that are significant in none of the samples are not shown.

We considered the number of significant expansions called at known pathogenic loci using the four different control sets. The choice of control makes a dramatic difference in the number of significant expansions called in pathogenic loci, with the total number across all the highvar samples ranging

from 16 in with the highvar controls to 50 with the unaffected controls (Table 4.4). In addition to considering a specific p value threshold, we also investigated the magnitude of the p values for these calls (Figure 4.13). For the SCA8 locus two samples received a significant p value while all others fell well below the threshold, and there was clear separation between the significant and non-significant p values. One of these calls was a confirmed SCA8 case described in the *STRetch* paper (Chapter 3). In contrast for the *FTDALS1* and *FXTAS* loci there was minimal separation in the p values between samples, giving us less confidence in the clinical significance of these calls. So that even when using a control set that is not well matched for variability, we can use plots like these to help distinguish possible false positives.

Table 4.4: Number of the 59 highly variable samples with significant expansions ($p < 0.01$) for each pathogenic STRs using each of the controls sets. Loci with no significant results are excluded.

Locus	Control set			
	all	highvar	lowvar	unaffected
DM1_DMPK	0	0	0	1
DM2_ZNF9	1	0	1	1
FRAXE_AFF2	1	1	1	2
FTDALS1_C9orf72	10	6	10	16
FXTAS_FMR1	2	0	2	16
HD_HTT AGC	1	0	2	2
HD_HTT CCG	0	0	1	1
SBMA_AR	1	1	1	1
SCA17_TBP	1	0	1	1
SCA1_ATXN1	1	0	1	1
SCA3/MJD_ATXN3	6	6	6	6
SCA8_ATXN8/ATXN8 OS	2	2	2	2

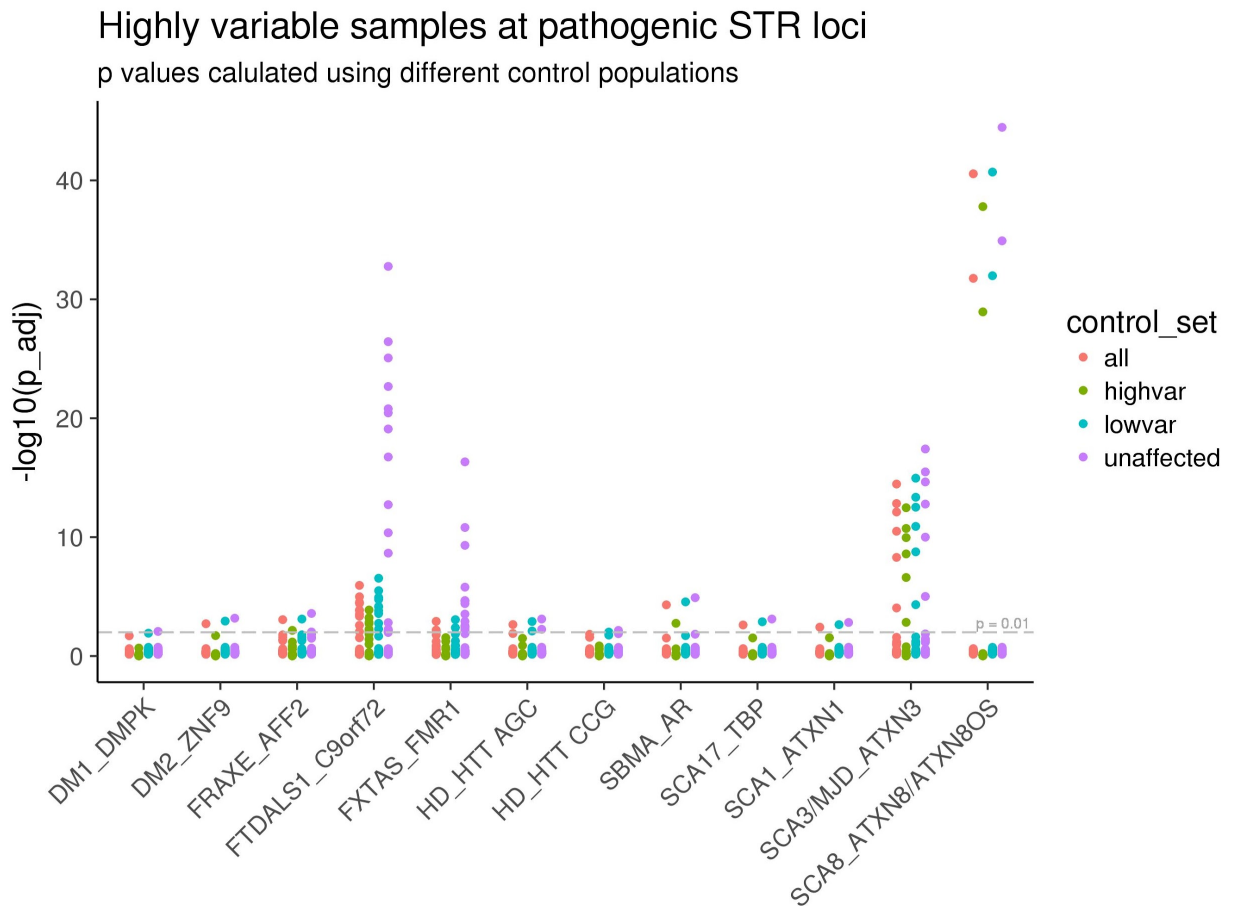


Figure 4.13: STRetch p values for all pathogenic loci with results in the 59 highly variable samples using each of the controls sets.

We also ran STRetch on the 125 unaffected parents' samples using the four different control sets. The largest number of variants called per sample was found when the control set was exactly matched to the samples, with the least variants called when using the "highvar" sample set (Figure 4.14A). The number of significant calls per sample was generally an order of magnitude lower for unaffected samples compared with the highvar samples. In contrast to the highvar samples, the percentage of samples sharing a significant locus was generally under 50% in the unaffected samples across all control sets and the choice of control didn't vary this number as dramatically (Figure 4.14B).

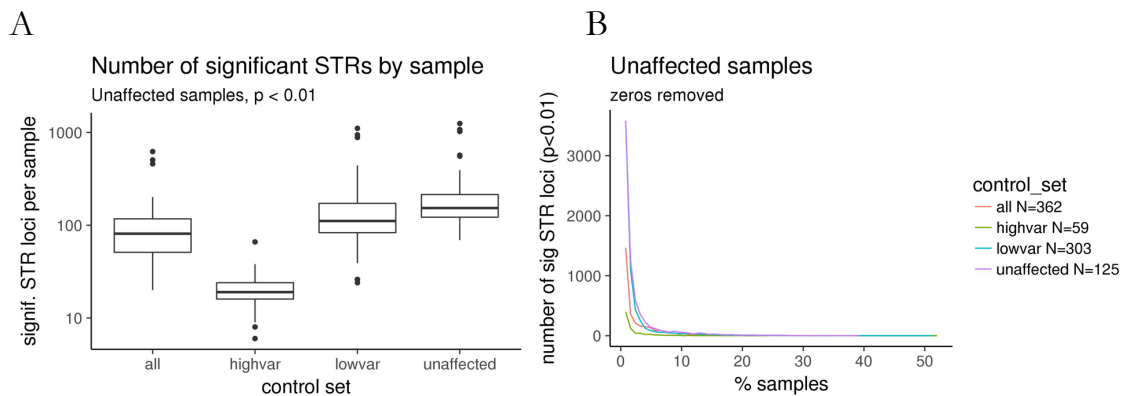


Figure 4.14: Comparing the number of significant STR loci ($p < 0.01$) called in the unaffected samples, using four different control sets: “all” samples, “highvar”, “lowvar” and the “unaffected” samples.

A) Number of significant STR loci per sample. B) Percentage of samples with a given number of significant STR loci. Loci that are significant in none of the samples are not shown.

As we previously described, the unaffected samples had a number of significant expansions in known pathogenic loci (Table 4.5). The largest number of significant calls in pathogenic loci (50 across the unaffected samples) was found using the unaffected controls, while only 15 are found using the highvar controls. The significant calls were again concentrated on the FTDALS1 and FXTAS loci, with FRAXE and SCA3 also making a substantial contribution. Looking at the magnitude of p values, we found very clear separation for the DM2 and SCA8 loci regardless of controls, and the same pattern to a lesser extent in the FRAXE, SBMA, and potentially SCA3 loci (Figure 4.15). While the FTDALS1 p values were closely clustered when using the “all”, “highvar” and “lowvar” controls, some separation became apparent using the “unaffected” controls. As before, most of these alleles were estimated to be below the pathogenic threshold (Figure 4.10B) and those that were in the pathogenic range didn’t fit the symptoms described for these families: SBMA is X linked so this female is not expected to be affected, and SCA8 can be quite late onset and there is some question as to the penetrance of these alleles, so this individual may not yet (or ever) display symptoms. Note that although p values vary when using different controls, STRetch allele

size estimates are only based on the counts in the sample, and so do not vary when changing controls.

Table 4.5: Number of unaffected samples with significant expansions ($p < 0.01$) for each pathogenic STRs using each of the controls sets.

Locus	Control set			
	all	highvar	lowvar	unaffected
DM2_ZNF9	1	1	1	1
FRAXE_AFF2	3	3	3	7
FTDALS1_C9orf72	11	6	16	20
FXTAS_FMR1	1	0	1	15
HD_HTT AGC	0	0	0	1
SBMA_AR	1	1	1	1
SCA3/MJD_ATXN3	3	3	4	4
SCA8_ATXN8/ATXN8 OS	1	1	1	1

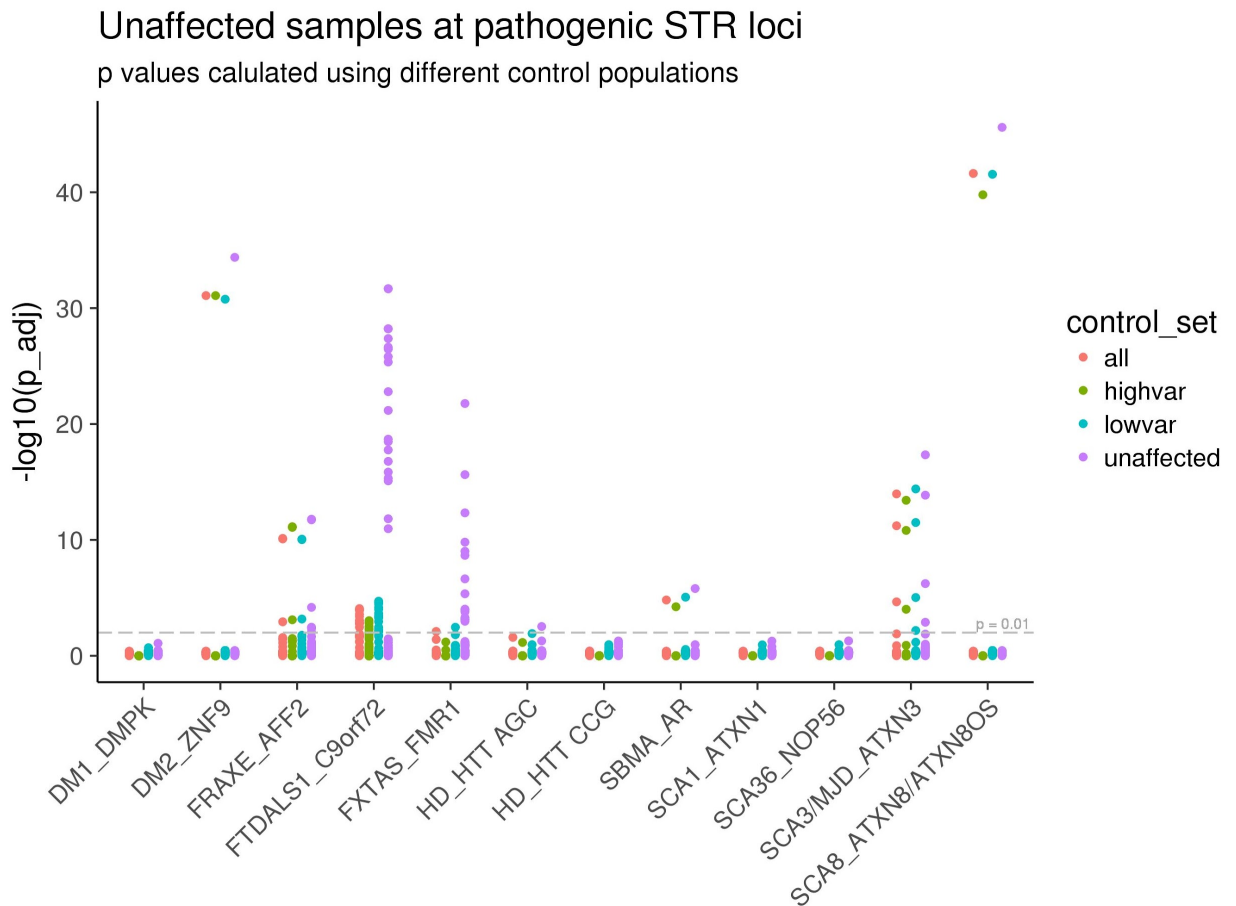


Figure 4.15: STRetch p values for all pathogenic loci with results in the 125 unaffected samples using each of the controls sets.

