Integrating Diverse Technologies for Genomic Variant Discovery

by

Alexandra M. Weber

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2022

Doctoral Committee:

Associate Professor Ryan Mills, Chair
Associate Professor Jeffery Kidd
Professor Jun Li
Professor Kayvan Najarian
Associate Professor Maureen Sartor

Alexandra M. Weber

aleweb@umich.edu

ORCID iD: 0000-0001-9137-6799

Acknowledgements

I owe thanks to many incredible individuals who have made this work possible. First and foremost, I would like to thank my advisor, Dr. Ryan Mills whose mentorship has not only been instrumental to my development as a scientist but also as an individual. I could not have completed this dissertation without his guidance not only in scientific endeavors but also his support and advice in all aspects of pursing a PhD. He has shown time and time again to be an advisor who truly cares about the well-being and development of his trainees. I am grateful to have worked in a lab environment that was not only collaborative and supportive but also fun.

I would also like to thank my past research mentors Dr. Daniel Schaid, Dr. Caitlin Pepperell, and Dr. Tatum Mortimer who introduced me to bioinformatics and were instrumental in my decision to pursue graduate school.

Additionally, I am grateful for the advice and insight provided by my committee members, Dr. Jeffery Kidd, Dr Jun Li, Dr. Kayvan Narjarian, and Dr. Maureen Sartor. I would like to thank my collaborators from the Todd, Boyle, and Oleksyk labs who played active roles in the development and progression of my projects. Thank you to the members of the Department of Computational Medicine and Bioinformatics who have provided support throughout my time in the program including Julia Eussen, Helen Severino and our graduate program chairs, Dr. Margit Burmeister and Dr. Maureen Sartor.

I am so grateful for the Mills lab who have not only been wonderful work colleagues but also great friends. Thank you to Xuefang Zhao, Tony Chung, Yifan Wang, Arthur Zhou, Marcus

Sherman, Chen Sun, Steve Ho and Wenjin Gu for all for your support, mentorship, chats, lab pranks, and delicious food. You have made the time fly by.

My time at Michigan has been made infinitely better by the friendships I have made here. Thank you to Ricardo D'Oliveira Albanus, Esra Ascigil, Allie Bouza, Christopher Castro, Ben Chandler, Joey Cicchese, Brad Crone, Marlena Duda, Danny Geiszler, Ben Hillebrand, Kevin Hu, Alex Kalinin, Zena Lapp, Cathy Smith, Matt Stewart, and Brooke Wolford. Our cookouts, tubing, trivia nights, intramural sports, and general shenanigans were an integral part of my time at Michigan. Also, thank you to Nick Joslyn and Sam Lima for being the ultimate pandemic squad.

I want to thank my family including my sisters, Erica and Justine, and my nieces who's facetime calls have been the perfect distraction after long workdays. And to my parents who's love and encouragement I have felt in everything I do. Thank you for all your support through this long educational journey. I would like to especially thank my mom for being a strong role model as a woman in computer science which allowed me to never question my place in this field.

And finally, I would like to thank Louis whose endless love, support, and belief in me means the world. Thank you for everything you do.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

Accurate detection of variation in the human genome is important for understanding diversity in the human species and for identifying the cause of genetic diseases. The technology for interrogating the genome has vastly improved since the sequencing of the first human genome, improving our ability to accurately detect and characterize more complex variation. However, there are still biases and limitations for all currently available technologies that we must work within. An integrative approach using multiple genotyping or sequencing platforms is a practical strategy that can work within these limitations while improving variation detection beyond what can be achieved with a single technology.

In this thesis, I apply an integrative approach to variant detection for different but related scenarios. First, I use Illumina short read sequencing and SNP microarrays to validate variant calls from BGI nanoball short read sequencing to provide a resource of variants present in individuals of Ukrainian decent, a previously underrepresented group in publicly available genome sequencing databases. Second, I study the ability to detect tandem repeat variation genome wide using both short and long read sequencing datasets through the comparison of multiple tandem repeat characterization methods. Lastly, I combine whole genome short read sequencing datasets to understand the relationship between SNP haplotypes and tandem repeat lengths to estimate tandem repeat lengths in individuals with ALS genotyped using SNP microarrays. Taken altogether, these examples represent case studies that demonstrate the utility of an integrative approach to genomic variant detection, analysis, and characterization.

Chapter 1 - Introduction and Background

1.1 MOTIVATION

Humans share 99.9% of DNA in the genome but it is the variation in the remaining 0.1% percent that makes us unique (1). Studying these variations can help us understand more about the evolution of modern humans, investigate differences across populations, and uncover the causes of disease which can lead to new therapeutic strategies and treatments (2).

The last 50 years has been spent developing technology to understand the sequence of the human genome and to characterize the genomic variation present in humans (3). In 2003, the use of shot-gun sequencing allowed the human genome to be sequenced for the first time (4). Genotyping microarrays aided in the discovery of many disease-associated variants because it provided a way to quickly detect known variations in thousands of samples which is required for genome wide association studies. Next generation sequencing (NGS) provided parallelization of the sequencing process which permitted the creation of high resolution genetic data for thousands of samples resulting in a better understanding of the diversity present in human species across variant classes. Finally, long read sequencing has given researchers the opportunity to interrogate regions of the genome which have been inaccessible with previous technologies. It stands to reason that as sequencing technologies continue to develop, they will similarly lead to new discoveries of the diversity of genome variation and what that means for our health.

Despite incredible strides, limitations of current sequencing technologies remain. For example, current methods struggle to capture and accurately characterize complex variations and highly repetitive regions of the genome (5,6). Additionally, with each advance in sequencing technology, there is an ever-present need for new bioinformatic approaches to process datasets to gain insights from these newer technologies (7).

Until we can achieve end-to-end genome sequencing at large scale without errors in a cost-efficient manner, we are only able to assess genome variations accessible with technology available at the present time. That said, the integration of multiple orthogonal techniques for variant detection can improve our ability to detect variation from the available technology. In this thesis, I integrate data from multiple sequencing platforms and employ various bioinformatics approaches to detect variation across the human genome.

1.2 BUILDING THE REFERENCE GENOME

The Structure of the Human Genome

The human genome encompasses approximately 3 billion base pairs and is comprised of 46 chromosomes – two copies of 22 autosomes and two sex chromosomes, either XX or XY. It is estimated that only one percent of the genome is made up of protein coding genes while the remaining DNA was originally thought of as 'junk DNA', but can actually serve several functions such as regulating the gene expression or organizing genome structure (8). The human reference genome serves as a linear representation of the DNA sequence that makes up the human genome.

Human Genome Project and the Reference Genome

The first draft of the human reference genome was created during the Human Genome Project – an international collaboration involving both the public and private sectors that was created in 1990 (4). The Human Genome Project used Sanger sequencing (discussed below) to determine and map the nucleotide sequences that make up the human genome. This pivotal work for genomic sciences concluded in 2003 with approximately 85% of the genome resolved (4). The sequence that makes up the reference genome is neither the complete genome of a single individual, nor is it representative of all the variation present in the human species. Approximately 93 percent of the current reference genome sequence is derived from 11 individuals with 70% deriving from a single male of African-European decent. The remaining 7% of sequence comes from over 50 different libraries (9). One of the main functions of the human reference genome is to act as an index of the genome for locating genes or other genomic features in genetic data generated from new samples (10,11). Additionally, as its name may suggest, the reference genome can act as a standard "reference" for which to classify variation. To identify variation in a particular genome, it must be in relation to how it is different from another genome - the reference genome provides a common sequence which can be used by scientists across different studies in order to characterize variation.

1.3 MOTIVATIONS FOR DETECTING VARIATION IN THE GENOME

After the completion of the first draft of the human genome, the next challenge was to understand the variation present across individuals. Classifying variation across diverse individuals can help researchers better understand our evolutionary history as well as identify variations responsible for genetic disease (12,13). While many variations in genome are benign, others can cause disease. Identifying genetic variations that cause or contribute to disease phenotypes can be the first step

in identifying the disease mechanisms and can lead to new therapeutic targets (14). A key component when developing personalized medicine strategies is to directly incorporate an individual's genetic make-up into their medical care. However, to ensure these efforts are accurate and available to everyone, it is important to understand the genetic diversity in our species (15).

Genotyping and Genome sequencing initiatives

Since the completion of the Human Genome Project, several large scale, sequencing initiatives have greatly contributed to our knowledge of the variation present across the human species. Each of these projects aimed to create a resource for the community to facilitate research into how variation affects different phenotypes, including disease phenotypes.

The HapMap Project

The HapMap project ran from 2002 until the final publication of results in 2010 (16). Due to the high cost of whole genome sequencing at the time, it was difficult to obtain whole genome sequencing for many diverse individuals. Instead of looking at every genomic position, a more cost-effective alternative was to use genotyping arrays (discussed in depth below) to look only at single nucleotide polymorphisms (SNPs – discussed in depth below) that were known to differ between individuals at an appreciable frequency, as humans share 99.9% of DNA sequence. SNPs that are nearby each other are generally inherited together, resulting in common segments of DNA referred to as haplotypes. These haplotypes can be represented by smaller numbers of 'tag' SNPs that could be directly interrogated using microarrays (17,18). The project went through three phases, with each phase subsequently increasing either the number of characterized SNP variants or increasing the number of samples included in the dataset. The project finished with 1.6 million

SNPs genotyped in 1,184 samples from 11 different global ancestry groups (19). This dataset has been used extensively in studies to determine regions of the genome that are associated with common disease (16,20–22)

1000 Genomes project

As the cost of genome sequencing declined, the 1000 Genomes Project (1KGP) was formed to sequence thousands of individuals to gain a more complete understanding of the genome variation present in different human populations. The 1KGP was another international collaboration and involved sequencing over 2,500 reportedly healthy individuals from 26 diverse populations around the world. The purpose of this initiative was to generate a publicly available database of genomic data and variation to facilitate research on the role genetic variation plays on health and disease (23). This project was created to build further on what was learned through the HapMap project and several of the samples overlap between the two projects. However, genome sequencing allowed for the characterization of variation beyond just SNPs. The 1KGP identified variations such as small insertions and deletions (indels) and structural variations (discussed in depth below). This project also consisted of multiple phases as the sequencing technology advanced and the cost continued to decline (24) The phase 3 dataset includes 2,504 samples with whole genome sequencing data, and scientists recently re-sequenced these samples using high coverage, whole genome Illumina sequencing for all samples (25). The findings presented by the 1KGP have provided avenues to study the amount of variation present in presumably healthy individuals. Additionally, studies aimed at elucidating novel pathogenic variations can identify variants present in disease samples that are not present in 1KGP samples. This allows the prioritization of variations more likely to result in a disease phenotype.