Together these findings show the importance of control set choice for STRetch. We found that a control set matched in terms of variability better controlled the number of significant calls in highly variable samples. However, when the test samples had little variability using a matched set may exacerbate potential false positives. Despite these findings, in both cases the smallest p values were quite distinct from other calls, so the choice of control set is unlikely to greatly diminish STRetch’s ability to detect large pathogenic expansions. Much of the behaviour of STRetch in relation to controls stems from STRetch’s use of robust estimates of variance; Huber’s robust estimate of variance tends to be much lower than a standard non-robust estimate for most loci (Appendix Figure 6.3). This is a choice that allows some false positives to ensure we call as many true positives as possible. Using these

robust estimates becomes particularly important when few controls are available or when many samples in a control set are affected. As we have previously shown (Chapter 3), STRetch is able to detect a true pathogenic expansion with as few as ten control samples, including multiple affected. The choice of control set may also depend on context. In a clinical setting while it is no doubt useful to reduce the false positive rate, sensitivity to true disease alleles is paramount. Therefore, using a less variable control set may be appropriate. In contrast, in the context of novel pathogenic locus discovery it may be more useful to reduce false positive rate, and therefore appropriate to use a more variable control set or one that closely matches the variability of the test data. In either case, examination of the distribution of p values (such as the visualisation in Figure 4.15) proved useful in distinguishing outlier loci from those that are significant in a larger number of individuals, and the magnitude of the p value for the confirmed pathogenic locus was robust to the choice of controls.

4.7 Discussion

Known pathogenic STR loci are predominantly found in coding regions, although many are found in introns and several are in UTRs (see Table 1.1 in the Introduction). Our survey of STRs annotated in the reference genome indicated a number of potentially promising loci to survey for disease association. For example, we found 623 STR loci overlapping CDS, 4730 over non-coding exons, and 8624 within 3000 bp upstream a TSS. We also found approximately half the annotated STRs in introns, consistent with almost 50% of the genome being annotated as intronic when considering all Gencode transcripts. The finding that RAN translation of STRs as a disease mechanism further widens the pool of potentially pathogenic STRs [75]. Together this makes STRs a potentially rich and largely untapped source of variation to consider for novel pathogenic loci.

We applied STRetch to the analysis of 362 PCR-free whole genomes. We found that many reads aligning to the STR decoys didn't correspond to any annotated STR locus. This may indicate STRs missing from the reference genome. The most obvious example of this was the huge number of STR decoy reads arising from the telomeres repeat unit, with telomeres known to be poorly represented in the hg19 reference genome.

A major challenge for both clinical and research genomics is to distinguish between common and rare (and therefore potentially interesting) variants. Population variation projects, in particular ExAC/gnomAD have become key to making such distinctions for SNVs and short indels. Therefore, we considered how to prioritise loci as more likely to be pathogenic using the data that we had. One way was to consider STR loci that show reduced variation in our samples, potential evidence of evolutionary constraint. We found that coding STRs showed more limited variability, as well as STRs in non-coding exons to a lesser extent. This is consistent with previous findings of constraint for short alleles in coding STRs [152]. This may make these loci a richer source of potentially pathogenic STRs. We also found reduced variability in known pathogenic STR loci, suggesting constraint of these loci. The same previous study found that many pathogenic loci were not under heavy constraint, however they only considered alleles less than the read length. This highlights the need to consider the full spectrum of allele sizes. Results from the limited number of samples in this study combined with the small number of STR loci in these categories showed tantalising trends, but no clear conclusions. Future work would ideally use thousands of unaffected individuals to draw out these ideas and get more accurate population variability estimates for loci.

One goal of this study was to determine the frequency of STR expansions in the population, and thus what are normal or at least common allele sizes. Across all samples, individuals typically had around 100 significant STR

expansions, with around one of these in pathogenic loci. Our subset of 125 unaffected and unrelated individuals had significant expansions in pathogenic loci in many samples, however most were below the pathogenic allele size threshold. Despite this, we were initially surprised to see large but sub-clinical expansions at pathogenic loci, as we would expect strong selection against such variation. An obvious possibility is that these are false positives, however in Chapter 3 we showed that we were able to validate several non-pathogenic expansions by PCR or long-read sequencing, suggesting that these results are also unlikely to be false positives. There is of course the potential for sample swaps, mislabelling or contamination, however an interesting possibility is that these loci truly are expanded but are for some reason not pathogenic. We know that many STR expansion diseases are relatively late-onset, with some individuals only showing symptoms at age 70 or more, making it very difficult to distinguish affected from unaffected individuals. There have also been reports of reduced penetrance, with individuals harbouring pathogenic-length alleles not going on to develop the disease. There may be other protective factors at play (genetic or otherwise) or the disease phenotype may be more variable than we thought. We observed 2/125 (1.6%) samples with significant STR expansions in the pathogenic range. This is consistent with the previous study which found 192/12632 (1.5%) of individuals to harbour a pathogenic-length allele [96].

Although short-read sequencing has limitations in detecting large STR expansions, current laboratory tests also have substantial limitations. The standard approach is to use PCR, however STRs are particularly difficult sequences to PCR effectively. They require a specific test to be developed for each STR locus, making the validation of novel loci a particularly expensive endeavour. They also suffer from limitations in the length of alleles that can be amplified, with many tests setting an estimated lower-bound on the allele size or only amplifying the non-pathogenic allele and thus inferring the

presence of a pathogenic one. This may be used as a screening test prior to using a more expensive assay such as a Southern.

The choice of control samples for STRetch can have a dramatic impact on the number of significant results. In general, more highly variable controls reduced the number of significant loci, while less variable controls increased it. Choosing a control set is therefore an important consideration and likely dependent on the context. For a clinical test a more sensitive test may be desirable, and a lab may wish to use the same set of controls for all samples for consistency. This may come with higher false positive rate. In contrast, in a research gene/variant discovery setting it may be more appropriate to use whatever controls best match the given samples to reduce the number of loci that need to be validated.

This study is limited by the limited dynamic range of STRetch. Many known pathogenic expansions are at our upper limit of ability to distinguish between moderate and very large expansions because the allele size is near to or greater than the sequencing insert size. STRetch is also limited to detecting alleles that are greater than the allele in the reference genome. A future strategy might be to combine a large STR-caller like STRetch with a caller designed to detect alleles within the read length such as LobSTR. Or to further develop STRetch to detect such smaller variants.

Chapter 5 Conclusion

In Chapter 1 I described the proliferation of exome and whole genome sequencing to both diagnose genetic disorders in the clinical setting, and to perform disease gene discovery in the research setting. With diagnostic rates in the range of 31-39% reported by many studies, we must question why we fail to diagnose some patients, and how we can increase diagnostic rates [2]. So why might we fail to make a genetic diagnosis? First, it's possible that the condition does not have a genetic basis, i.e., a phenocopy. Second, we may not have sequenced the causal variant (e.g. an intergenic variant when doing exome sequencing). Third, we sequenced the variant but the bioinformatic analysis failed to call the variant (e.g. STRs). Finally, we may have genotyped the variant, but filtered it out, or failed to prioritise it as causative. This can happen when we lack understanding of the impact of a specific variant.

In this thesis I address the issue of increasing genetic diagnostic rates from multiple perspectives. First, in Chapter 2, I describe a sequencing and analysis strategy to prioritise *de novo* variants, and thus reduce the number of variants that need to be curated. In Chapter 3 I describe a method that I have developed to detect potentially pathogenic STR expansions from short-read data, which are currently not routinely tested for in exome or genome sequencing. Finally, in Chapter 4, I explore data-driven methods to prioritise STR expansions when doing gene-discovery. In the following section I will describe in detail the scientific contributions I made in each chapter.

Scientific contributions

In Chapter 2 I developed and tested a powerful new sequencing strategy for prioritising *de novo* variants using pooled-parent exome sequencing. This method provides a trade-off between singleton and trio sequencing. Several

probands with suspected *de novo* disease are sequenced as individual exomes, then a pool of the DNA from all their parents is sequenced together in a single exome capture. This allows us to filter out variants seen in the parent pool as being inherited and so reduce the number of proband variants that need to be considered.

Using simulations, I showed that pooling together up to ten samples still allowed recall of 94-98% of variants, and that recall rate was increased by sequencing more deeply. I found that the GATK HaplotypeCaller individual variant calling strategy provided the best recall for singleton variants (those that only occur once in the pool). In an analysis of a real pooled-parent sequencing experiment I showed that over 81% of variants in the probands could be filtered using the pool and that the parent filter was complementary to a population frequency variant filter using gnomAD.

The pooled-parents exome strategy addresses a critical tension between the diagnostic value of sequencing parents and the need to sequence as many patients as possible within a given budget. It provides a way to reduce the number of variants that need to be curated for a given patient while keeping costs to only slightly more than that of a single exome.

The largest portion of this thesis focused on Short Tandem Repeat (STR) expansions. In Chapter 3 I introduced STRetch, a new method that I developed to detect large STR expansions from short-read Illumina sequencing data, which has been published in *Genome Biology* [104]. STRetch takes a completely novel approach to detecting reads arising from STR expansions. It creates “STR decoy chromosomes”, one for each of all possible STR repeat units, and adds them to the reference genome. Reads arising from expanded STR alleles preferentially map to these decoy chromosomes. STRetch then assigns these to annotated STR loci using a uniquely mapping read-pair. STRetch then uses counts of these reads for each

locus to both estimate the allele size and to compare across individuals to find significantly expanded STR loci in this sample relative to others.

I described the validation of STRetch on a set of individuals with known STR disease loci. STRetch showed a sensitivity of 0.778, specificity of 0.974, and a false discovery rate of 0.025 on known STR disease loci. I then showed that STRetch could detect expansions at STR loci that have not been associated with disease. STRetch's predictions were confirmed using long-read sequencing, as well as PCR. I further ran STRetch on a set of 97 individuals to assess the frequency of expansions at known pathogenic loci. Although STRetch was originally developed for whole genome sequencing data, it has been since been shown to be competitively sensitive on exome sequencing data, compared to other tools [115].

I described how STRetch was able to identify a pathogenic SCA8 allele in a previously undiagnosed patient. This was confirmed with PCR and led to a formal genetic diagnosis for this individual and their affected sibling, after a diagnostic odyssey lasting many years. This patient had already undergone extensive genetic testing for individual genes and loci (including PCR tests for other STR loci), finally culminating in whole genome sequencing and analysis for SNVs and indels by the MacArthur lab. This shows the value of detecting STR expansions from short-read sequencing data, as it enables simultaneous testing for STRs alongside SNVs, indels and other variants. As a disease known to be caused by expansions in more than 10 STR loci as well as numerous short variants in other genes, ataxia is a classic example of a disease where whole genome sequencing with multiple methods of analysis may result in a faster and potentially cheaper diagnosis than traditional diagnostic methods.

The STRetch code and reference data have been made freely available to the research and clinical communities at <https://github.com/Oshlack/STRetch>.

It can be difficult to estimate the impact of bioinformatics software until many years later, as research projects can take years to publish, or may not cite the method. However, I have already had several personal communications from researchers describing how STRetch has been useful.

In particular, STRetch has already been used in the discovery of a novel pathogenic STR locus. Researchers at the University of Washington used STRetch to detect a pathogenic GGC expansion in exon 1 of the XYLT1 gene, and found this variant to cause changes to methylation, resulting in Baratela-Scott Syndrome [163]. This pathogenic STR locus is *de novo*; the STR sequence is not found in the reference genome, only in the individuals. Even though STRetch relies on the reference genome to determine the positions of STR loci, the authors were able to use it to detect and visualise reads arising from this expansion using the STRetch decoy chromosomes. Once found, this locus could be added to reference annotation, and STRetch could generate calls for the novel locus. This highlights the potential of the STR decoy chromosome method to detect reads from STR expansions that are not in the reference genome. This concept is elaborated upon in Chapter 4.

STRetch has also aided in the discovery of another *de novo* pathogenic STR expansion that causes cerebellar ataxia with neuropathy and bilateral vestibular areflexia syndrome (CANVAS). The novel RFC1 gene pathogenic STR locus actually arises from the replacement of an AAAAG STR locus in the genome with a AAGGG expansion carried on an Alu element in the affected individuals [164]. In this case, rather than detecting the *de novo* STR locus, after discovery the locus was added to the reference and STRetch was used to detect the locus in a collection of samples and controls to confirm its cosegregation with disease.

Both the Baratela-Scott Syndrome and CANVAS stories demonstrate that the search for pathogenic STR loci should not be limited to STR loci in the

reference genome. We have seen that STRetch's STR decoy chromosomes can be used to identify reads that arose from *de novo* STR expansions. Future work on STRetch is planned to extend the algorithm to detect *de novo* STR expansions. The informative reads are already captured by the current method; however, development would be required to assign reads from novel expansions to a specific genomic location.