TopMed and other Disease Genomic Databases

While previous large initiatives were aimed at sequencing and genotyping presumably healthy populations to create reference sets of background genomic variation, other projects have focused on sequencing disease samples. One such initiative is the Trans-omics for precision Medicine (TopMed). The TopMed database not only contains genomic data, but also includes additional omics datasets such as transcriptomics and epigenomics for every sample. Additionally, the database provides clinical and environmental data in order to facilitate precision medicine research. TopMed focuses on diseases related to the heart, lung, blood, and sleep (26). Many other initiatives have been created for specific diseases or specific classes of disease such as Answer ALS (27), Alzheimer's Disease Sequencing Project (28), and the T2D-Genes consortium (29) to name a few.

Genome Aggregation Database

As more sequencing initiatives were created, the Genome Aggregation Database (gnomAD) consortium was created to systematically aggregate and reanalyze these available datasets in a uniform manner. By providing uniform analysis, gnomAD created an even larger overview of the variation present in the human species, providing context for variation in both (presumably) healthy individuals as well those diagnosed with disease. Key outcomes of this initiative include catalogs of variation, for both small variants and larger structural variants, that have been generated as a public resource for studying genetic diversity and disease (30).

Variation Databases

As genetic variation is continually discovered in the sequencing datasets generated in these initiatives, several databases have been generated to provide a central location to store these variations. dbSNP (31) and dbVar (32) are databases containing small and large variations that have been discovered across studies. ClinVar is a database specifically for the relationship between genome variation and phenotypes (33).

These are just a few of the main initiatives that have been undertaken to create publicly available datasets to aid in genomic research. While these provide incredible resources for the genomics and medical research committees, there are still areas for additional work. Through these initiatives, the importance of using population matched controls for detecting pathogenic variations in disease samples became clear to prevent false positive results stemming from identity by descent (34). While there are many populations represented in these datasets, there are still underrepresented groups and it is important that we continue expanding the diversity of samples included in these resources (35). Further, as sequencing technology has advanced, new and more accurate variations have been detected and integrated into these datasets. Whole genome sequencing on a large scale has allowed researchers to systematically examine larger, structural variations in individual samples. With the introduction and optimization of long read sequencing, we should be able to incorporate information about various complex variations, including repetitive regions. In particular, long read sequencing has the potential to interrogate and characterize variation in the

hard-to-reach areas of the genome, including regions such as centromeres which have been challenging to sequence with previous technologies (36). Below, we cover the various classes of genetic variation that are present across the genome.

1.4 CLASSES OF GENETIC VARIATION

SNPs

Single nucleotide polymorphisms (SNPs) are the variation of a single DNA base pair. They are the most common type of variation in the human genome and occur about one in every 1,000 base pairs (37). The mutation rate of SNPs is estimated to be ($1.0\text{-}1.25*10^{-8}$) per site per generation (38–40). While many SNPs do not create pathogenic variations capable of causing disease, others can affect the function of cells and thereby cause disease. One example of a pathogenic SNP is missense mutations (41). Missense mutations result from SNPs in the coding region of a gene that generates a change in the amino acid that gene normally codes for as part of the translation process. This missense mutation will produce a different, potentially harmful protein. A single base pair substitution in a coding sequence for hemoglobin sufficiently alters the protein generated from the sequence resulting in sickle-cell disease (42,43).

Indels

The insertion or deletion of up to 50 base pairs in a genome (relative to a reference) are collectively referred to as indels (44), as it is impossible to determine whether the underlying variant was deleted or inserted without an ancestral outgroup for comparison. While many of these variants can be harmless, insertions or deletions in coding sequence can be damaging, particularly if the insertion or deletion of base pairs is not a multiple of three, thereby resulting in a frame shift

mutation (45). Because the amino acids used to create protein sequences are coded using three consecutive base pairs, frame shift mutations can interrupt the frame in which the DNA sequence is read, creating a significantly different protein product than intended and often truncating the protein early due to the creating of a nonsense stop codon. Furthermore, even non-frame shifted mutations can have dire consequences. For example, an indel resulting in the deletion of a single codon within a coding sequence can result in the genetic disease cystic fibrosis (46).

Structural Variation

Variations affecting greater than 50 base pairs are considered structural variation by convention. Compared to SNPs and indels, there are fewer structural variations in a given genome, but these types of variation collectively affect a larger percentage of the genome than the smaller variations (24).

Copy Number Variations

Large variations which result in a change in the amount of DNA are considered copy number variation (CNVs). This includes the insertion, deletion, or duplication of DNA sequences in the genome. Like SNPs, some variants can be harmless, while others can affect the function of an otherwise healthy cell. In the worst cases, CNVs can cause the deletion or duplication of entire genes, which can be particularly harmful. For example, Prader-Willi and Angelman syndrome are caused by deletions on chromosome 15 (47).

Inversions

Inversions are a class of structural variation that does not result in a change in the amount of DNA present and are considered balanced variations. Inversions involve a region of the genome occurring in a different orientation than expected. They can cause issues if the breakpoints of the inversion occur within important regions of the genome. A key inversion in the factor VIII gene is a common cause of hemophilia A, which can lead to spontaneous bleeding and an inability for blood to clot following injury in those who are afflicted (48).

Translocations

Translocations occur when a segment of a chromosome is transferred to a new location in the genome. These variations can either be unbalanced, resulting in loss or gain of genetic material, or balanced, resulting in no change in genetic material. Balanced, or reciprocal translocations, occur when two segments for a chromosome switch locations in the genome. Both balanced and unbalanced translocations can cause disease phenotypes. When translocation breakpoints occur within genes, this can alter the function of the gene. Translocations are quite prevalent in cancer (49). In particular, oncogenic fusion genes formed by translocations have been identified in both hematological and solid tumors (50–53).

Mobile Element Insertions

Transposable elements are segments of DNA which are capable of replicating and creating copies of itself at new regions of the genome. Transposable elements that have recently been mobile in the genome are classified as mobile element insertions (MEIs). These include L1s, a sub-class of long interspersed nuclear elements (LINE), *Alu* elements, a subclass of short, interspersed nuclear elements (SINEs), and SINE-VNTR-*Alu* (SVA) elements. Together, sequences derived from these

elements make up over 35% of the genome (6). NUMTs, nuclear mitochondrial DNA, is the insertion of mitochondrial DNA into the nuclear genome (54). While there are sequences from these elements present in the genome, it is the novel insertions that are of interest. Like other genomic variation, the insertion of MEIs into important regions of the genome can cause problems. The exonic insertion of an AluY in *FIX* gene has been associated with Hemophilia B (55).

Tandem Repeat Expansion

Tandem repeats are a specific class of copy number variation that includes a repetitive motif of bases that are repeated in succession in the genome. These are often divided into two classes based on the size of the repeat motif. Short tandem repeats (STRs) have repeat motifs up to 6 base pairs in length and variable number tandem repeats (VNTRs) have repeat motifs larger than 6 base pairs. The primary mechanism of mutation for tandem repeats is strand slippage, which involves the displacement of DNA strands during DNA replication (56). Expansions of these repetitive regions are thought to occur in a step-wise fashion, i.e. the expansions increase in size in subsequent generations or potentially in subsequent replications of the cell (57). Tandem repeats have several orders of magnitude higher mutation rates (10^{-6} to 10^{-2}) than point mutations (58) and are prevalent in neurodegenerative disease. Trinucleotide repeat expansions are the known cause of Huntington's Disease and Fragile X syndrome (59,60).

1.5 TECHNOLOGY FOR DETECTING GENOMIC VARIATION

DNA Microarrays

DNA Microarrays are a genotyping technology and offer a method for identifying genetic variation. DNA microarrays are a less expensive alternative to DNA sequencing but are more limited in what they can provide. Microarrays are created by attaching single stranded DNA probes onto a small chip. These probes contain the sequence for different variations of regions in the genome. Single stranded DNA from a sample is washed over the microarray chip to observe which probe the sample DNA hybridizes with, determining the sequence present in the sample (61). While microarrays can be applied for a variety of scientific uses, this thesis will focus on the use of microarrays for SNP detection. Microarrays are commonly employed for genotyping common SNPs in large populations. They are popular due to their high accuracy, low cost, and quick turn-around time; however, microarrays are unable to detect novel SNP variation (or rare SNP variants) and are not straightforward in detecting larger structural variations. As discussed below, other whole genome sequencing methods are required for these types of studies. Additionally, due to their popularity and their length of existence, there is an abundance of previously generated datasets available. DNA microarrays datasets are analyzed as part of the analysis performed in Chapter 4.

Sanger Sequencing

The first human genome was sequenced using the Sanger chain termination method (4). In Sanger sequencing, template strands of DNA undergo in vitro DNA replication but the random incorporation of specific fluorescently labeled chain-terminating nucleotides halt replication resulting in DNA strands of different lengths. These DNA strands undergo size separation, and the sequence of base pairs can be identified by observing the order of the different colors indicating the final base pair incorporated in the strands of increasing size (62). When considering read

length, there is an upper bound of around 800 base pairs due to size separation becoming challenging for molecules beyond this length. Despite improvements to Sanger sequencing since its inception, it remains a time-consuming method and has been superseded by Next Generation Sequencing methods for most routine sequencing. However, the high accuracy and the length of the reads generated make Sanger sequencing amenable for validation of variation calls made using other sequencing methods with shorter or lower accuracy reads. In Chapter 3, we use Sanger sequencing datasets as validation or 'truth sets' to compare variant calls generated from multiple sequencing technologies.

Next Generation Sequencing

The invention of next generation sequencing (NGS) methods was pivotal for studying the genome, allowing a human genome to be sequenced in under a day (63). Compared to Sanger sequencing, which took decades to create a draft of the first human genome, NGS was a dramatic advancement. NGS methods massively parallelize the sequencing process by allowing many regions of the genome to be analyzed at once. However, as a trade-off, the length of the reads generated by these methods are shorter (~150 base pairs) than what can be achieved with Sanger sequencing (64). This makes downstream analysis of the data more complex and limits the type of variations that can be captured with the data (65). Furthermore, larger variations, as well as mutations in more complex regions of the genome, often escape detection when using short read sequencing methods (5). Paired end sequencing can improve sequence alignment and provide features to help uncover large variations from short read data. Paired end sequencing involves sequencing from both ends of a DNA fragment and thereby creating two reads from each fragment. Together, these paired reads provide additional spatial information which can aide downstream analysis (66). In general,

NGS methods are not as accurate as Sanger sequencing but the ability to sequence to a higher coverage at low cost can improve the error rate (67).

Illumina

Currently, Illumina sequencing is the most used NGS platform for short-read sequencing (68). DNA molecules are sheared into ~400 base pairs fragments and attached to a flow-cell via an adapter. Each molecule is PCR amplified using bridge amplification to create a cluster of identical single-strands at the same location on the flow-cell. Sequencing by synthesis occurs for each cluster as fluorescently tagged nucleotides are incorporated one at a time, releasing a color corresponding the nucleotide that was just added to the sequence. Pictures are taken of the flow-cell after each iteration of the process to capture the light emitted by each cluster. The sequence of bases is determined based on the sequence of colors emitted by each cluster (69). In paired-end Illumina sequencing, the opposite end of the fragments is attached to the flow-cell so that the process can be repeated to create a read for the other end of the fragment. Illumina datasets are used in each chapter in this thesis.

BGI

BGI genomics also provides short-read sequencing. BGI sequencing differs from Illumina in that it uses DNA nanoballs to amplify the DNA instead of bridge amplification to enhance the signal during sequencing. As with Illumina sequencing, the process starts with sheared DNA fragments but instead of being attached to a flow-cell right away, adapters are attached to the fragments and the two ends of the adapters are ligated together to form a circle (70). Rolling circle replication is

used to amplify small fragments of genomic DNA creating DNA nanoballs containing hundreds of copies of the DNA fragment. Because each replication is based on the original template, nanoball amplification has low amplification bias compared to PCR bridge amplification. The DNA nanoballs are attached to a flow-cell for sequencing and sequencing of the strands occurs in the same manner as Illumina (using sequencing by synthesis). BGI sequencing datasets are utilized in Chapter 2 of this thesis.

Long Read Sequencing

The last decade has seen the introduction of a new generation of sequencing methods with the ability to generate the longest reads lengths yet, with individual sequencing reads being generated often over 10,000 base pairs in length. These longer reads vastly improve the ability to both definitively align reads to a reference genome as well as construct accurate genome assemblies and have shown the capability to identify many of the variant classes that have been inaccessible using short-read methods (71–73). However, as with many new technologies, long-read sequencing is more expensive per base pair than the well-established short read methods. Additionally, these methods tend to have a higher error rate compared to short read methods (72,74). Currently, the high cost often prohibits generating the higher coverage needed to decrease the relatively higher base calling error rate of long-sequencing technology. There are two main long read sequencing technologies being used today.

<u>PacBio</u>

Single molecule, real-time sequencing developed by Pacific Biosciences, referred to as PacBio sequencing, is one technology used to generate long read sequencing data. A single stranded

circular DNA molecule is created by ligating hairpin adapters to either end of a double stranded template DNA fragment (75). This circular molecule is then loaded into a well on a chip where a polymerase, the module used to replicate DNA in a cell natively, is immobilized at the bottom. The polymerase performs replication of the single stranded molecule using nucleotides fluorescently labeled to indicate which of the four nucleotides is incorporated by the polymerase. The light emitted throughout the sequencing process is captured by a camera and translated into the base pair sequences. Because the molecule is circular, the polymerase can continuously sequence the molecule until it ceases to function, producing multiple copies – subreads – of the same template sequence. These subreads can be combined to create higher accuracy consensus reads. This process occurs in parallel across the chip, allowing many molecules to be sequenced at once. In Chapter 3, a PacBio sequencing dataset is used to compare several bioinformatics approaches.

Oxford Nanopore Technologies

The other common long read sequencing technology is from Oxford Nanopore Technologies, referred to as nanopore sequencing. Long, single-stranded molecules are fed through a nanopore with an electric current running through it (76). As the molecule passes through, the different base pair sequences disrupt the current in different, and predictable, ways. These changes in current are recorded and converted into base pair sequences. While not directly utilized within this thesis, targeted sequencing approaches, such as those offered by Oxford Nanopore Technologies, can be used as validation for hypothesis generating studies such as the work presented in Chapter 4.

1.6 BIOINFORMATIC APPROACHES FOR IDENTIFYING GENOMIC VARIATION

Genome Alignment

For all sequencing methods, the genome of a sample is fragmented into pieces, the length depending on the technology being used. In order to interpret the sequences contained in the reads, they are aligned to the reference genome. Genome alignment involves identifying what part of the genome a read is derived from by matching each sequencing read to the most similar sequence in the reference genome. The length of the read and the error rate play important roles in genome alignment. Genome alignment methods have been optimized to work with both short, accurate reads as well as long, error-prone reads. Alignment methods used in this thesis include minimap2 (77) and BWA (78). Once these reads have been aligned to the reference genome, the regions of the genome that differ from the reference genome can be identified using a number of bioinformatic approaches that have been developed to detect the different classes of variation.

Variant Detection

SNPs and Indels

Detecting SNPs from short read sequencing is highly effective given the lower error rate of the technology and because the variants are completely contained within a single read allowing direct comparison between reads and the reference sequence. GATK (Genome Analysis Tool Kit) is the industry standard for this analysis and provides best practice pipelines set up for processing short read data and detecting SNP and indels in samples (79–81).

Detection of these smaller variants can be more challenging to detect with long read sequencing given the higher error rate, specifically given the high prevalence of small insertion and deletion

errors that occur with long read sequencing technologies. Methods to detect SNPs and indels have been developed specifically to handle the higher error rates associated with this these technologies (82).

Microarrays are specifically designed to identify SNP and indel variations and are highly effective at this task. As opposed to detecting variation from sequencing datasets, detecting variation from microarrays does not involve genome alignment as each probe is designed to distinguish between the known alleles at a given location in the genome. SNP and indel alleles are determined directly by reading the signal intensity indicating hybridization at each probe.

CNVs and Inversions

While large variations can be challenging to detect using short reads given the variations can exceed the length of a single read, several methods have been developed to detect large variation using a combination of signatures of the alignments of paired-end reads to the reference genome. These signatures include read depth to detect changes in copy number, distance and orientation in which read pairs align to the reference genome to identify insertions or deletions relative to the reference sequence, and split reads to identify breakpoints of structural variants where split reads are reads that only partially align to the reference genome (79–81).

Long read sequencing has an advantage over short read sequencing for the detection of structural variation as more variations can be fully contained within the ~10,000 base pair read lengths. Differences between a read and the reference sequence can be directly interrogated for variations

that are completely spanned by a long read. Mapping signatures used in short reads such as split reads can also be used to detect variations spanning across multiple reads (83–85).

While there are microarrays designed to detect changes in the number of copies of regions of DNA, these are not used in this thesis.

MEI

Methods for detecting novel MEIs and NUMTs from short reads, such as MELT and dinumt, use similar approaches to methods for detecting CNVs (86,87). Pairs of reads where one read maps to the reference genome while the mate-pair maps to the known sequence of a mobile element identify regions of the genome with novel insertions and split reads are used to refine location of the insertion. Long read sequencing methods can identify mobile element insertions that are undetectable in using short read datasets, specifically in repetitive regions of the genome (88).

Tandem Repeats

Tandem repeats variation has historically been very challenging to characterize using current sequencing technologies. This is due to the difficulty of aligning reads containing repetitive sequence to the reference genome. This is particularly apparent with short reads which may not span the entirety of a repetitive region. However, several methods have been developed to characterize tandem repeats from short read data (89,90). These often again take advantage of the properties of mapped paired end reads to estimate the length of the repeat expansions compared to the reference genome.

Long read sequences have been particularly helpful in the characterization of repetitive variation as spanning the repeat and having unique sequence flanking the repetitive region can be helpful in mapping the read to the correct location in the reference genome and allowing comparison of the lengths between the read and reference sequences (91–93).

1.7 CROSS PLATFORM VARIANT DETECTION STRATEGIES

As each of the approaches discussed above has limitations, there is still room for advancement in detecting variation across the genome. It has been shown on many occasions that integrating multiple sequencing platforms can enhance variation detection (94–97). Often times, these approaches look to leverage information from a more accurate but difficult to obtain datasets to improve or enhance data from more accessible datasets – accessible being either already generated data or data that is less expensive to generate.

Variant Call Validation

While some sequencing technologies such as Sanger sequencing or long read sequencing provide advantages over others for detection of variation, they can be cost prohibitive to perform on a large scale. Instead, these technologies are often used as validation for variant calls made from technologies that may be less accurate (94,98). Targeted Sanger sequencing can be used to validate large, complex, or repetitive variation detected by bioinformatic approaches for short read data. This validation can be automated such as was done in previous work from our group, including myself, using long read sequencing data to validate structural variation calls made from short read datasets (99) (Figure 1).

Variant Imputation

The concept of imputation takes advantage of knowledge taken from short read sequencing data to increase the SNP density for samples genotypes using SNP microarrays. Sequencing datasets provide higher resolution data for variant detection compared to SNP microarrays and provide higher density variant datasets (100). However, variant imputation can be performed to infer what variants are present in a sample without directly interrogating the variants. This technique relies on linkage disequilibrium - the phenomenon that variants are not typically inherited independently of one another and instead, variations near one another are often inherited together (101,102). For example, if an allele in a sample genotyped with a SNP microarray matches closely with an allele in a WGS sample present in a reference population, it can be inferred that the untyped variants in the microarray genotyped sample match the reference sample for the region. While the imputed variants cannot be determined as accurately as directly genotyping the sample, this technique can provide the opportunity to increase the variant density of SNP microarray datasets.