In Chapter 4 I took the STRetch method and many of the ideas developed in Chapter 3, and extended them to a more extensive analysis of STR expansions in a larger set of more than 300 samples. I first described the distribution of all STRs annotated in the human genome. I then used the counts of reads aligning to the STR decoy chromosomes to detect repeat units that are observed more frequently in the samples than we would expect based on the reference genome. For example, as expected, the telomeric repeat unit was prevalent in the samples, even though the reads were not allocated to specific loci, because they are under-represented in the reference genome. More surprisingly however, I found an excess of C/G homopolymers in these samples, even though these loci are rare in the reference genome. This may point to a systematic under-representation of these sequences in the reference genome, possibly due to their difficulty in sequencing due to high GC content.

I used STRetch to explore the frequency of STR expansions in a set of 362 exomes, predominantly from families with suspected genetic disease, including a subset of 125 unaffected and unrelated individuals. On average, individuals had around 100 significant STR expansions. Many samples had a significant expansion in a known pathogenic STR, however the majority of these were below the pathogenic range. I further used this data to explore the value of genomic context and constraint (low variance) as a strategy to prioritise potentially pathogenic STRs in a gene-discovery setting. Finally, I commented on the impact of control choice when running STRetch in research and clinical contexts. A major contribution of Chapter 4, is the idea that known

pathogenic STR loci tend to be less variable, and so constraint at other STR loci may be evidence to support pathogenicity of an expansion at that loci.

Discussion

This thesis highlights the central role of bioinformatics, and in particular bioinformatics methods development, in our efforts to increase genetic diagnostic rates in clinical settings, and to discover new disease genes. From read alignment and variant calling, through to variant prioritisation, bioinformatics has been critical in enabling the very foundations of modern clinical genomics. Many of the bioinformatics tools used in diagnostic laboratories today are still under active development, with new algorithms constantly transitioning from the bioinformatics research community into the clinic. There is still huge scope for the development of new bioinformatics methods. For example, I have described the rush to develop algorithms to effectively detect STR expansions, and the many gaps that still need to be filled in this area, especially with respect to the detection of *de novo* STRs.

More broadly, short next-generation sequencing reads have proven a challenge to the detection of repetitive variants such as transposable elements, pseudogenes, centromeres, and longer tandem repeats. While short reads limit our ability to detect some variants, they also provide the opportunity for ingenious bioinformatics solutions to making inferences from this incomplete data. Some of the strategies described in this thesis to detect STR expansions could be applied more broadly. For example, transposable element decoy sequences could be used to detect reads arising from novel insertion sites and then paired read information could locate the insertion site.

Many researchers look to new technologies, such as long-read sequencing, as the answer to our current inability to detect large or complex variants, such as STR expansions or structural rearrangements. Although these technologies are

currently cost-prohibitive, they certainly provide a valuable view of variation. However, we must remember that these methods come with their own issues and biases. As bioinformaticians, it is our role to critically analyse these biases and to develop algorithms to extract and distinguish biologically meaningful information from noise.

In many diagnostic laboratories, variants found by next-generation sequencing are being validated by secondary methods, such as PCR. As we move towards large-scale uptake of genomics in the diagnostic setting, this labour-intensive validation is becoming untenable. We need to move towards high-throughput sequencing as the primary test. For this to be possible we need to be rigorous in the quality of our bioinformatics software as it moves from research to the clinical arena. We need to be mindful of testing our software just as thoroughly as we would test a new diagnostic PCR assay. We need to be informed about the way we do genomics at scale, especially the suitability of controls and population references, and the unintended biases we introduce when we make these decisions. As I described in Chapter 1, even the human reference genome, the very foundation of many of our bioinformatics methods, has known limitations. We must be vigilant in constantly reassessing the quality of our data and of our assumptions as we increasingly rely on genomics to make medical decisions.

As genomic sequencing moves from the cutting-edge of research into everyday vernacular and experience, and as the number of individuals sequenced world-wide increases exponentially, we march towards a society where children may be sequenced at birth. As researchers we feel the keen responsibility to contribute the best science and the best software that we are capable of. Bioinformatics is the unsung hero of the genomics revolution.

References

1. Stevenson DA, Carey JC. Contribution of malformations and genetic disorders to mortality in a children's hospital. *Am J Med Genet* [Internet]. 2004 [cited 2019 May 28];126A:393–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15098237>
2. Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, et al. A meta-analysis of the diagnostic sensitivity and clinical utility of genome sequencing, exome sequencing and chromosomal microarray in children with suspected genetic diseases. *bioRxiv* [Internet]. Cold Spring Harbor Laboratory; 2018 [cited 2018 Jul 9];255299. Available from: <https://www.biorxiv.org/content/early/2018/01/30/255299>
3. Stark Z, Tan TY, Chong B, Brett GR, Yap P, Walsh M, et al. A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genet Med* [Internet]. Murdoch Childrens Research Institute, Melbourne, Australia. University of Melbourne, Melbourne, Australia. Melbourne Genomics Health Alliance, Melbourne, Australia. Royal Children's Hospital, Melbourne, Australia. Royal Women's Hospital, Melbourne, Australia; 2016;18:1090–6. Available from: <http://dx.doi.org/10.1038/gim.2016.1>
4. Bird TD. Hereditary Ataxia Overview [Internet]. GeneReviews®. University of Washington, Seattle; 1993 [cited 2019 Sep 26]. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20301317>
5. Osoegawa K, Mammoser AG, Wu C, Frengen E, Zeng C, Catanese JJ, et al. A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res* [Internet]. Cold Spring Harbor Laboratory Press; 2001 [cited 2019 May 24];11:483–96. Available from:

<http://www.ncbi.nlm.nih.gov/pubmed/11230172>

6. Consortium T 1000 GP, Consortium 1000 Genomes Project, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* [Internet]. Nature Publishing Group; 2012 [cited 2019 May 24];491:56–65. Available from: <http://dx.doi.org/10.1038/nature11632>

7. Kehr B, Helgadóttir A, Melsted P, Jonsson H, Helgason H, Jonasdóttir A, et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nat Genet* [Internet]. Nature Publishing Group; 2017 [cited 2019 May 24];49:588–93. Available from: <http://www.nature.com/articles/ng.3801>

8. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* [Internet]. Cold Spring Harbor Laboratory Press; 2017 [cited 2019 May 24];27:849–64. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28396521>

9. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon DS, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* [Internet]. 2017 [cited 2019 May 24];677–85. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27895111>

10. Karki R, Pandya D, Elston RC, Ferlini C. Defining “mutation” and “polymorphism” in the era of personal genomics. *BMC Med Genomics* [Internet]. BioMed Central; 2015;8:37. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/26173390>

11. Tattini L, D’Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. *Front Bioeng Biotechnol* [Internet].

Frontiers Media SA; 2015 [cited 2019 May 24];3:92. Available from:
<http://www.ncbi.nlm.nih.gov/pubmed/26161383>

12. Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nat Rev Genet* [Internet]. NIH Public Access; 2016 [cited 2019 Jun 14];17:19–32. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26593421>

13. Mattick JS, Makunin I V. Non-coding RNA. [cited 2019 May 30]; Available from: https://academic.oup.com/hmg/article-abstract/15/suppl_1/R17/632705

14. Ensembl. Calculated consequences [Internet]. [cited 2019 Jun 1]. Available from:
http://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences

15. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* [Internet]. National Academy of Sciences; 1977 [cited 2019 May 21];74:5463–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/271968>

16. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics* [Internet]. Elsevier; 2016 [cited 2019 May 21];107:1–8. Available from:
<http://www.ncbi.nlm.nih.gov/pubmed/26554401>

17. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl* [Internet]. NIH Public Access; 2014 [cited 2019 May 21];1. Available from:
<http://www.ncbi.nlm.nih.gov/pubmed/25699289>

18. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* [Internet]. Narnia; 2012

[cited 2019 May 21];40:e72–e72. Available from:

<https://academic.oup.com/nar/article/40/10/e72/2411059>

19. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol* [Internet]. 2011 [cited 2019 May 21];12:R112. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22067484>

20. Gymrek M. PCR-free library preparation greatly reduces stutter noise at short tandem repeats [Internet]. 2016 Mar. Available from: <http://biorxiv.org/lookup/doi/10.1101/043448>

21. Huptas C, Scherer S, Wenning M. Optimized Illumina PCR-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly. *BMC Res Notes* [Internet]. BioMed Central; 2016 [cited 2019 May 21];9:269. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27176120>

22. Van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology A Brief History of Sequencing Technology. 2018 [cited 2019 May 21]; Available from: <https://doi.org/10.1016/j.tig.2018.05.008>

23. Travers KJ, Chin C-S, Rank DR, Eid JS, Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res* [Internet]. Narnia; 2010 [cited 2019 May 21];38:e159–e159. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq543>

24. Jain M, Tyson JR, Loose M, Ip CLC, Eccles DA, O’Grady J, et al. MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Research* [Internet]. 2017 [cited 2019 May 21];6:760. Available from: <https://f1000research.com/articles/6-760/v1>

25. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. *Hum Mol Genet* [Internet]. Narnia; 2018 [cited 2019 May 21];27:R234–41. Available from: <https://academic.oup.com/hmg/article/27/R2/R234/4996216>
26. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet* [Internet]. Nature Publishing Group; 2013 [cited 2019 May 28];14:295–300. Available from: <http://www.nature.com/articles/nrg3463>
27. Petrikin JE, Willig LK, Smith LD, Kingsmore SF. Rapid whole genome sequencing and precision neonatology. *Semin Perinatol* [Internet]. 2015 [cited 2019 May 28];39:623–31. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26521050>
28. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* [Internet]. Nature Publishing Group; 2015 [cited 2019 May 24];17:405–23. Available from: <http://www.nature.com/articles/gim201530>
29. Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, et al. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife*. eLife Sciences Publications, Ltd; 2019;8.
30. Dietz HC. New Therapeutic Approaches to Mendelian Disorders. Feero WG, Guttmacher AE, editors. *N Engl J Med* [Internet]. Massachusetts Medical Society ; 2010 [cited 2019 May 28];363:852–63. Available from: <http://www.nejm.org/doi/10.1056/NEJMra0907180>
31. Sadedin SPSP, Dashnow H, James PAPA, Bahlo M, Bauer DCDC, Lonie A, et al. Cpipe: a shared variant detection pipeline designed for diagnostic

settings. *Genome Med* [Internet]. BioMed Central; 2015;7:68. Available from: <http://dx.doi.org/10.1186/s13073-015-0191-x>

32. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr* [Internet]. 2013;00:3. Available from: <http://arxiv.org/abs/1303.3997>

33. Gymrek M. A genomic view of short tandem repeats. *Curr Opin Genet Dev* [Internet]. Elsevier Current Trends; 2017 [cited 2019 May 23];44:9–16. Available from: <http://dx.doi.org/10.1016/j.gde.2017.01.012>

34. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples [Internet]. *bioRxiv*. 2018. Available from: <https://www.biorxiv.org/content/early/2018/07/24/201178>

35. Institute DSP@ B. GATK | BP Doc #11145 | Germline short variant discovery (SNPs + Indels) [Internet]. 2018. Available from: <https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145>

36. Consortium 1000 Genomes Project, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature* [Internet]. 2015;526:68–74. Available from: <http://dx.doi.org/10.1038/nature15393>

37. Lek M, Karczewski KJ, Minikel E V, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* [Internet]. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. School of Paediatrics and Chil; 2016;536:285–91. Available from: <http://dx.doi.org/10.1038/nature19057>

38. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* [Internet]. 2010 [cited 2013 Nov 7];38:e164. Available from: <http://nar.oxfordjournals.org/content/38/16/e164.abstract>
39. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol* [Internet]. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. wm2@ebi.ac.uk. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, C; 2016;17:122. Available from: <http://dx.doi.org/10.1186/s13059-016-0974-4>
40. Vaser R, Adusumalli S, Ngak Leng S, Sikic M, Ng PC. SIFT missense predictions for genomes. *Nat Protoc* [Internet]. 2015 [cited 2019 May 30];11. Available from: <http://sift-dna.org/sift4g>,
41. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods* [Internet]. 2010 [cited 2019 May 30];7:248–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20354512>
42. González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* [Internet]. Research Programme on Biomedical Informatics, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona. abel.gonzalez@upf.edu; 2011;88:440–9. Available from: <http://dx.doi.org/10.1016/j.ajhg.2011.03.004>
43. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* [Internet]. 1] Department of Genome Sciences, University

of Washington, Seattle, Washington, USA. [2].: Nature Pub. Co.; 2014;46:310–

5. Available from:

<https://ezp.lib.unimelb.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=mnh&AN=24487276&site=eds-live&scope=site>

44. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* [Internet]. 2018 [cited 2019 May 30];46:D1062–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/29165669>

45. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Hum Mutat*. John Wiley and Sons Inc.; 2015;36:915–21.

46. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP) [Internet]. Seattle, WA; 2018. Available from: <http://evs.gs.washington.edu/EVS/>

47. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes [Internet]. *bioRxiv*. 2019. Available from: <https://www.biorxiv.org/content/10.1101/531210v2>

48. Robinson PN, Köhler S, Oellrich A, Sanger Mouse Genetics Project SMG, Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res* [Internet]. Cold Spring Harbor Laboratory Press; 2014 [cited 2019 May 30];24:340–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24162188>

49. Stark Z, Dashnow H, Lunke S, Tan TYTY, Yeung A, Sadedin S, et al. A clinically driven variant prioritization framework outperforms purely computational approaches for the diagnostic analysis of singleton WES data.

Eur J Hum Genet [Internet]. Nature Publishing Group; 2017 [cited 2019 May 24];25:1268–72. Available from:

<http://www.nature.com/articles/ejhg2017123>

50. Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat [Internet]. John Wiley & Sons, Ltd; 2011 [cited 2019 Jun 1];32:557–63. Available from: <http://doi.wiley.com/10.1002/humu.21438>

51. seqr [Internet]. 2019. Available from: <https://github.com/macarthurlab/seqr>

52. Online Mendelian Inheritance in Man, OMIM® [Internet]. McKusick-Nathans Inst. Genet. Med. Johns Hopkins Univ. (Baltimore, MD). [cited 2016 Feb 26]. Available from: <http://omim.org/>

53. NLM. Microsatellite Repeats [Internet]. Med. Subj. Head. National Institutes of Health; 2011. Available from: http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?mode=&term=Microsatellite+Repeats&field=entry#TreeG02.111.570.080.708.800.500

54. NLM. Minisatellite Repeats [Internet]. Med. Subj. Head. National Institutes of Health; 2011. Available from: http://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?mode=&term=Variable+Number+of+Tandem+Repeats

55. Gemayel R, Vincens MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet [Internet]. 2010 [cited 2014 Jan 21];44:445–77. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20809801>

56. Urquhart a, Kimpton CP, Downes TJ, Gill P. Variation in short tandem repeat sequences--a survey of twelve microsatellite loci for use as forensic

identification markers. *Int J Legal Med* [Internet]. Central Research and Support Establishment, Forensic Science Service, Birmingham, UK.;

1994;107:13–20. Available from:

<http://www.ncbi.nlm.nih.gov/pubmed/7999641>

57. Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* [Internet]. 2003;4:R13. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=151303&tool=pmcentrez&rendertype=abstract>

58. Subirana J a, Messeguer X. The most frequent short sequences in non-coding DNA. *Nucleic Acids Res* [Internet]. 2010 [cited 2014 Jan 21];38:1172–81. Available from:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2831315&tool=pmcentrez&rendertype=abstract>

59. Legendre M, Pochet N, Pak T, Verstrepen KJ. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res*. 2007;17:1787–96.

60. Hamada H, Seidman M, Howard BH, Gorman CM. Enhanced gene expression by the poly (dT-dG). poly (dC-dA) sequence. *Mol Cell Biol*. 1984;4:2622–30.

61. Sawaya SM, Bagshaw AT, Buschiazzo E, Gemmell NJ. Promoter microsatellites as modulators of human gene expression. *Tandem Repeat Polymorphisms*. Springer; 2012. p. 41–54.

62. Sawaya SM, Lennon D, Buschiazzo E, Gemmell N, Minin VN. Measuring microsatellite conservation in mammalian evolution with a phylogenetic birth–death model. *Genome Biol Evol*. Oxford University Press; 2012;4:748–59.

63. Young LJ, Wang ZX. The neurobiology of pair bonding. *Nat Neurosci* [Internet]. 2004 [cited 2014 Jan 22];7:1048–54. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15452576>
64. Hammock E a D, Young LJ. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* (80-) [Internet]. 2005;308:1630–4. Available from: <http://www.sciencemag.org/cgi/doi/10.1126/science.1111427>
65. Walum H, Westberg L, Henningsson S, Neiderhiser JM, Reiss D, Igl W, et al. Genetic variation in the vasopressin receptor 1a gene (AVPR1A) associates with pair-bonding behavior in humans. *Proc Natl Acad Sci*. 2008;105:14153–6.
66. Tansey K, Hill M, Cochrane L, Gill M, Anney R, Gallagher L. Functionality of promoter microsatellites of arginine vasopressin receptor 1A (AVPR1A): implications for autism. *Mol Autism* [Internet]. 2011;2:3. Available from: <http://www.molecularautism.com/content/2/1/3>
67. Knafo A, Israel S, Darvasi A, Bachner-Melman R, Uzefovsky F, Cohen L, et al. Individual differences in allocation of funds in the dictator game associated with length of the arginine vasopressin 1a receptor RS3 promoter region and correlation between RS3 length and hippocampal mRNA. *Genes, Brain Behav*. 2008;7:266–75.
68. Maher BS, Vladimirov VI, Latendresse SJ, Thiselton DL, McNamee R, Kang M, et al. The AVPR1A gene and substance use disorders: association, replication, and functional evidence. *Biol Psychiatry*. 2011;70:519–27.
69. Oh D-B, Kim Y-G, Rich A. Z-DNA-binding proteins can act as potent effectors of gene expression in vivo. *Proc Natl Acad Sci*. 2002;99:16666–71.
70. Rothenburg S, Koch-Nolte F, Rich A, Haag F. A polymorphic

- dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proc Natl Acad Sci U S A*. 2001;98:8985–90.
71. van Ham SM, van Alphen L, Mooi FR, van Putten JPM. Phase variation of *H. influenzae* fimbriae: transcriptional control of two divergent genes through a variable combined promoter region. *Cell*. 1993;73:1187–96.
72. Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet*. 2009;10:161–72.
73. Gymrek M, Willems T, Zeng H, Markus B, Daly MJ, Price AL, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet* [Internet]. Nature Publishing Group; 2016;48:22–9. Available from:
<http://biorxiv.org/content/early/2015/04/02/017459.abstract>
<http://www.ncbi.nlm.nih.gov/pubmed/26642241>
74. van Eyk CL, Richards RI. Dynamic Mutations. *Tandem Repeat Polymorphisms*. Springer; 2012. p. 55–77.
75. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nat Publ Gr* [Internet]. Nature Publishing Group; 2018; Available from: <http://dx.doi.org/10.1038/nrg.2017.115>
76. Mirkin SM. Expandable DNA repeats and human disease. *Nature* [Internet]. 2007 [cited 2014 Jan 13];447:932–40. Available from:
<http://www.ncbi.nlm.nih.gov/pubmed/17581576>
77. Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Stephens K, et al. *GeneReviews* [Internet]. Univ. Washington, Seattle. 2018. Available from:
<https://www.ncbi.nlm.nih.gov/books/NBK1116/>
78. Sherman SL, Jacobs PA, Morton NE, Froster-Iskenius U, Howard-Peebles PN, Nielsen KB, et al. Further segregation analysis of the fragile X

syndrome with special reference to transmitting males. *Hum Genet.* 1985;69:289–99.

79. Hannan AJ. Tandem Repeat Polymorphisms: Genetic Plasticity, Neural Diversity and Disease. Hannan AJ, editor. *Adv. Exp. Med. Biol.* Austin/New York: Landes Bioscience/Springer Science+Business Media; 2012.

80. Paradisi I, Ikonomu V, Arias S. Huntington disease-like 2 (HDL2) in Venezuela: frequency and ethnic origin. *J Hum Genet.* Nature Publishing Group; 2013;58:3.

81. Laffita-Mesa JM, Pupo JMR, Sera RM, Mojena YV, Kour\`i V, Laguna-Salvia L, et al. De novo mutations in ataxin-2 gene and ALS risk. *PLoS One.* Public Library of Science; 2013;8:e70560.

82. Gan S-R, Ni W, Dong Y, Wang N, Wu Z-Y. Population genetics and new insight into range of CAG repeats of spinocerebellar ataxia type 3 in the Han Chinese population. *PLoS One.* Public Library of Science; 2015;10:e0134405.

83. Veneziano L, Mantuano E, Catalli C, Gellera C, Durr A, Romano S, et al. A shared haplotype for dentatorubropallidolusian atrophy (DRPLA) in Italian families testifies of the recent introduction of the mutation. *J Hum Genet.* Nature Publishing Group; 2014;59:153.

84. Polling S, Hill AF, Hatters DM. Polyglutamine aggregation in Huntington and related diseases. *Tandem Repeat Polymorphisms.* Springer; 2012. p. 125–40.

85. Shoubridge C, Gecz J. Polyalanine tract disorders and neurocognitive phenotypes. *Tandem Repeat Polymorphisms.* Springer; 2012. p. 185–203.

86. Zhu L, Li Z, Wang Y, Zhang C, Liu Y, Qu X. Microsatellite instability and survival in gastric cancer: A systematic review and meta-analysis. *Mol Clin Oncol [Internet].* 2015;3:699–705. Available from:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4471517&tool=pmcentrez&rendertype=abstract>

87. Karran P. Microsatellite instability and DNA mismatch repair in human cancer. *Semin Cancer Biol* [Internet]. 1996;7:15–24. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8695762>

88. Shah SN, Hile SE, Eckert KA. Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. *Cancer Res*. 2010;70:431–5.

89. Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, Rüschoff J, et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J Natl Cancer Inst*. Oxford University Press; 2004;96:261–8.

90. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, et al. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res*. AACR; 1998;58:5248–57.

91. Ni Huang M, McPherson JR, Cutcutache I, Teh BT, Tan P, Rozen SG. MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations. *Sci Rep* [Internet]. Nature Publishing Group; 2015 [cited 2019 May 30];5:13321. Available from: <http://www.nature.com/articles/srep13321>

92. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* [Internet]. Narnia; 2014 [cited 2019 May 30];30:1015–6. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt755>

93. Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC. Microsatellite instability detection by next generation sequencing. *Clin Chem [Internet]. Clinical Chemistry*; 2014 [cited 2019 May 30];60:1192–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24987110>
94. Guichoux E, Lagache L, Wagner S, Chaumeil P, LÉGer P, Lepais O, et al. Current trends in microsatellite genotyping. *Mol Ecol Resour [Internet]*. 2011 [cited 2014 Jan 22];11:591–611. Available from: <https://ezp.lib.unimelb.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=61268164&site=eds-live>
95. Pompanon F, Bonin A, Bellemain E, Taberlet P. Genotyping errors: causes, consequences and solutions. *Nat Rev Genet [Internet]. Nature Publishing Group*; 2005 [cited 2019 May 23];6:847. Available from: www.nature.com/reviews/genetics
96. Tang H, Kirkness EF, Lippert C, Biggs WH, Fabani M, Guzman E, et al. Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *Am J Hum Genet [Internet]*. 2017 [cited 2019 May 17];101:700–15. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/29100084>
97. Shinde D, Lai Y, Sun F, Arnheim N. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)_n and (A/T)_n microsatellites. *Nucleic Acids Res [Internet]*. 2003 [cited 2014 Jan 23];31:974–80. Available from: <http://nar.oxfordjournals.org/content/31/3/974.abstract>
98. Margolis RL. The spinocerebellar ataxias: Order emerges from chaos. *Curr Neurol Neurosci Rep [Internet]*. 2002;2:447–56. Available from: <http://dx.doi.org/10.1007/s11910-002-0072-8>
99. Lelieveld SH, Spielmann M, Mundlos S, Veltman J a, Gilissen C. Comparison of Exome and Genome Sequencing Technologies for the

Complete Capture of Protein-Coding Regions. *Hum Mutat* [Internet].

2015;36:815–22. Available from:

<http://www.ncbi.nlm.nih.gov/pubmed/25973577>

100. Ummat A, Bashir A. Resolving complex tandem repeats with long reads. *Bioinformatics*. 2014;30:3491–8.

101. Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, et al. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene.

Genome Res [Internet]. 2012/10/16. Biochemistry and Molecular Medicine, University of California, Davis, School of Medicine; 2013 [cited 2014 Feb

3];23:121–8. Available from:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3530672&tool=pmcentrez&rendertype=abstract>

102. Doi K, Monjo T, Hoang PH, Yoshimura J, Yurino H, Mitsui J, et al.

Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. *Bioinformatics*. 2014;30:815–22.

103. Mousavi N, Shleizer-Burko S, Gymrek M, Yanicky R, Gymrek M.

Profiling the genome-wide landscape of tandem repeat expansions [Internet].

bioRxiv Cold Spring Harbor Laboratory; Sep 5, 2018 p. 361162. Available

from: <https://www.biorxiv.org/content/10.1101/361162v2>

104. Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, et al.

STretch: detecting and discovering pathogenic short tandem repeat

expansions. *Genome Biol* [Internet]. BioMed Central; 2018 [cited 2019 May

17];19:121. Available from:

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1505-2>

105. Warshauer DH, Lin D, Hari K, Jain R, Davis C, LaRue B, et al. STRait

Razor: a length-based forensic STR allele-calling tool for use with second

generation sequencing data. *Forensic Sci Int Genet.* Elsevier; 2013;7:409–17.

106. Warshauer DH, King JL, Budowle B. STRait Razor v2. 0: the improved STR allele identification tool–Razor. *Forensic Sci Int Genet.* Elsevier; 2015;14:182–6.