1.8 DISSERTATION SUMMARY

In this thesis, I have developed and employed multiple bioinformatic approaches to analyze datasets produced by multiple sequencing technologies. In Chapter 2, I compare variant calls generated from BGI sequencing of individuals of Ukrainian descent using the more established short read sequencing platform Illumina and SNP microarrays. Additionally, we identify the key genomic variation that differentiates Ukrainians from others of European ancestry. In Chapter 3, I perform a comparison of various bioinformatics and sequencing approaches for detecting tandem

repeat variation and provide guidelines for when to use each technique. In Chapter 4, I utilize the relationship between tandem repeat lengths and surrounding SNP haplotypes gleaned from whole genome, short reads sequencing to estimate the lengths of tandem repeats in disease samples genotyped using SNP microarrays. We identify repeat expansions of interest relating to ALS from both the samples with WGS and SNP microarray data that should be further investigated using targeted sequencing approaches. This thesis highlights the important role executing multi-platform approaches can play in detecting genomic variation across both disease and diversity studies.

1.9 FIGURES

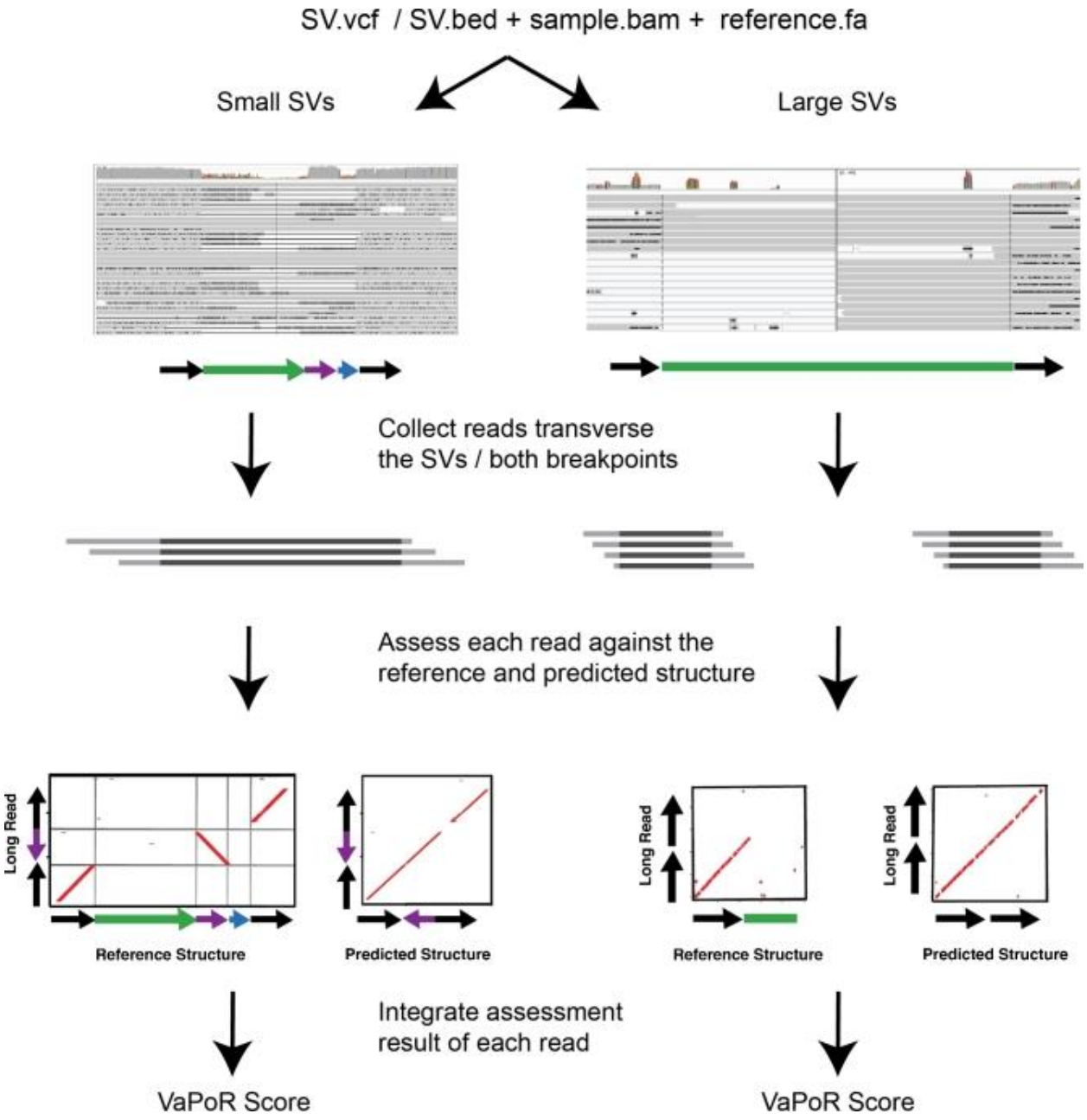

Figure 1-1 Flowchart describing the VaPoR algorithm.

As input, the algorithm requires a set of structural variants in either VCF or BED format, a series of long reads and/or sequence contigs in BAM format, and the corresponding reference sequence. VaPoR then interrogates each variant individually at its corresponding reference location, assesses the quality of the region, and assigns a score.

Chapter 2 - Characterizing Population Scale Variation Using Multiple Sequencing Platforms

This chapter is a published work:
Taras K Oleksyk*, Walter W Wolfsberger,* Alexandra M Weber*, Khrystyna Shchubelka*, Olga T Oleksyk, Olga Levchuk, Alla Patrus, Nelya Lazar, Stephanie O Castro-Marquez, Yaroslava Hasynets, Patricia Boldyzhar, Mikhailo Neymet, Alina Urbanovych, Viktoriya Stakhovska, Kateryna Malyar, Svitlana Chervyakova, Olena Podoroha, Natalia Kovalchuk, Juan L Rodriguez-Flores, Weichen Zhou, Sarah Medley, Fabia Battistuzzi, Ryan Liu, Yong Hou, Siru Chen, Huanming Yang, Meredith Yeager, Michael Dean, Ryan E Mills, Volodymyr Smolanka, Genome diversity in Ukraine, GigaScience, Volume 10, Issue 1, January 2021, giaa159, https://doi.org/10.1093/gigascience/giaa159
(*indicates co-first authorship)

The work presented in this chapter of the dissertation was published in GigaScience. The project conceptualization, methodology and project administration were conceived by Dr. Taras Oleksyk. I performed the bioinformatic analysis required for the creation of the variant callsets as well as the comparison analysis of variations detected across the different sequencing and genotyping methods. Walter Wolfsberger identified functionally and medically relevant variants and performed the population structure analyses. Khrystyna Shchubelka was involved in the acquisition and sequencing of the biological samples. Please see the published paper for the contributions of the remaining authors.

2.1 ABSTRACT

Background

The main goal of this collaborative effort is to provide genome-wide data for the previously underrepresented population in Eastern Europe, and to provide cross-validation of the data from genome sequences and genotypes of the same individuals acquired by different technologies. We collected 97 genome-grade DNA samples from consented individuals representing major regions of Ukraine that were consented for public data release. BGISEQ-500 sequence data and genotypes

by an Illumina GWAS chip were cross-validated on multiple samples and additionally referenced to one sample that has been re-sequenced by Illumina NovaSeq6000 S4 at high coverage.

Results

The genome data have been searched for genomic variation represented in this population, and a number of variants have been reported: large structural variants, indels, copy number variations, single-nucleotide polymorphisms, and microsatellites. To our knowledge, this study provides the largest to-date survey of genetic variation in Ukraine, creating a public reference resource aiming to provide data for medical research in a large understudied population.

Conclusions

Our results indicate that the genetic diversity of the Ukrainian population is uniquely shaped by evolutionary and demographic forces and cannot be ignored in future genetic and biomedical studies. These data will contribute a wealth of new information bringing forth novel, endemic and medically related alleles.

2.2 INTRODUCTION

Context

Ukraine is the largest country located fully in Europe, with a population that was formed as a result of several millennia of migration and admixture. It occupies the intersection between the westernmost reach of the great steppe and the easternmost extent of the great forests that spread across Europe, at the crossroads of the great trade routes from “Variangians to the Greeks” along the river Dnipro, which the ancient Greeks referred to as Borysthenes, and the Silk Road linking

civilizations of Europe and Asia (103). This land has seen the great human migrations of the Middle Ages sweeping from across the great plains, and even before that in the more distant past, of the early farmers (104) and the nomads who first domesticated the horse (105–108). Here, at the dawn of the modern human expansion, our ancestors met the Neanderthals who used to hunt the great game along the glacier of the Ice Age (109,110).

The rich history shaped genetic diversity in the population living in the country of Ukraine today. As people have moved and settled across this land, they have contributed unique genetic variation that varies across the country. While the ethnic Ukrainians constitute approximately more than three-quarters of the total population, this majority is not uniform. A large Russian minority compose approximately one-fifth of the total population, with higher concentration in the southeast of the country. Smaller minority groups are historically present in different parts of the country: Belarusians, Bulgarians, Crimean Tatars, Greeks, Gagauz, Hungarians, Jews, Moldovans, Poles, Romanians, Roma (Gypsies), and others (111).

This study offers genome data from 97 individuals from Ukraine (Ukrainians from Ukraine [UAU]) to the scientific community to help fill the gaps in the current knowledge about genomic variation in Eastern Europe, a part of the world that has been largely and consistently overlooked in global genomic surveys (112). To our knowledge, this was the first effort to describe and evaluate the genome-wide diversity in Ukraine. Samples were successfully sequenced using BGI's DNA Nanoball (DNBSEQ™) sequencing technology and cross-validated by Illumina sequencing and genotyping. The major objectives of this study were to demonstrate the importance of studying local variation in the region and to demonstrate the distinct and unique genetic components of this

population. Of particular interest were the medically related variants, especially those with allele frequencies that differed with the neighboring populations. As a result, we present and describe an annotated dataset of genome-wide variation in genomes of healthy adults sampled across the country.

Dataset

The new dataset includes 97 whole genomes of self-reported UAU at 30× coverage sequenced using BGISEQ-500 (one of the range of DNBSEQ™ sequencers; BGI Inc., Shenzhen, China) and annotated for genomic variants: single-nucleotide polymorphisms (SNPs), indels, structural variants, and mobile elements. The samples were collected across the entire territory of Ukraine, after obtaining institutional review board (IRB) approval for the entire study design and informed consent from each participating volunteer. Each participant in this study had an opportunity to review the informed consent, received an explanation of the nature of the genome data, and made a personal decision about making it public.