107. Van Neste C, Vandewoestyne M, Van Criekinge W, Deforce D, Van Nieuwerburgh F. My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing. *Forensic Sci Int Genet.* Elsevier; 2014;9:1–8.

108. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* [Internet]. 2012;22:1154–62. Available from:

<http://genome.cshlp.org/content/early/2012/04/19/gr.135780.111.abstract>

109. Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* [Internet]. 2013 [cited 2014 Feb 6];41:e32. Available from:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3592458&tool=pmcentrez&rendertype=abstract>

110. Tae H, Kim DY, McCormick J, Settlage RE, Garner HR. Discretized Gaussian mixture for genotyping of microsatellite loci containing homopolymer runs. *Bioinformatics.* 2014;30:652–9.

111. Fungtammasan A, Ananda G, Hile SE, Su MS, Sun C, Harris R, et al. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. 2015;1–14.

112. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. *Nat Publ Gr*

[Internet]. Nature Publishing Group; 2016 [cited 2019 May 23];14:590–2.

Available from: <http://biorxiv.org/lookup/doi/10.1101/077727>

113. Cao MD, Tasker E, Willadsen K, Imelfort M, Vishwanathan S, Sureshkumar S, et al. Inferring short tandem repeat variation from paired-end short reads. *Nucleic Acids Res* [Internet]. 2014 [cited 2014 Dec 11];42:e16.

Available from:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3919575&tool=pmcentrez&rendertype=abstract>

114. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res* [Internet]. 2017;27:1895–903.

Available from: <http://genome.cshlp.org/content/27/11/1895.abstract>

115. Tankard RM, Bennett MF, Degorski P, Delatycki MB, Lockhart PJ, Bahlo M. Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. *Am J Hum Genet* [Internet]. Cell Press; 2018 [cited 2019 May 17];103:858–73. Available from:

<https://www.sciencedirect.com/science/article/pii/S0002929718303677>

116. Dashnow H, Tan S, Das D, Eastel S, Oshlack A. Genotyping microsatellites in next-generation sequencing data. *BMC Bioinformatics* [Internet]. BioMed Central Ltd; 2015 [cited 2019 May 24];16:A5. Available from: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S2-A5>

117. Gymrek M. ScienceDirect A genomic view of short tandem repeats. *Curr Opin Genet Dev* [Internet]. Elsevier Ltd; 2017;44:9–16. Available from: <http://dx.doi.org/10.1016/j.gde.2017.01.012>

118. Bahlo M, Bennett MF, Degorski P, Tankard RM, Delatycki MB, Lockhart PJ. Recent advances in the detection of repeat expansions with

short-read next-generation sequencing. F1000Research [Internet]. Faculty of 1000 Ltd; 2018 [cited 2019 May 17];7:1–11. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/29946432>

119. Dashnow H, Bell KM, Stark Z, Tan TY, White SM, Oshlack A. Pooled-parent exome sequencing to prioritise de novo variants in genetic disease [Internet]. bioRxiv. 2019. Available from: <https://www.biorxiv.org/content/10.1101/601740v1>

120. Glessner JT, Bick AG, Ito K, Homsy J, Rodriguez-Murillo L, Fromer M, et al. Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. *Circ Res* [Internet]. The Center for Applied Genomics, Children’s Hospital of Philadelphia, Philadelphia, PA 19104, USA. Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA. Genetics, Harvard Medical School, Boston, MA 02115, USA. *Mi*; 2014;115:884–96. Available from: <http://dx.doi.org/10.1161/CIRCRESAHA.115.304458>

121. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, et al. De novo mutations in histone-modifying genes in congenital heart disease. *Nature* [Internet]. Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06510, USA.; 2013;498:220–3. Available from: <http://dx.doi.org/10.1038/nature12141>

122. Brown TL, Meloche TM. Exome sequencing a review of new strategies for rare genomic disease research. *Genomics* [Internet]. Department of Neurology, Columbia University, USA. Electronic address: tbrown@post.harvard.edu. Major General Hugh G. Robinson Center for Neuropsychiatric Studies.; 2016;108:109–14. Available from: <http://dx.doi.org/10.1016/j.ygeno.2016.06.003>

123. Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* (80-) [Internet]. Institute for Systems Biology, Seattle, WA 98103, USA.; 2010;328:636–9. Available from: <http://dx.doi.org/10.1126/science.1186802>
124. de Ligt J, Willemsen MH, van Bon BWM, Kleefstra T, Yntema HG, Kroes T, et al. Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. <http://dx.doi.org/101056/NEJMoa1206524> [Internet]. Massachusetts Medical Society; 2012 [cited 2018 Jul 9]; Available from: <https://www.nejm.org/doi/full/10.1056/nejmoa1206524>
125. O’Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* [Internet]. Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA.: Nature Publishing Group; 2011 [cited 2018 Jul 9];43:585–9. Available from: <http://www.nature.com/doifinder/10.1038/ng.835>
126. Rauch A, Wieczorek D, Graf E, Wieland T, Ende S, Schwarzmayr T, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* [Internet]. Institute of Medical Genetics, University of Zurich, Schwerzenbach-Zurich, Switzerland.; 2012;380:1674–82. Available from: [http://dx.doi.org/10.1016/S0140-6736\(12\)61480-9](http://dx.doi.org/10.1016/S0140-6736(12)61480-9)
127. Vissers LELM, de Ligt J, Gilissen C, Janssen I, Stehouwer M, de Vries P, et al. A de novo paradigm for mental retardation. *Nat Genet* [Internet]. Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences and Institute for Genetic and Metabolic Disorders, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands.; 2010;42:1109–12. Available from: <http://dx.doi.org/10.1038/ng.712>

128. Veltman JA, Brunner HG. De novo mutations in human genetic disease. *Nat Rev Genet* 2012 138 [Internet]. Nature Publishing Group; 2012;13:565. Available from: <https://www.nature.com/articles/nrg3241>
129. Wright CF, FitzPatrick DR, Firth H V. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet* [Internet]. Nature Publishing Group; 2018 [cited 2018 Jul 9];19:253. Available from: <https://www.nature.com/articles/nrg.2017.116>
130. Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nat Rev Genet* [Internet]. Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, 1210 Vienna, Austria. 1] Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, 1210 Vienna, Austria. [2] Vienna Graduate School of Population Genetics.; 2014;15:749–63. Available from: <http://dx.doi.org/10.1038/nrg3803>
131. Popp B, Ekici AB, Thiel CT, Hoyer J, Wiesener A, Kraus C, et al. Exome Pool-Seq in neurodevelopmental disorders. *Eur J Hum Genet* [Internet]. Institute of Human Genetics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91054, Erlangen, Germany. bernt.popp@uk-erlangen.de. Institute of Human Genetics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 91054, Erlangen, Germany. *Inst*; 2017;25:1364–76. Available from: <http://dx.doi.org/10.1038/s41431-017-0022-1>
132. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing [Internet]. *arXiv [q-bio.GN]*. 2012. Available from: <http://arxiv.org/abs/1207.3907>
133. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* [Internet]. Simons Foundation Autism Research Initiative, New York, NY 10010, USA.

gf@simonsfoundation.org; 2010;68:192–5. Available from:
<http://dx.doi.org/10.1016/j.neuron.2010.10.006>

134. Huang HW, Program NCS, Mullikin JC, Hansen NF. Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics* [Internet]. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. hhuang58@jhup.edu. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. mullikin@mail.nih.gov. National Human Genome Research; 2015;16:235. Available from: <http://dx.doi.org/10.1186/s12859-015-0624-y>

135. Jakaitiene A, Avino M, Guarracino MR. Beta-Binomial Model for the Detection of Rare Mutations in Pooled Next-Generation Sequencing Experiments. *J Comput Biol* [Internet]. 1 Bioinformatics and Biostatistics Center, Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University , Vilnius, Lithuania . 2 High Performance Computing and Networking Institute , National Research Council, Naples, Italy .; 2017;24:357–67. Available from: <http://dx.doi.org/10.1089/cmb.2016.0106>

136. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* [Internet]. Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. depristo@broadinstitute.org; 2011;43:491–8. Available from: <http://dx.doi.org/10.1038/ng.806>

137. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* [Internet]. Program in Medical and Population Genetics, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.: Cold Spring Harbor Lab;

2010;20:1297–303. Available from: <http://dx.doi.org/10.1101/gr.107524.110>

138. Institute DSP@ B. UnifiedGenotyper [Internet]. GATK | Tool Doc. Index. 2018. Available from:

https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_genotyper_UnifiedGenotyper.php

139. Analysis of genome evolution over time with GATK [Internet]. 2018. Available from:

<https://gatkforums.broadinstitute.org/dsde/discussion/8407/analysis-of-genome-evolution-over-time-with-gatk>

140. Institute DSP@ B. GATK | Doc #11097 | Can't use VQSR on non-model organism or small dataset [Internet]. 2018. Available from:

<https://software.broadinstitute.org/gatk/documentation/article?id=11097>

141. Francioli L. gnomAD v2.1 [Internet]. 2018. Available from:

<https://macarthurlab.org/2018/10/17/gnomad-v2-1/>

142. Sadedin SP, Pope B, Oshlack A. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics* [Internet]. 2012;28:1525–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22500002>

143. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* [Internet]. 2009;25:1754–60. Available from: <http://dx.doi.org/10.1093/bioinformatics/btp324>

144. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* [Internet]. 2009;25:2078–9. Available from:

<http://bioinformatics.oxfordjournals.org/content/25/16/2078.abstract>

145. Institute B. Picard Tools [Internet]. Broad Institute, GitHub Repos.

Available from: <http://broadinstitute.github.io/picard/>

146. Garrison E. freebayes [Internet]. 2018. Available from:
<https://github.com/ekg/freebayes>
147. Casbon J. PyVCF [Internet]. 2019. Available from:
<https://github.com/jamescasbon/PyVCF>
148. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. England; 2010;26:841–2.
149. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol* [Internet]. Department of Human Genetics, University of Utah, Salt Lake City, UT, 84105, USA. bpederse@gmail.com. USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT, 84105, USA. bpederse@gmail.com. Department of Biomedical Informatics, Univers; 2016;17:118. Available from: <http://dx.doi.org/10.1186/s13059-016-0973-5>
150. Dashnow H, Lek M, Phipson B, Halman A, Davis M, Lamont P, et al. STRetch: detecting and discovering pathogenic short tandem repeats expansions. *bioRxiv* [Internet]. 2017; Available from:
<http://biorxiv.org/content/early/2017/07/04/159228.abstract>
151. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. *Nat Genet* 2014 469 [Internet]. Nature Publishing Group; 2014 [cited 2018 Jul 9];46:944. Available from: <https://www.nature.com/articles/ng.3050>
152. Gymrek M, Willems T, Reich D, Erlich Y. Interpreting short tandem repeat variations in humans using mutational constraint. 2017 [cited 2019 May 23]; Available from: <https://www-nature-com.ezp.lib.unimelb.edu.au/articles/ng.3952.pdf>
153. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et

al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* [Internet]. Nature Publishing Group; 2014 [cited 2019 Jun 15];32:246–51. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24531798>

154. Gremme G, Steinbiss S, Kurtz S. GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Trans Comput Biol Bioinforma* [Internet]. 2013 [cited 2019 Jun 14];10:645–56. Available from: <http://ieeexplore.ieee.org/document/6529082/>

155. Sadedin SP, Oshlack A. Bazam: a rapid method for read extraction and realignment of high-throughput sequencing data. *Genome Biol* [Internet]. BioMed Central; 2019 [cited 2019 Jun 14];20:78. Available from: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1688-1>

156. Gogvadze E, Buzdin A. Retroelements and their impact on genome evolution and functioning. *Cell Mol Life Sci* [Internet]. SP Birkhäuser Verlag Basel; 2009 [cited 2019 Jun 14];66:3727–42. Available from: <http://link.springer.com/10.1007/s00018-009-0107-2>

157. Moyzis RK, Buckingham JM, Cram LS, Dani M, Deaven LL, Jones MD, et al. A highly conserved repetitive DNA sequence, (TTAGGG)_n, present at the telomeres of human chromosomes. *Proc Natl Acad Sci U S A* [Internet]. National Academy of Sciences; 1988 [cited 2019 Jun 14];85:6622–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3413114>

158. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;517:608–

11. Available from: <http://dx.doi.org/10.1038/nature13907>
159. Cruts M, Engelborghs S, van der Zee J, Van Broeckhoven C. C9orf72-Related Amyotrophic Lateral Sclerosis and Frontotemporal Dementia [Internet]. GeneReviews®. University of Washington, Seattle; 1993 [cited 2019 Jun 14]. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25577942>
160. Saul RA, Tarleton JC. FMR1-Related Disorders [Internet]. GeneReviews®. University of Washington, Seattle; 1993 [cited 2019 Jun 14]. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20301558>
161. La Spada A. Spinal and Bulbar Muscular Atrophy [Internet]. GeneReviews®. University of Washington, Seattle; 1993 [cited 2019 Jun 14]. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20301508>
162. Ayhan F, Ikeda Y, Dalton JC, Day JW, Ranum LP. Spinocerebellar Ataxia Type 8 [Internet]. GeneReviews®. University of Washington, Seattle; 1993 [cited 2019 Jun 14]. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20301445>
163. LaCroix AJ, Stabley D, Sahraoui R, Adam MP, Mehaffey M, Kernan K, et al. GGC Repeat Expansion and Exon 1 Methylation of XYLT1 Is a Common Pathogenic Variant in Baratela-Scott Syndrome. *Am J Hum Genet* [Internet]. Cell Press; 2019 [cited 2019 May 17];104:35–44. Available from: <https://www.sciencedirect.com/science/article/pii/S0002929718304087>
164. Cortese A, Simone R, Sullivan R, Vandrovcova J, Yau WY, Humphrey J, et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. 2019;51:649–58.

Chapter 6 Appendices

Appendix A Pooled-parent supplementary materials

This appendix contains supplementary materials for Chapter 2 Pooled-Parent Exome Sequencing to Prioritise *de Novo* Variants in Genetic Disease.

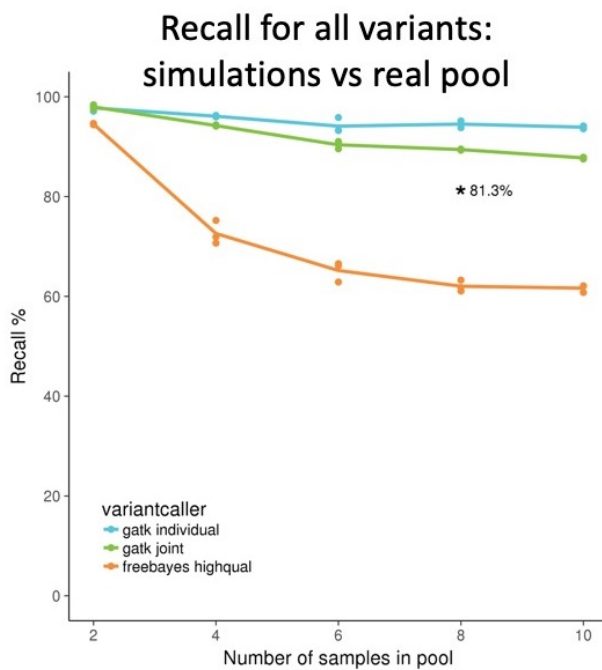


Figure 6.1: Recall for all constant depth simulations from Figure 1, with * to indicate recall for all variants in the real pools (percentage of variants found in all probands that were also detected in the pool).

Table 6.1: Mean recall rate (over three replicates) for all variants as a percentage in simulated pools of two, four, six, eight and ten individuals.

Recall % is the percentage of variants called in all individuals that make up the pool that are also called in the pool.

		Number of samples in pool				
sim_type	variantcaller	2	4	6	8	10
constant depth	gatk individual	97.8	96.1	94.1	94.5	93.9
	gatk joint	98	94.2	90.4	89.4	87.7
	freebayes highqual	94.5	72.6	65.2	62	61.6
additive depth	gatk individual	97.8	96.7	95.3	95.7	95.6
	gatk joint	98	95.3	91.7	90.3	88.8

Table 6.2: Mean recall rate (over three replicates) for singleton variants as a percentage in simulated pools of two, four, six, eight and ten individuals.

Recall % is the percentage of variants called in all individuals that make up the pool that are also called in the pool.

		Number of samples in pool				
sim_type	variantcaller	2	4	6	8	10
constant depth	gatk individual	95	86.2	71.5	69.4	63.9
	gatk joint	94.5	73.6	39.9	19.1	8.7
	freebayes highqual	89.4	57.9	38.2	27.2	18.8
additive depth	gatk individual	95	90.5	80	81.4	80.5
	gatk joint	94.5	79	46.1	18.8	6.7

Table 6.3: Mean variants called per pool (over three replicates) in simulated pools of two, four, six, eight and ten individuals.

		Number of samples in pool				
sim_type	variantcaller	2	4	6	8	10
constant depth	gatk individual	54699	74020	90934	102604	124668
	gatk joint	52627	68560	78770	85135	93724
	freebayes highqual	310388	498533	655845	790594	825890
additive depth	gatk individual	54699	74040	88815	96747	111552
	gatk joint	52627	68553	78762	85128	93708

Table 6.4: Mean false positive rate (over three replicates) as a percentage in simulated pools of two, four, six, eight and ten individuals.

False positive % is the percentage of variants called in the pool that are not called in any of the individuals that make up that pool.

		Number of samples in pool				
sim_type	variantcaller	2	4	6	8	10
constant depth	gatk individual	3.25	6.08	9.68	13.9	21.6
	gatk joint	0.0735	0.0545	0.0385	0.0349	0.0299
	freebayes highqual	14.8	11	8.14	6.53	5.33
additive depth	gatk individual	3.25	6.1	7.54	8.64	12.5
	gatk joint	0.0735	0.0643	0.066	0.0606	0.0598

Table 6.5: Mean variants called per individual (diploid) sample used in the simulations for each variant caller.

Variantcaller		
gatk individual	gatk joint	freebayes highqual
38934	39113	165067

Table 6.6: Ethnicity of probands, as identified by the patient's family.

Proband	Ethnicity
Proband 1	European (Greek)
Proband 2	Caucasian (Anglo)
Proband 3	Pacific Islander
Proband 4	European (Greek)

Table 6.7: SRA run IDs for the 111 exome samples used in pooling simulations.

All samples are parents from the Simons Simplex Collection.

SRR1272235	SRR1301418	SRR1301506	SRR1301624	SRR1301797	SRR1301865
SRR1272236	SRR1301419	SRR1301517	SRR1301627	SRR1301801	SRR1301872
SRR1272246	SRR1301422	SRR1301530	SRR1301631	SRR1301802	SRR1301873
SRR1272259	SRR1301423	SRR1301533	SRR1301632	SRR1301805	SRR1301876
SRR1272260	SRR1301430	SRR1301541	SRR1301653	SRR1301806	SRR1301877
SRR1272263	SRR1301431	SRR1301542	SRR1301654	SRR1301810	SRR1301887
SRR1272293	SRR1301438	SRR1301545	SRR1301665	SRR1301821	SRR1301888
SRR1272294	SRR1301439	SRR1301546	SRR1301698	SRR1301825	SRR1301894
SRR1301242	SRR1301457	SRR1301555	SRR1301699	SRR1301833	SRR1301898
SRR1301243	SRR1301458	SRR1301556	SRR1301718	SRR1301834	SRR1301899
SRR1301266	SRR1301461	SRR1301559	SRR1301726	SRR1301841	SRR1301902
SRR1301267	SRR1301462	SRR1301560	SRR1301737	SRR1301842	SRR1301903
SRR1301299	SRR1301468	SRR1301567	SRR1301749	SRR1301849	SRR1301918
SRR1301338	SRR1301469	SRR1301568	SRR1301763	SRR1301850	SRR1301919
SRR1301339	SRR1301487	SRR1301615	SRR1301764	SRR1301856	SRR1515931
SRR1301374	SRR1301488	SRR1301616	SRR1301771	SRR1301857	SRR1515932
SRR1301375	SRR1301491	SRR1301619	SRR1301772	SRR1301860	
SRR1301403	SRR1301492	SRR1301620	SRR1301775	SRR1301861	
SRR1301404	SRR1301505	SRR1301623	SRR1301776	SRR1301864	

Table 6.8: List of disease genes excluded from the real pooled parents analysis.

PSEN2	ALS2	C9orf72	SPG20	FA2H	VAPB
PARK7	CHMP2B	VCP	ATP7B	MAPT	PANK2
TARDBP	UCHL1	SETX	ANG	GRN	PRNP
GBA	SNCA	VPS13A	PSEN1	NPC1	
ATP13A2	SNCAIP	OPTN	NPC2	APOE	
PINK1	PARK2	TH	SPG11	FTL	
HTRA2	FIG4	LRRK2	FUS	NOTCH3	
UBQLN2	XK	PLA2G6	SOD1	APP	

Supplementary Materials

Appendix B STRetch paper supplementary materials

This appendix contains supplementary materials for Chapter 3 STRetch: detecting and discovering pathogenic short tandem repeat expansions.

Table S1: Comparison of allele sizes (number of repeat units) called by STRetch, LobSTR, HipSTR and ExpansionHunter in the 10 individuals with known STR alleles.

Sample	disease	gene	repeat unit	type	position	allele ref	allele PCR	allele STRetch	allele LobSTR	allele HipSTR	allele Expansion Hunter
Sample 1	SCA1	ATXN1	CAG	coding	chr6:16327865-16327955	30.3	51	50.5	31.3/31.3	36.3/39.3	31/70
Sample 2	Unaffected relative of sample 1	ATXN1	CAG	coding	chr6:16327865-16327955	30.3	29/32	32.2	30.3/35.3	30.3/35.3	30/35
Sample 3	SCA3	ATXN3	CAG	coding	chr14:92537355-92537396	14	73	65.4	28/28	no call	25/70
Sample 4	SCA6	CACNA1A	CAG	coding	chr19:13318673-13318712	13.3	22	no call- 0 STR reads in all samples	11.3/24.3	11.3/22.3	11/22
Sample 5	SBMA	AR	CAG	coding	chrX:66765159-66765261	33.3	41	43.3	no call	no call	52/55
Sample 6	SBMA	AR	CAG	coding	chrX:66765159-66765261	33.3	47	35.1 (0 STR reads)	no call	no call	57/63
Sample 7	FTDALS1	C9orf72	GGGGC C	intronic	chr9:27573482-27573544	10.8	>50	41.5	no call	no call	12/980
Sample 8	DM2	ZNF9	CCTG	intronic	chr3:128891419-128891502	20.8	>75	38.4	14.8/14.8	no call	14/48
Sample 9	DM1	DMPK	CAG	non-coding	chr19:46273462-46273524	20.7	>150	79.3	5.7/5.7	20.7/5.7	5/140
Sample 10	FRDA	FXN	GAA	intronic	chr9:71652203-71652205	6	~850	17.7*	No call	No call	83/117

Table S2: Raw data for sensitivity, specificity and FDR calculations. The ten true positive samples (Samples 1-10) and eight of the 97 control samples (Controls 1-8) were tested using STRetch at 22 known pathogenic loci (FRDA was omitted as it is not annotated in the reference genome). Cells contain the STRetch p-values (using reference controls) for all significant hits ($p < 0.05$).

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10	Control 1	Control 2	Control 3	Control 4	Control 5	Control 6	Control 7	Control 8
CCHS_PHOX2B																		
DM1_DMPK									1.13E-32									
DM2_ZNF9								4.05E-16										
DRPLA_ATN1																		
FRA12A_DIP2B																		
FRAXE_AFF2																		
FTDALS1_C9orf72	5.08E-08	3.88E-07				5.30E-08	1.20E-06	9.91E-03					3.57E-04	1.07E-05				
FXTAS_FMR1																		
HD_HIT									1.12E-04									
HDL2_JPH3																		
OPMD_PAPBN1																		
SBMA_AR					5.19E-05													
SCA1_ATXN1	2.46E-14																	
SCA2_ATXN2																		
SCA3/MJD_ATXN3			4.15E-09		5.75E-09				3.20E-08									
SCA6_CACNA1A																		
SCA7_ATXN7																		
SCA8_ATXN8/ATXN8OS											4.24E-24							
SCA10_ATXN10																		
SCA12_PPP2R2B																		
SCA17_TBP																		
SCA36_NOP56									1.73E-02									

Count Colour Description

7		True positive (Significant in STRetch and pathogenic expansion in PCR)
64		True negative (Not significant in STRetch, normal length on PCR)
2		False negative (Not significant in STRetch, pathogenic expansion by PCR)
7		Likely false positive (Significant in STRetch, no PCR confirmation)
3		Possible true positives (Significant in STRetch, non-pathogenic expansion confirmed by PCR and/or ExpansionHunter)
315		Likely true negative (Not significant in STRetch, no PCR confirmation)

Table S3: Pathogenic STR loci, positions in hg19. Also available as the bed file hg19.STR_disease_loci.bed on Figshare along with the other reference data at <https://figshare.com/s/1a39be9282c90c4860cd>. FRDA is not annotated as an STR in the reference genome and so was excluded from most analyses.