The majority of samples in this study (86 of 97) were additionally genotyped using Illumina Global Screening Array (Illumina Inc., San Diego, CA, USA) to confirm the accuracy of base calling between the 2 platforms. In addition, one sample (EG600036) was also sequenced on the Illumina NovaSeq 6000 S4 (2 × 150 bp; ~60× coverage) and used for validation of the variant calls (see summary in Table 2-2 and full sequencing statistics for individual samples in Table 2-3). The list of the cross-validated samples and the source technology of the data is presented in Table 2-1.

The present dataset contains locations and frequencies of >13 million unique variants in UAU that are further interrogated for functional impact and relevance to medically related phenotypes (Table 2-4 and data in GigaDB (113)). As much as 3.7% of these alleles, or 478,000, are novel genomic SNPs that have never been previously registered in the Genome Aggregation Database (gnomAD) (30) (Table 2-4). This number is similar in magnitude to what was reported earlier in 2 populations from European Russia (3–4% (114)). Many of the discovered variants (12.6%) are also currently missing from the global survey of genomic diversity in the 1000 Genomes Project (1KG) (24). The majority of these described variants are rare or very rare (<5% Appendix Figure A 1).

Because other indigenous ethnic groups from Ukraine (such as the Crimean Tatars or the Gagauz) are not included in the study, increasing the aforementioned sample size from 100 to 1,000 individuals is not likely to greatly contribute to discovery of novel mutations (115). The proportion of the novel structural variants and mobile elements compared to the earlier databases is even higher: almost 1M (909,991) complex indels, regions of simultaneous deletions and insertions of DNA fragments of different sizes that lead to net a change in length, the majority of which are novel (Table 2-4). Many of the newly discovered variants are functional and potentially contribute to the phenotype (classified in Table 2-5). We report many important variants that are overlooked or require special modifications in the commonly used resources and tools in genomic research and diagnostics. This wealth of novel variation underscores the importance of variant discovery in local populations that cannot be ignored in biomedical studies.

2.3 RESULTS

Variant calling and confirmation

For each sample in the database, we estimated the number of passing bi-allelic SNP calls (i.e., loci with the non-reference genotypes relative to the most current major human genome assembly, GRCh38 (116)) (Table 2-4). Then ~12% of these were filtered out on the basis of excess heterozygosity and low variant quality scores (Table 2-3). For the indels, we also estimated the number of passing calls compared to GRCh38 and excluded 4% of those that did not pass filtering. The total number of the unique SNPs, small and large indels (Table 2-4) was calculated from the raw read alignments of all 97 sequenced genomes (Total Unique SNPs, Table 2-3) with the exception of those filtered out for low variant quality scores and containing excess heterozygosity (Filtered Count; Table 2-3). In addition, we filtered out 4,135,903 variants that only appeared once in a single sample (for both indels and SNPs) and designated them as "singletons."

We report a good correspondence between the SNP calls made using BGISEQ-500 and NovaSeq 6000 S4 data. A comparison of the variants detected using these 3 platforms for sample EG600036 is summarized in Figure 2-1A. The SNP concordance for samples with both BGISEQ-500 and SNP array data is summarized in Figure 2-1C. The cross-platform comparison shows a very good overlap across all 3 technologies: >3.5 M (97.7%) of the SNPs identified in the BGISEQ-500 were also verified in the whole-genome sequence of EG600036 sequenced by the Illumina NovaSeq 6000 S4. The correspondence with the Illumina SNP Array for sample EG600036 was also very good: 95.8% of all the SNP genotypes called by the Illumina method were also detected by the BGISEQ-500 (Figure 2-1A, right, and C, right). The concordance between the non-reference alleles between the 2 platforms in all 86 samples was nearly linear (r2= 0.985, Figure 2-1C, left).

The transition/transversion (TITV) ratio for the novel SNPs (estimated with TiTvtools (117) and visualized by plotTiTv in Supplementary Figure 2-2) was lower than the TITV ratio for SNPs in the dbSNPs database (1.9 vs 2.2; (31)). Similarly, the insertions to deletions (ins/del) ratio for novel indels is lower than for the indels already reported in the dbSNP database (0.63 vs 0.75). This observation likely reflects our improved ability to detect small insertions in newer sequencing technologies compared to many platforms that historically submitted variation to dbSNP.

We have defined the multi-allelic SNPs as observations of genomic positions having 2 or more alternative alleles (118). These are important variants that are overlooked or require special modifications in the commonly used resources and tools in genomic research and diagnostics. We report a total of 343,696 multiallelic sites in the sequences from our sample, of which 2.0% are at locations unreported in the gnomAD database (30) (Table 2-4).

In addition to the SNPs, we have identified and quantified major classes of structural variations in the Ukrainian population: small indels (insertions and deletions <50 bp), large structural variants (deletions, duplications, and inversions > 50 bp), and mobile element insertions (MEI) (Alu elements [ALU], L1 elements, non-autonomous retroelements [SVA], and nuclear mitochondrial DNA [NUMT] copies). A number of structural elements were reported, including common and novel ones. While among the small variants most were common (6–9%), a large proportion of large variants and MEIs (38–52%) have not been reported previously in the 1GP Database (Table 2-4).

Once more, there is a significant correspondence between the calls made using BGISEQ-500 and Illumina NovaSeq 6000 S4 data. The 2 sequencing platforms show a significant overlap in calling indels (DEL): 87.9% of the variants called by the BGISEQ-500 were also detected by the Illumina platform. At the same time, there were 822 deletions, or 33.8% of all the indels called by the Illumina that were not detected by the BGISEQ-500 (Figure 2-1B). A similar picture, where BGISEQ-500 performs competitively well, is also observed for inversions (INV) (Figure 2-1B) and LINE1 transposable elements (Figure 2-1D). At the same time, there were more duplications (DUP) (Figure 2-1B) and the 2 classes of transposable elements evaluated: ALU and SVA (Figure 2-1D). Evaluation tests show that current algorithms are platform dependent, in the sense that they exhibit their best performance for specific types of structural variation, as well as for specific size ranges (119), and the algorithms designed for detection and archived datasets are predominantly for Illumina pair-end sequencing (120,121). While it is possible that these results indicate Illumina's superiority at detecting structural variation, it can also be the consequence of the bioinformatics tools for calling structural variants developed using mainly the Illumina data, as suggested by previous comparative evaluations of the 2 technologies (122,123). Additionally, higher coverage of the Illumina data (60×) could have contributed to the differences observed between the platforms.

The database was compared to the existing global resources of population variation such as gnomAD (30) and the 1KG (24). Specifically, under our search criteria, the small variants (SNPs and small indels) were considered “novel” if they were absent from all the samples in the 2 global datasets (gnomAD and 1KG; Table 2-4). The large structural variants and MEIs were considered novel if the variant was not present in the gnomAD and 1KG databases. To determine whether a

given variant was present in 1 of the databases, a variant of the same type in the database had to overlap the Ukrainian variant with a minimum fraction of 0.95. We observed no significant deviation of the rate at which reference bases were observed at REF/alt heterozygous SNP sites (reference bias was near 50%).

Collection of functional variants

A particular interest in this study is the distribution of functional variation, not in the least due to the potential impact on phenotypes, especially to those with medical relevance (124). As much as 97.5% of all annotated variation was discovered outside of the known functional elements (upstream, downstream, intron, and intergenic). These results are similar to the expected distributions of mutations shown with the simulated data (125). Nevertheless, there were >8,000 mutations discovered within exons of each individual on average (top half of Table 2-5). We annotated several classes of functional mutations within the coding regions (bottom half of Table 2-5). As expected, the nonsense mutations classified in the annotation file as "disruptive in-frame indel," "start lost," "stop gained," and "stop loss" were rare, while categories with minimal effect on the function, such as "synonymous," "motif," "protein folding," and "missense," were more common. Some of the mutations listed in the annotation file can be classified in >1 category (e.g., "synonymous variants" can also be counted in "exonic variants").

In addition to the gene-coding mutations, we report a number of regulatory variants. For example, the database contains a total of 2,229 transcription factor binding site ablation (TFBS) mutations (bottom half of Table 2-5). A summary of functional variation discovered in this study is presented

in Table 2-5. The full annotation database with classifications is available alongside the associated data deposited in GigaDB (113).

Collection of medically relevant variants

Many of the reported variants are already known to be medically related and are listed either in genome-wide association studies (GWAS) (126) or ClinVar (an NCBI archive of reports of the relationships among human variations and phenotypes with supporting evidence)(33) catalogues (Table 2-6). Our database contains a total of 43,892 benign mutations in medically related genes but also 189 unique pathogenic or likely pathogenic variants, as well as 20 protective or likely protective alleles as defined in ClinVar (33,127). Each individual in this study carries 19 pathogenic and 12 protective mutations on average. While at least some individuals were homozygous for the pathogenic allele, none of the associated disease phenotypes have been reported, which could be largely attributed to heterozygosity, age-dependent penetrance, expressivity, and gene-by-environment interactions (128,129).

As expected, our study shared a lot more variants with the GWAS (126) than with the ClinVar (33) catalogue. While GWAS has recently become the tool of choice to identify genetic variants associated with complex disease and other phenotypes of interest (130), because the amount of genetic variance explained by these variants is low, they are generally not very useful for predicting pathogenic phenotypes (131). It is also important to note that not all ClinVar variants carry the same weight of supporting evidence; attributing disease causation to prioritized variants remains an inexact process and some of the reported associations eventually are proven to be spurious (132). Nevertheless, the importance of the unique set of mutations published here is difficult to

overemphasize because it constitutes the first published set of pathological variants in an understudied population, an important step towards a local catalogue of medically relevant mutations. In addition, as the attention in the genomic community is shifting from monogenic to polygenic traits, many of these may become relevant in future research and exploration (133).

Disease variants with frequencies that differed between the Ukrainians and the neighboring populations are of particular interest to the medical community. It is well established that differences in allele frequencies are a consequence of evolutionary forces acting in populations (such as drift, mutation, migration, nonrandom mating, and natural selection) and that certain diseases and heritable traits display marked differences in frequency between populations (134). With this in mind, we created a list of the known disease variants whose frequencies differ between Ukrainians and other European populations (the combined European sample [EUR] from the 1KG, comprising Utah residents [CEU] with Northern and Western European ancestry, Toscani in Italy [TSI], Finnish in Finland [FIN], British in England and Scotland [GBR], Iberian population in Spain [IBS] (24,135), and French population from Human Genome Diversity Project [HGDP] [FRA] (136)) and Russians from HGDP (RUS) (136). Several examples of these variants are presented in Table 2-7. Among these are variants involved in a number of medical conditions such as hyperglycinuria/iminoglycinuria (rs35329108,SLC6A19), efficacy of bisphosphonate response (rs2297480,FDPS), autism (rs7794745, CNTNAP2), Leber congenital amaurosis (rs10151259,RPGRIP1), and breast cancer susceptibility in BRCA1 and BRCA2 carriers (rs1801320,RAD51) (Table 2-7).