Chromosome	Start	End	Disease	Gene
chr3	63898361	63898392	SCA7	ATXN7
chr3	128891419	128891502	DM2	ZNF9
chr4	3076604	3076695	HD	HTT
chr4	41747993	41748039	CCHS	PHOX2B
chr5	146258291	146258322	SCA12	PPP2R2B
chr6	16327865	16327955	SCA1	ATXN1
chr6	170870995	170871105	SCA17	TBP
chr9	27573482	27573544	FTDALS1	C9orf72
chr9	71652203	71652205	FRDA	FXN
chr12	7045880	7045938	DRPLA	ATN1
chr12	50898785	50898805	FRA12A	DIP2B
chr12	112036754	112036823	SCA2	ATXN2
chr13	70713484	70713561	SCA8	ATXN8/ATXN8OS
chr14	23790681	23790701	OPMD	PAPBN1
chr14	92537355	92537397	SCA3/MJD	ATXN3
chr16	87637889	87637935	HDL2	JPH3
chr19	13318673	13318712	SCA6	CACNA1A
chr19	46273462	46273524	DM1	DMPK
chr20	2633379	2633421	SCA36	NOP56
chr22	46191235	46191304	SCA10	ATXN10
chrX	66765159	66765261	SBMA	AR
chrX	146993555	146993629	FXTAS	FMR1
chrX	147582125	147582273	FRAXE	AFF2

Table S4: PacBio validation of STRetch calls on CHM1 and CHM13 mixed Illumina data. This table lists all the loci that were called significant by STRetch compared to the 97 reference controls.

gene	repeat unit	type	position	allele ref	allele PacBio	sample	allele STRetch	rank (reference control)	p-val (reference control)
HIST2H2BA	A	upstream	chr1:120898564-120898595	31	NA	CHM1-CHM13-mix2	50.7	11	3.84E-02
BC038779	AG	upstream	chr2:62892328-62892385	28.5	72.5	CHM1-CHM13-mix2	46.0	3	7.78E-07
LPP	AGATAT	intronic	chr3:188331032-188331103	11.8	13.8	CHM1-CHM13-mix1	15.8	7	4.83E-02
OPA1	AAAAG	intronic	chr3:193314083-193314212	25.8	31.2	CHM1-CHM13-mix2	30.8	7	4.05E-04
-	AAAAG	intergenic	chr4:99901543-99901596	10.8	24.8	CHM1-CHM13-mix1	15.6	5	1.23E-03
-	AAAAG	intergenic	chr4:99901543-99901596	10.8	24.8	CHM1-CHM13-mix2	15.8	8	4.05E-04
-	A	intergenic	chr5:60901988-60902016	28	NA	CHM1-CHM13-mix1	52.0	4	1.23E-03
MRDS1	AAG	intronic	chr6:9795375-9795631	86	98	CHM1-CHM13-mix1	94.0	3	8.01E-04
MRDS1	AAG	intronic	chr6:9795375-9795631	86	98	CHM1-CHM13-mix2	94.3	5	2.45E-04
GNAI2	ACATC	intronic	chr7:2852270-2852331	12.2	28.2	CHM1-CHM13-mix2	31.9	9	1.12E-03
-	A	intergenic	chr9:76933570-76933599	29	262	CHM1-CHM13-mix1	78.1	2	2.56E-06
-	A	intergenic	chr9:76933570-76933599	29	262	CHM1-CHM13-mix2	69.3	4	5.73E-05
ABCA1	ACCCCC	intronic	chr9:107624982-107625036	9	18	CHM1-CHM13-mix2	14.0	6	4.05E-04
GABRG3	AGAT	intronic	chr15:27390742-27390811	17.8	23.55	CHM1-CHM13-mix1	23.8	6	3.50E-03
CHD3	CCG	intronic	chr17:7788625-7788663	12.7	30.7	CHM1-CHM13-mix1	24.9	1	1.88E-06
CHD3	CCG	intronic	chr17:7788625-7788663	12.7	30.7	CHM1-CHM13-mix2	27.9	2	3.62E-09
RNA5-8S5/ LOC100507412	ACCCCC	intronic	chrUn_g1000220:120825-120864	7.8	NA	CHM1-CHM13-mix2	23.2	1	1.72E-16
RNA5-8S5/ LOC100507412	AGGC	intronic	chrUn_g1000220:121453-121502	12.2	NA	CHM1-CHM13-mix2	17.1	10	1.48E-02

Table S5: Primers used for amplification and sequencing of the AACT repeat in an intron of the MTHFD2 gene (chr2:74430970-74431055).

	Primer name	Sequence (5'-3')	Position (chr2, hg19)	Length	%GC	Tm (°C)
PCR	MTHFD2 STR JC1 - FWD	TGGTGGGTGCCTGTATTC TCAG	+/ 74430660	22	54.5	58.1
	MTHFD2 STR JC2 - REV	TGCTTGAGGTCAGGAGT TCCAG	-/ 74431258	22	54.5	58
Sequencing	MTHFD2 STR JC1 seq - FWD	AAGAGGAGATTACTTCA TTGGTC	+/ 74430873	23	39.1	51.8
	MTHFD2 STR JC2 seq - REV	CATGGCAAACCCCGTC TCTG	-/ 74431223	21	57.1	58.1

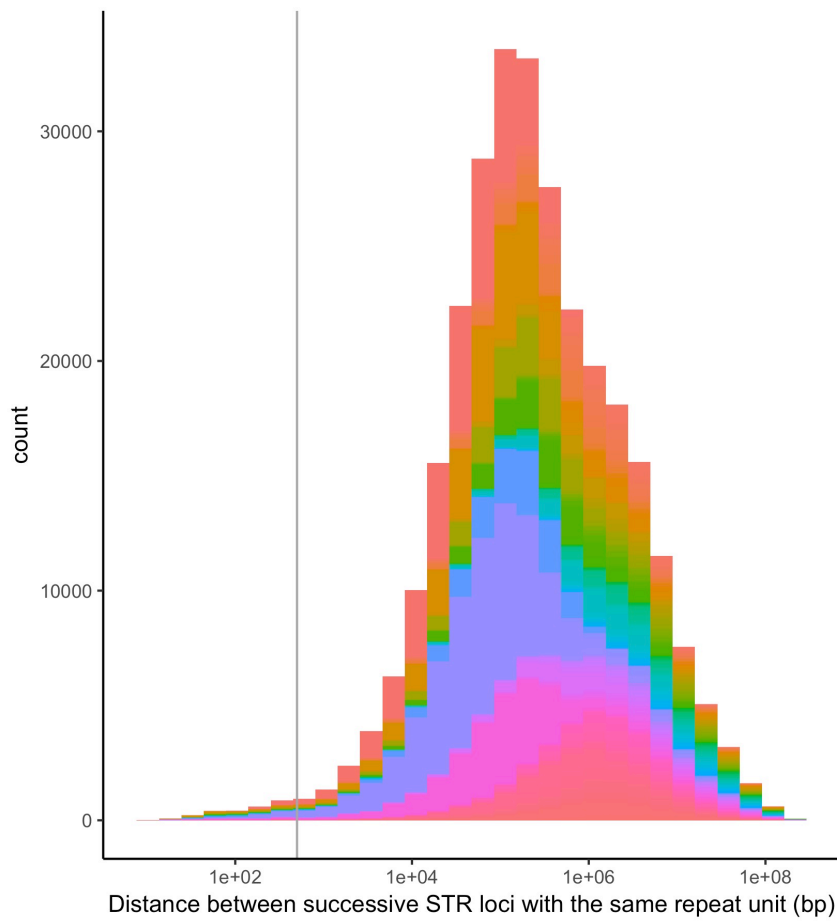


Fig. S1: The distribution of distances between STR loci with the same repeat unit in hg19. 0.93% of STR loci are within 500 bp of another STR locus with the same repeat unit (loci to the left of the solid vertical line at 500 bp). Colours indicate the 501 different STR repeat units.

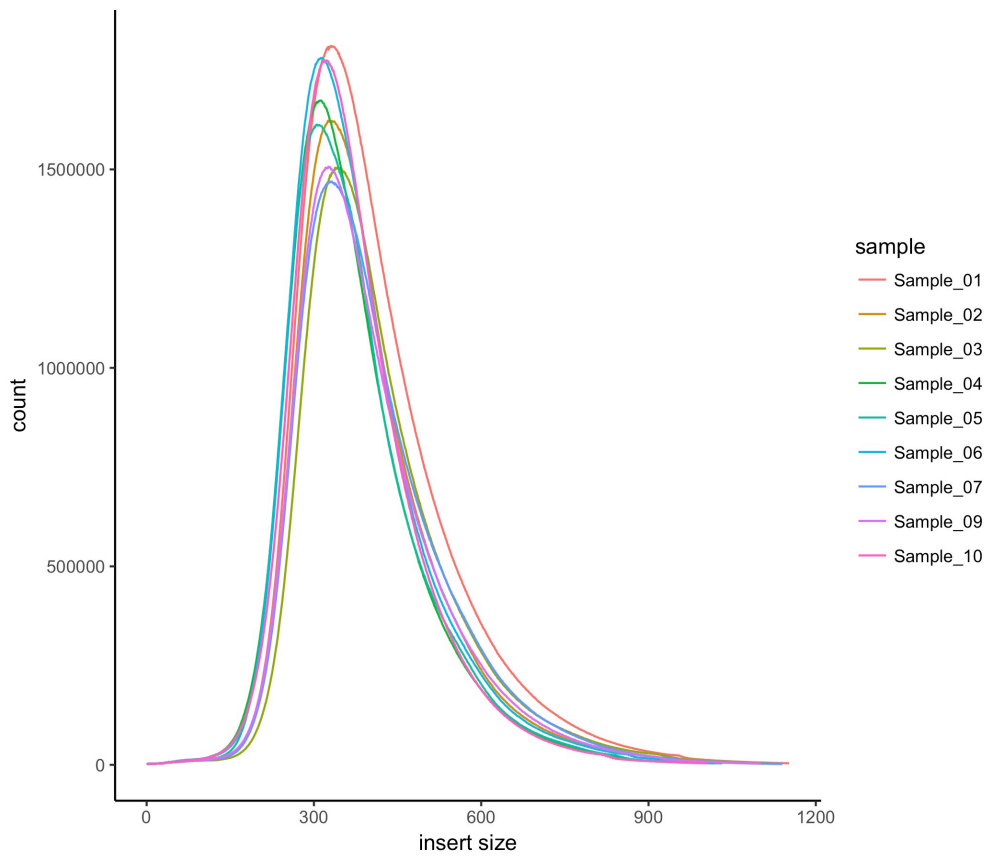


Fig. S2: Insert sizes of the ten true positive samples. The mean insert size ranged from 372 to 415 bp (standard deviation range: 130-150 bp).

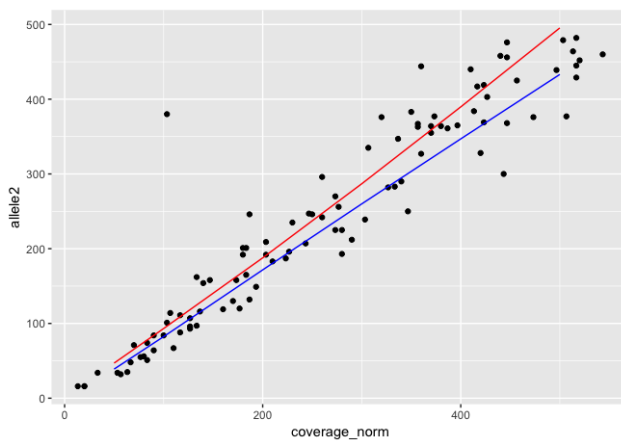


Fig. S3: The plot shows the simulated data (black points) with the upper (red) and lower (blue) bounds for the linear fit indicated. A plot of the number of reads mapping to the AGC decoy chromosome against the number of AGC repeat units inserted into the ATXN8 locus shows a clear linear relationship between these two variables.

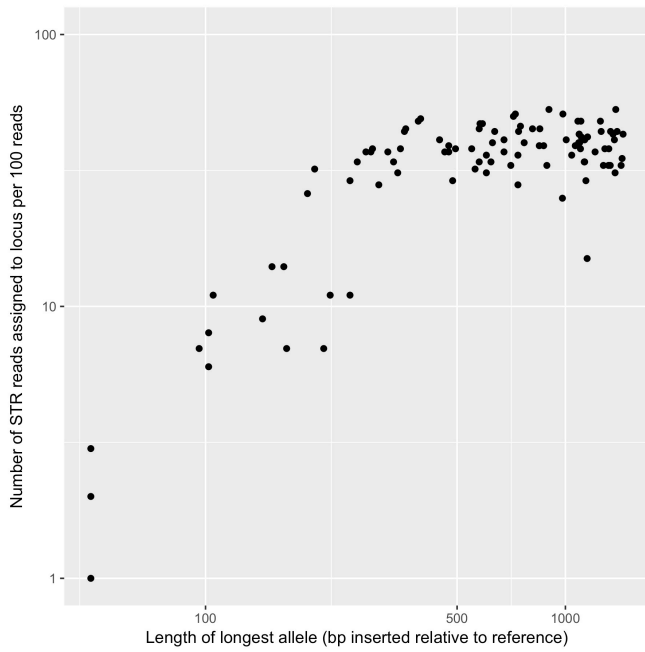


Fig. S4: The number of STR reads assigned to a locus does not increase beyond the insert size in simulated data, so STRetch will tend to underestimate alleles greater than the insert size.