Of course, not all the medically related variants are currently known, and many remain to be discovered and verified in local populations. This is, to some extent, a consequence of underreporting of allelic endemism within understudied populations, particularly in Eastern Europe (112) but also elsewhere (137,138). By offering public annotations of functional mutations in a population sampled across the territory of Ukraine, our database contributes a number of candidates to direct future research in medical genomics. We chose only the markers with the highest non-reference allele frequency differences compared to the neighboring populations EUR(24) and RUS (136), evaluated by the Fisher exact test, and listed them in Table 2-8.

Population structure and ancestry informative markers

We performed several population analyses, but only to demonstrate the uniqueness and usefulness of this new dataset. Our results indicate that genetic diversity of the Ukrainian population is uniquely shaped by evolutionary and demographic forces and cannot be ignored in future genetic studies. However, we do not evaluate any historical hypotheses on the timing of origins, founding, migration, and admixture of this population and use only the naive approaches, based on the statistical models.

To demonstrate the extent to which our dataset contributes to the genetic map of Europe, we explored genetic relationships between Ukrainian individuals within our sample and evaluated genetic differences between this population and its immediate neighbors on the European continent for which population data of full genome sequences were publicly available. A principal component analysis (PCA) of the merged dataset of 654 samples included European populations from the 1KG (Utah residents [CEU] with Northern and Western European ancestry, Toscani in

Italy [TSI], Finnish in Finland [FIN], British in England and Scotland [GBR], Iberian population in Spain [IBS] (24,135)) and French (FRA) and Russian (RUS) populations from the HGDP (136), as well as the relevant high-coverage human genomes from the Estonian Biocentre Human Genome Diversity Panel (EGDP: Croatians [CRO], Estonians [EST], Germans [GER], Moldovans [MOL], Polish [POL], and Ukrainians [UKR]) (139) and Simons Genome Diversity Project (Czechs [CZ], Estonians [EST], French [FRA], Greeks [GRE], and Polish [POL]) (139) (Figure 2-3). The latter article also identifies "Cossacks" as a separate self-identified ethnic group within Russians (Cossacks [RUS]) or Ukrainians (Cossacks [UKR]) (140) (Table 2-1).

Ukrainian genomes from this as well as other studies (139,140) form a single cluster positioned between the Northern (Russians, Estonians) on 1 side, and Western European populations on the other (CEU, French, British, and Germans, Figure 2-3). There was a significant overlap with the other Central and Eastern European populations, such as Czechs, Polish, and the people from the Balkans (Croats, Greeks, and Moldovans). This is not surprising; in addition to the close geographic distance between these populations, this may also reflect the insufficient representation of samples from the surrounding populations (see data in GigaDB (113)). Similarly, the admixture analysis demonstrates distinctiveness of our dataset but also demonstrates unique combinations of genetic components that may have shaped this population (Figure 2-4 and Appendix Figure A 2).

Addition of the new genomic data will most likely add to the resolution of the genetic map of this region and further reveal differences between the populations of Eastern and Central Europe. Our dataset showed a limited amount of inbreeding and contains information for future population studies. This database can be a starting point for association studies, as the candidate ancestry

informative markers (AIMs) (141) can be used for mapping disease alleles by admixture disequilibrium (142,143).

To provide a more extended view of the genetic components contributing to the Ukrainian population, we used the population structure plots using the ADMIXTURE package (144). This allowed us to construct a preliminary picture of putative ancestry contributions and population admixture. To identify the optimal K, we implied the 10-fold cross-validation function in range of K = 2−6. The results with the optimal K = 3 shown in Figure 2-4 illustrate similarity and the difference of Ukrainian population compared to the other populations in Central and Eastern Europe (Figure 2-4, second row). While the higher values of K (Appendix Figure A 2) show an increasing number of clusters, they also show an increasing amount of error in the cross-validation function. This analysis already shows the potential of the present database in helping to resolve population structure in Eastern Europe, but additional genome-wide data from neighboring populations would be helpful to refine the picture in this geographical region. Unfortunately, valuable genome-wide data collected from 3 populations in Russia have been retracted from public databases after publication (114).

Despite the fact that all of the samples were collected from self-identified ethnic Ukrainians, there were 2 notable outliers: sample EG600048 clustered with the Southern Europeans (Iberia and Italian populations) Figure 2-3. This illustrates an important point that ignoring the unique composition of the population will result in ascertainment bias in biomedical studies. Genetics is not a reliable determinant of ethnicity but can be used to evaluate individual contributions of ancestry.

People of Ukraine carry many previously known and several novel genetic variants with clinical and functional importance that in many cases show allele frequencies different from neighboring populations in the rest of Europe, including Poland to the west, Romania to the south, the Baltics to the north, and Russia to the northeast. While several large genome projects already exist contributing to the understanding of global genetic variation, many rare and endemic alleles have not yet been identified by international databases such as 1KG and are currently not available in standard genotyping panels for association testing for human diseases, and glaring white spots still exist on the genetic maps in local populations of Eastern Europe (112). We fully expect future sampling and sequencing to continue to improve and complete the detailed picture of genomic diversity in people across the country and contribute to the further development of genetic approaches in biomedical research and applications.

2.4 METHODS

Sampling strategy

The collection and consent procedure was approved as part of the "Genome Diversity in Ukraine" project by the IRB of Uzhhorod National University in Uzhhorod, Ukraine (Protocol 1 from 09/18/2018, Supplementary File S1). We employed doctors and medical professionals from different regions of Ukraine to oversee collection of blood samples at hospitals. Healthy (non-hospitalized) volunteers were contacted through advertisements and invited for personal interviews at outpatient offices. During the visit the volunteers were familiarized with the study and the collection procedure and gave full consent to participate and let their genotypic and phenotypic data be freely and publicly available. During each interview, the volunteer participants

also completed a questionnaire indicating self-reported region of origin, place of birth of all 4 grandparents (if remembered), sex, and several phenotypical features, such as daily history of disease (Table 2-1). The hard copies of the consents and personal interviews remain sealed and stored at the Biology Department of Uzhhorod National University. After the conclusion of the interview and sample collection, all personal identifiers were removed from the vials containing blood samples, except for an alphanumeric identifier and a barcode. All the subsequent analysis and publication was done in a blind design where neither the participants nor the researchers could identify the person who donated the sample.

At the conclusion of the interview a whole-blood sample was collected from a vein into two 5-mL EDTA tubes by a certified nurse or a phlebotomist, assigned a barcode number, and shipped by courier on dry ice to a biomedical laboratory certified to handle blood samples in Uzhhorod, Ukraine (Astra Dia Inc.), for DNA extraction immediately on arrival. The excess of the blood and DNA from samples remaining after the genetic analysis is stored frozen at the biobank of the Biology Department, Uzhhorod National University, Ukraine. As a result, blood samples were collected from a total 113 individuals.

DNA extraction

Immediately upon arrival to the laboratory, DNA isolation from 200 μL of blood was carried out with the innuPREP DNA Blood Minikit (AAnalytik Jena GmbH, Jena, 07745, Germany). High molecular weight genomic DNA was lightly fragmented by vortexing. The initial DNA concentration was measured with the Implen C40 Nanophotometer (München, Germany), and quality was verified visually on a 2% agarose gel. The 97 successfully extracted DNA samples

were normalized to 20–30 ng/μL concentration for downstream application. After extraction the samples were recoded and sent to NIH for the genotyping procedure, whence the aliquots were further shipped to a BGI facility (BGI Shenzhen, China) or to Psomagen Inc. (Gaithersburg, MD, USA) for the whole-genome sequencing (WGS). The remaining ~2 mL was frozen for future use.

Sequencing and genotyping

All 97 individuals in this study were sequenced with BGISEQ-500 and 88 individuals were cross-validated by genotyping using Illumina Global Screening Array. The record of which individual samples have been cross-validated by both technologies is presented in Table 2-1. In addition, a single sample (EG600036) was also sequenced on Illumina NovaSeq 6000 S4 (~60× coverage).

Sequencing with BGISEQ-500

All 97 DNA samples were sequenced on BGISEQ-500 (BGI Shenzhen, China). Upon receipt at the BGI facility, and prior to sequencing, samples were checked again for quality. Concentration was once more detected by fluorometer or Microplate Reader (e.g., Qubit Fluorometer, Invitrogen). Sample integrity and purity were detected by agarose gel electrophoresis (concentration of agarose gel: 1%; voltage: 150 V; electrophoresis time: 40 min). Aliquots of 1 μg genomic DNA were fragmented by Covaris. The fragmented genomic DNA was selected by Agencourt AMPure XP-Medium kit to a mean size of 200–400 bp. Fragments were end-repaired and then 3′-adenylated. Adaptors were ligated to the ends of these 3′-adenylated fragments. PCR products were purified by the Agencourt AMPure XP-Medium kit. The double-stranded PCR products were heat denatured and circularized by the splint oligo sequence. The single-strand circle DNA was formatted as the final library. The qualified libraries were sequenced by BGISEQ-500:

the single-strand circle DNA molecule formed a DNA nanoball (DNB) containing >300 copies through a rolling-cycle replication. The DNBs were loaded into the patterned nanoarray by using high-density DNA nanochip technology. Finally, pair-end 100-bp reads were obtained by combinatorial probe-anchor synthesis. Raw reads were filtered to remove adaptor sequences, contamination, and low-quality reads. Sequencing of all 97 full genome samples submitted for sequencing at BGI was successful.

Short-read sequencing with Illumina NovaSeek6000

One individual was resequenced by Illumina NovaSeq6000 S4 at Psomagen Inc. (Gaithersburg, MD, USA). The library was prepared using TruSeq DNA PCR Free 350 bp protocol by Illumina. The library was sequenced at ~64× depth, producing 150-bp–long reads, resulting in 241.7 Gb of data.