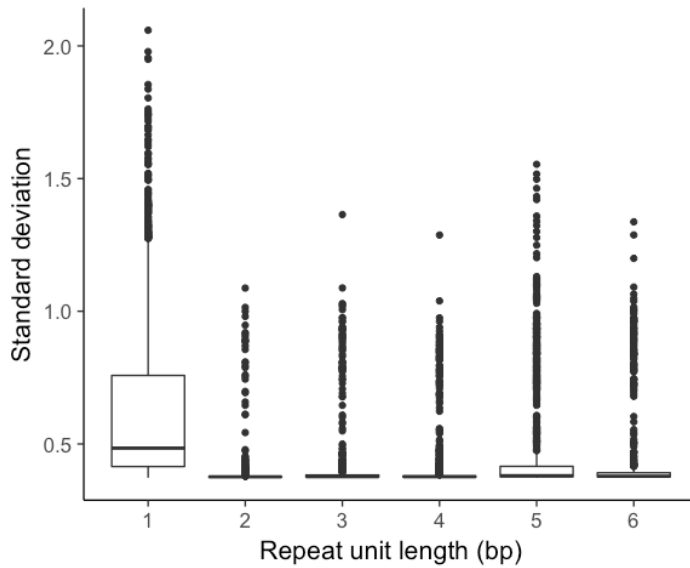


Fig. S5: Robust standard deviation of STR reads assigned to each locus across 97 WGS samples. Homopolymer (1 bp repeat unit) loci are the most variable between individuals.

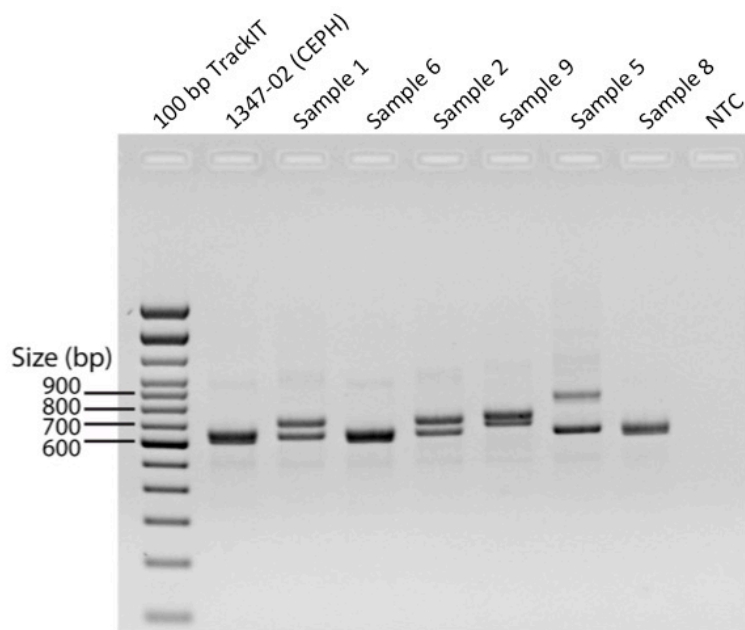


Fig. S6: PCR amplification of the AACT repeat in an intron of the *MTHFD2* gene (chr2:74430970-74431055). Sample 5 is predicted by STRetch to have a significant expansion while Samples 1, 2, 6 and 8 are not. Sample 1347-02 (CEPH control DNA) has an unknown genotype at this locus. 2% agarose gel, see primer sequences in Supplementary Methods.

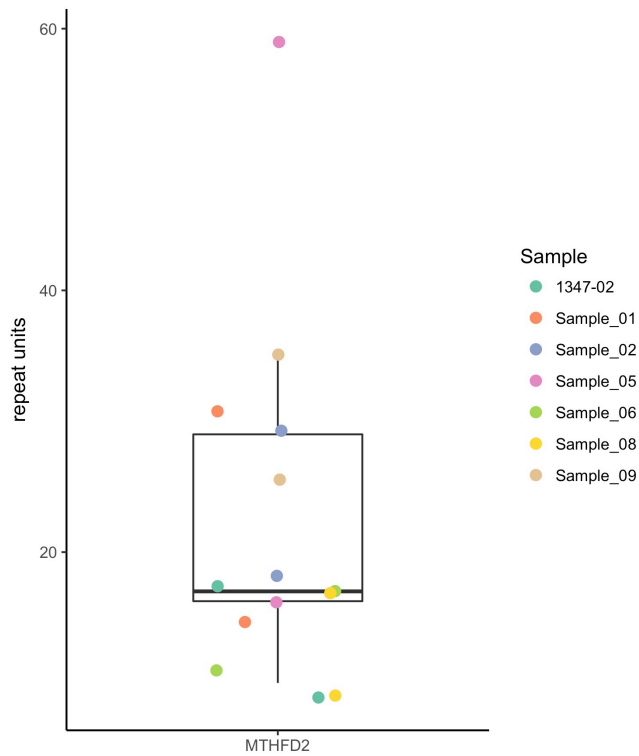


Fig. S7: Repeat sizes determined from Sanger sequencing of the AACT repeat in an intron of the *MTHFD2* gene (chr2:74430970-74431055). Sample 5 is predicted by STRetch to have a significant expansion while Samples 1, 2, 6 and 8 are not. The 1347-02 control DNA sample has an unknown genotype at this locus. See primer sequences in Supplementary Methods.

Supplementary Methods

Runtime

In single-threaded mode STRetch takes approximately 3 hours, 50 minutes (of which 3 hours, 40 minutes is re-alignment) to perform analysis of a 40X PCR-free whole genome using an aligned cram file as input. This can be reduced to approximately 1 hour, 50 minutes by running STRetch with 12X concurrency. Samples can be run in parallel. STRetch is built on Bpipe, which interfaces with most common compute cluster environments.

ExpansionHunter custom repeat specifications (hg38):

```
{
  "RepeatId": "ATXN8",
  "RepeatUnit": "CTG",
  "CommonUnit": "true",
  "TargetRegion": "chr13:70139384-70139428"
}
{
  "RepeatId": "MTHFD2",
  "RepeatUnit": "TAGTT",
  "CommonUnit": "true",
  "TargetRegion": "chr2:74203844-74203928"
}
{
  "RepeatId": "NOP56",
  "RepeatUnit": "GGGCCT",
  "CommonUnit": "true",
  "TargetRegion": "chr20:2652733-2652775"
}
{
  "CommonUnit": "true",
  "RepeatId": "ZNF9",
  "RepeatUnit": "CAGG",
  "TargetRegion": "chr3:129172577-129172656"
}
```

Appendix C STR variation supplementary materials

This appendix contains supplementary materials for Chapter 4 Frequency and variability of STR expansions.

Table 6.9: The proportion of the hg19 genome covered by various genomic features based on the Gencode v19 annotation.

Only canonical chromosomes are considered (chr 1-22, X, Y). Note that these values don't sum to 100% because some features overlap (e.g. different transcripts).

genomic feature	% of the genome
CDS	1.15
exon (inc. CDS)	3.94
intron	48.50
intergenic	49.38

Table 6.10: The top ten STR decoy chromosomes ranked by the number of reads aligned to them across all the 303 least variable samples.

Repeat unit	Mean decoy counts per sample	Mean decoy counts / sequencing depth	Percent of decoy reads
C	92912	2662	79.4
AACCCT	8604	245	7.36
A	4948	141	4.23
AATGG	2706	77.9	2.31
ACTCC	1802	51.6	1.54
ACC	628	18	0.537
AT	553	16.2	0.473
ACACT	519	14.9	0.444
AGGG	341	9.8	0.291
AG	293	8.4	0.251

Table 6.11: Summary of normal, intermediate and pathogenic allele size ranges for known pathogenic loci for which expansions were identified in the unaffected samples.

***intermediate = premutation/incomplete penetrance/uncertain pathogenicity**

disease	gene	repeat unit	inheritance	type	normal	intermediate *	pathogenic
DM1	DMPK	CAG	AD	coding	5–34	35-49	50-2000
DM2	ZNF9	CCTG	AD	intronic	11–26	27-74	75-11,000
FRAXE	AFF2/FMR2	CGG	XR	5'UTR	6-25		200-2000
FTDALS1	C9orf72	GGGGCC	AD	intronic	3–25	20-60	>60
FXS/FXTAS/POF1	FMR1	CGG	XD	5'UTR	5–44	45-200	200-2000
HD	HTT	CAG	AD	coding	6–26	27-39	40–250
SBMA	AR	CAG	XR	coding	9–34	36-37	38–68
SCA1	ATXN1	CAG	AD	coding	6–35	36-38	39–88
SCA3/MJD	ATXN3	CAG	AD	coding	12–44	45-59	60-87
SCA36	NOP56	GGCCTG	AD	intronic	3 to 14	15-649	650-2500
SCA8	ATXN8/ ATXN8OS	CAG	AD	untranslated exon	15–50	50-70	71-1300

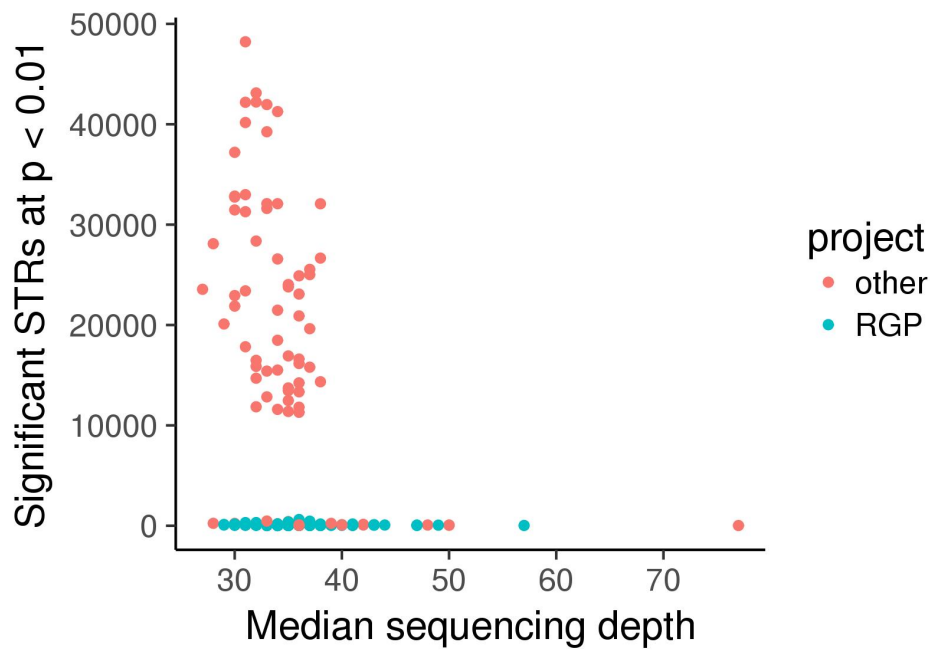


Figure 6.2: Number of significant STR loci at $p < 0.01$ called by STRetch vs. median sequencing depth across the genome for each sample shows not increased rate of significant calls in more deeply sequenced samples.

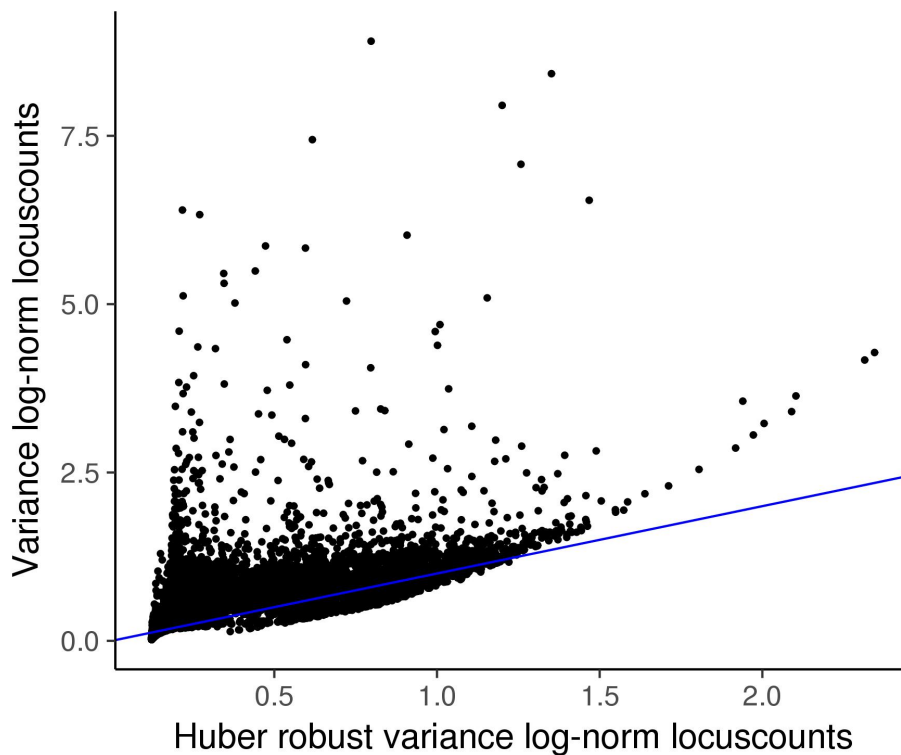


Figure 6.3: Variance vs. Huber's robust variance estimate of log normalised locuscounts for all samples.



Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Dashnow, Harriet

Title:

Bioinformatics methods and approaches to discover disease variants from DNA sequencing data

Date:

2019

Persistent Link:

<http://hdl.handle.net/11343/230817>

File Description:

Final thesis file

Terms and Conditions:

Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.