Genotyping with the Illumina Infinium Global Screening Array

We attempted to genotype all 97 of the collected samples using the Illumina Infinium Global Screening BeadChip Array-24 v1.0 (GSAMD-24v1–0) for 700,078 loci at the National Cancer Institute's Division of Cancer Epidemiology and Genetics (Bethesda, MD, USA) (145). Data were analyzed by using the standard Illumina microarray data analysis workflow. During quality control (QC), samples were filtered for contamination, completion rate, and relatedness. As part of QC, we performed ancestry assessment using SNPweights software (141) with a reference panel consisting of 3 populations (European, West African, and East Asian). All samples were attributed to the European ancestry group. After QC and sample exclusion, 87 (86 samples and 1 QC) samples with 689,918 loci and completion rate of 99.9% were retained for further analysis.

Variant Calling

Variant Calling of the BGISEQ-500 data

The sequencing data produced using the BGISEQ-500 platform for 97 samples were analyzed using the Sentieon tools (Sentieon Inc., San Jose, CA, USA) high-performance implementation of the BWA/GATK best practices pipeline on servers hosted by the Cornell University Biotechnology Resource Center. Reads were aligned to the GRCh38 human reference genome using BWA-MEM (Version: 0.7.16a-r1181), and mapped reads were prepared for variant calling using GATK (v3.8-1-0-gf15c1c3ef by Broad), including marking duplicates (picard MarkDuplicates, Version 2.12.1), indel realignment (GATK RealignerTargetCreator, IndelRealigner, Version 3.7-0), and base quality score recalibration (GATK BaseRecalibrator, PrintReads, Version 3.7-0). SNP and indel discovery were performed for each individual using GATK HaplotypeCaller and merged into a single pVCF using bcftools. Sample EG600036 was also run without joint calling, which was used when calculating concordance between the Illumina and BGISeq variant callsets, estimated with TiTvtools and visualized by plotTiTv (117).

Repetitive variant calling

Mobile element discovery was performed using MELT (Version 2.2.0) (86) and structural variant discovery using lumpy-sv with Smoove (Version 0.2.5) (146). Short tandem repeats were called using GangSTR (Version 2.4.2) (89) and nuclear mitochondrial DNA using dinumt (87).

Data validation and quality control

Variant files were compared for consistency across the 3 different platforms: BGISEQ-500 sequencing, Illumina genotyping, and Illumina NovaSeq6000 S4 sequencing. Illumina genotyping was performed on 86 of the 97 samples previously sequenced with BGISEQ-500. Additionally, 1 sample (EG600036) was also sequenced with Illumina NovaSeq6000 S4. The variant detection programs were rerun without joint calling for the BGISEQ-500 sequencing for sample EG600036 for comparison with the single Illumina-sequenced sample. In this sample, the SNPs derived from the WGS platforms were compared to those identified using the Illumina SNP array for both matching position and matching genotype. Structural variants and MEIs were compared between the WGS platforms in EG600036. Variants were considered the same if they had 95% reciprocal overlap. Overall, we found that Illumina identified a higher number of larger variants than BGISEQ-500. This could potentially be due to its higher coverage (~60×) compared to BGISEQ-500 (~30×). However, because both have high coverage, we may see diminishing returns for coverage >30×. An alternative explanation is that the variant identification tools have been built to detect variation from Illumina sequencing data and therefore may not be able detect variants in BGISEQ-500 data as accurately.

Annotation

Sequence variant files were annotated using ANNOVAR (ANNOVAR, RRID:SCR_012821) (147) and SNPEff (SNPEff, RRID:SCR_005191) (148) software using GRCh38 reference databases. The following databases were used for the For ANNOVAR annotations: RefSeq Gene, 1GP superpopulation, dbSNP150 with allelic splitting and left-normalization. For annotation of the medically related and functional variants we used ClinVar Version 20200316 (33), InterVar gnomeAd Version 3.0 (30), and dbnsfp Version 35c (149). For SNPEff, the default GRCh38

annotation database (150) was complemented with ClinVar (ClinVar, RRID:SCR_006169) (33) and GWAS catalog (126) database annotation using the snpSift tool (snpSift, RRID:SCR_015624) (151).

Population analysis

Principal component analysis

For PCA, we used WGS variants of our samples and merged them with samples from neighboring countries available from the European samples from 1KG (Utah residents [CEU] with Northern and Western European ancestry, Toscani in Italy [TSI], Finnish in Finland [FIN], British in England and Scotland [GBR], Iberian population in Spain [IBS] (24,135)) and French (FRA) and Russians (RUS) from HGDP (136), as well as the relevant high-coverage human genomes Croatian (CRO), Czech (CZ), Estonian (EST), German (GER), Greek (GRE), Hungarian (HUN), Moldovan (MOL), Polish (POL), Russian Cossack (RUS), and Ukrainian (UKR) from the EGDP (139), and the Simons Genome Diversity Project (140). The analysis was performed with Eigensoft (Eigensoft, RRID:SCR_004965) (152).

To produce a meaningful number of alleles to analyze, the resulting dataset was filtered by genotyping rate (1) and pruned for variants in linkage disequilibrium by excluding those with high pairwise correlation within a moving window (–indep-pairwise 50 10 0.5). This resulted in 677 samples with 208,945 variants. We used EIGENSOFT (152) to calculate the eigenvectors, of which PC1 and PC2 were visualized using Python programming language, with pandas, matplotlib, and seaborn libraries (153). Two extreme outlier samples (EG600056, and EG600052)

were excluded from the visible range of the PCA plot because they clustered with each other far away from any known European group.

Model-based population structure analysis

For the naive (model-based) structure analysis, we used the same dataset described in the PCA (above). The analysis was performed using ADMIXTURE software (ADMIXTURE, RRID:SCR_001263) (144). For identification of the optimal K parameter, we used the 10-fold cross-validation function of ADMIXTURE in the range 2–6, with K = 3 resulting in the lowest error, deeming it optimal. The results were visualized using Python programming language, with pandas, matplotlib, and seaborn libraries (153,154) to construct a population structure plot using samples from the 1KG (Utah residents [CEU] with Northern and Western European ancestry, Toscani in Italy [TSI], Finnish in Finland [FIN], British in England and Scotland [GBR], and Iberian population in Spain [IBS]) and French population (FRA) (24,135) and Russians (RUS) from HGDP (136), as well as the relevant high-coverage human genomes from the EGDP (139), and Simons Genome Diversity Project (140). The resulting plot with K = 3 is presented in Figure 2-4, and plots with K = 4 to K=8 are in Appendix Figure A 2.

Inbreeding estimates

We estimated inbreeding coefficients for all the genotype samples in the same dataset. For this analysis the samples were pruned for genotyping rate (>0.9) and linkage disequilibrium by excluding those with high pairwise correlation within a moving window (PLINK parameter –indep-pairwise 50 10 0.1). Using the resulting dataset containing the remaining 117,641 loci from 84 samples, we performed several inbreeding estimates: (i) method-of-moments F-coefficient

estimates, (ii) variance-standardized relationship minus 1 estimates, and (iii) F-estimates based on correlation between uniting gametes (155).

2.5 DISCUSSION

Reuse potential

Since the publication of the first human genome (4,156) and the first surveys of worldwide variation such as 1KG (24,135), efforts have been directed outward, to expanding the exploration of human diversity across the world and filling out more and more "white spots" of genome variation (114,140), as well as inward, to fill the remaining white spots in the human genome itself: to map the remaining gaps in the chromosome assembly and identify new structural and functional variation (10) and to map the 3D structure of the human genome (157). The new data present a valuable addition to the former and represent the first exploration of the genome landscape in the important component of European genomic diversity.

The genome diversity of Ukraine is an important clue to advance modern genome studies of the population history of Europe. The country is positioned in the crossroads of the early migration of modern humans and the westward expansion of the Indo-Europeans, and represents an aftermath of centuries of migration, admixture, and demographic and selective processes. As wave after wave of great human migrations moved across this land for millennia, they were followed by the exchange of cultural knowledge and technology along the great trade routes that continue to transect this territory until the present day.

The justifications for collecting, sequencing, and analyzing populations from this part of Europe have been outlined previously (112,113), and the new database is a step in that direction. Given its unique history, the genome diversity data from Ukraine will contribute a wealth of new information, bringing forth different risk and/or protective alleles that neither exist nor associate with disease elsewhere in the world. This project identified 13M variants in Ukrainians of which 478,000 were novel genomic SNPs currently missing from global surveys of genomic diversity (30,114). We also report almost 1M (909,991) complex indels, regions of simultaneous deletions and insertions of DNA fragments of different sizes that lead to a net change in length, with only 713,858 previously reported in gnomAD (30) (Table 2-4). The newly discovered local variants can be used to augment the current genotyping arrays and used to screen individuals with genetic disorders in GWAS, in clinical trials, and in genome assessment of proliferating cancer cells.

The present project is built upon the open release/access philosophy. The data have been released and can be used to search for population ancestry markers, as well as medically related variants, in subsequent studies. The public nature of the data deposited on the specially created web resource located at Uzhhorod National University will ensure that the nation's biomedical researchers will receive access to a useful information resource for future projects in genomics, bioinformatics, and personalized medicine. Engaging local Ukrainian scientists in this collaborative international project lays the foundation for future studies and ensures their participation in the worldwide research community.

Data Availability

The raw reads are available at the SRA (Project PRJNA661978, SUB7904361). All other datasets mentioned in this project are available in *GigaScience* GigaDB.

2.6 FIGURES AND TABLES

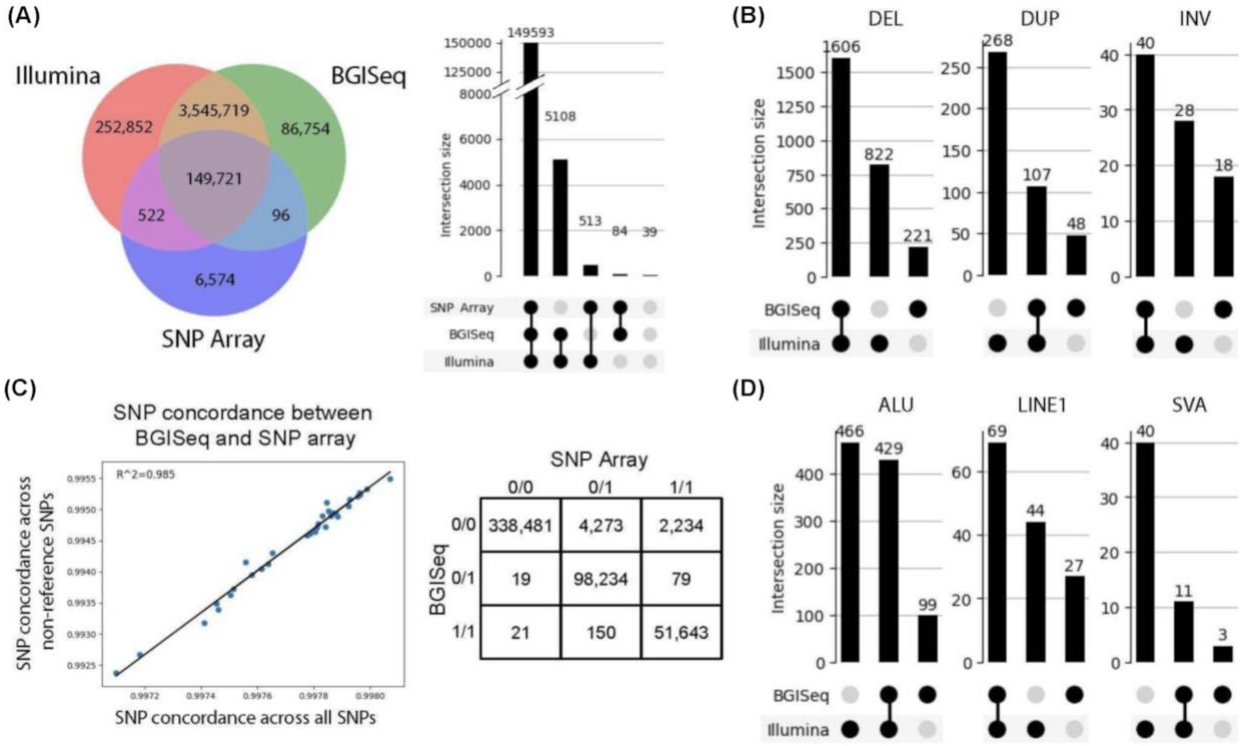

Figure 2-1 Variant concordance across the 3 sequencing/genotype methods

Left: Overlap of SNP positions identified in 1 sample (EG600036) using each of the 3 platforms. Right: Concordance of SNP genotypes in 1 sample derived from each of the 3 platforms. This only includes the subset of SNPs with alternate alleles included in the Illumina genotyping array (the smallest of the 3 variant sets). The variants indicated as belonging to none of the categories are variants whose genotypes differ between all 3 platforms. (B) Left: The percentage of concordance between the Illumina SNP array and BGISEQ-500 for all SNPs compared to the percentage concordance of only SNPs with non-reference alleles in the Illumina SNP array for the 86 samples genotyped on both platforms. Right: Concordance of SNP genotypes between BGISEQ-500 and Illumina SNP Array for 1 sample (EG600036). (C) Overlap within the numbers of the 3 major structural variants detected in 1 sample using the 2 whole-genome sequencing datasets. (D) Overlap within the numbers of the 3 major mobile element insertions detected in 1 sample using the 2 whole-genome sequencing datasets.

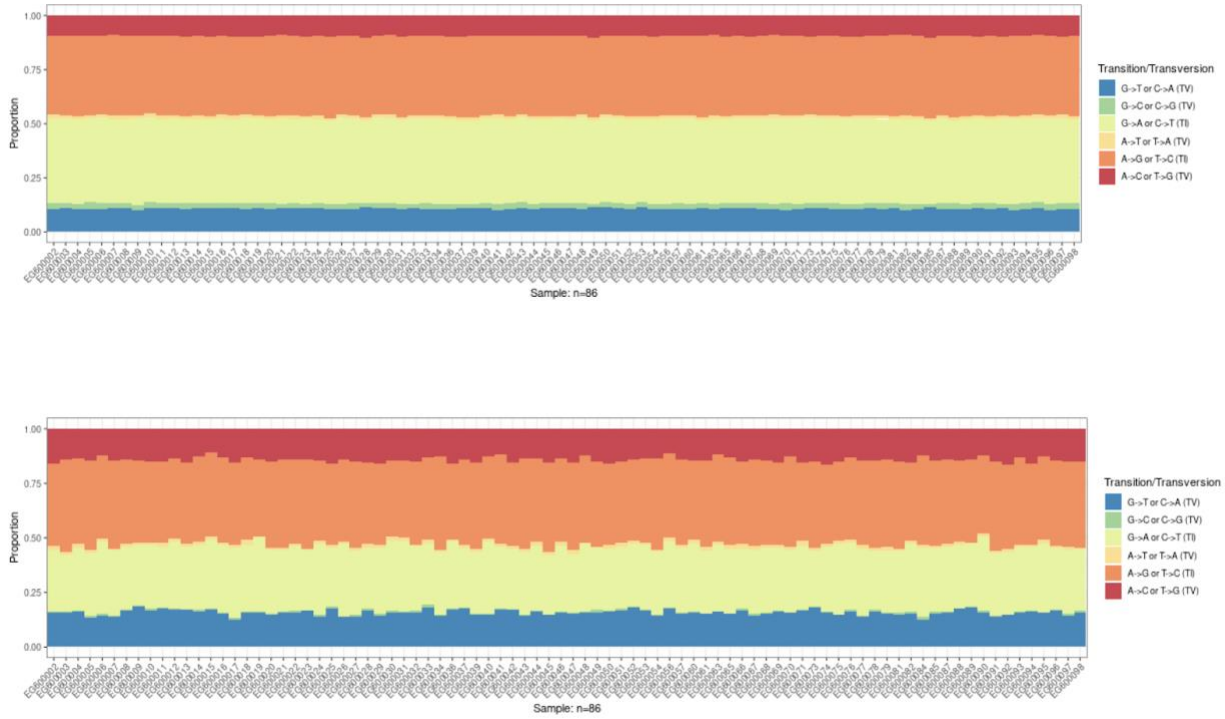

Figure 2-2 Transition/Transversion ratio (or TITV ratio) for the novel SNPs

(estimated with TiTvtools and visualized by plotTiTv) (top) for the SNPs where Illumina SNP array identified more alternate haplotypes than BGI (top right triangle in Figure 1C) and (bottom) for the SNPs where BGISeq identified more alternate haplotypes than Illumina SNP Array (bottom left triangle on Figure 1C table).

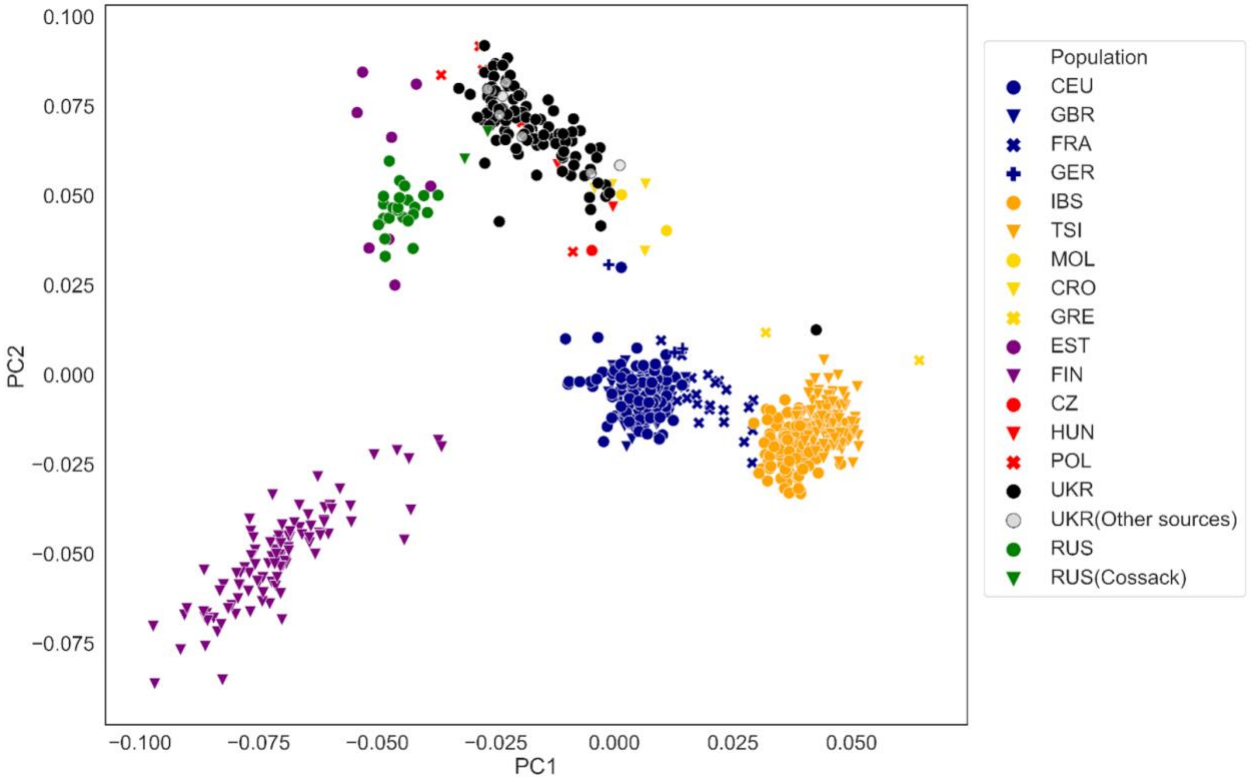

Figure 2-3 Principal component (PC) analysis of genetic merged dataset, containing European populations.

Colors reflect prior population assignments from the European samples from the 1KG (Utah residents [CEU] with Northern and Western European ancestry, Toscani in Italy [TSI], Finnish in Finland [FIN], British in England and Scotland [GBR], Iberian population in Spain [IBS])(24,135) and French (FRA) and Russians (RUS) from HGDP (RUS)(136), as well as the relevant high-coverage human genomes Croatian (CRO), Czech (CZ), Estonian (EST), German (GER), Greek (GRE), Hungarian (HUN), Moldovan (MOL), Polish (POL), Russian Cossack (RUS), and Ukrainian (UKR) from the Estonian Biocentre Human Genome Diversity Panel (EGDP) (139)as well as Simons Genome Diversity Project (140). The analysis was performed with Eigensoft (158).

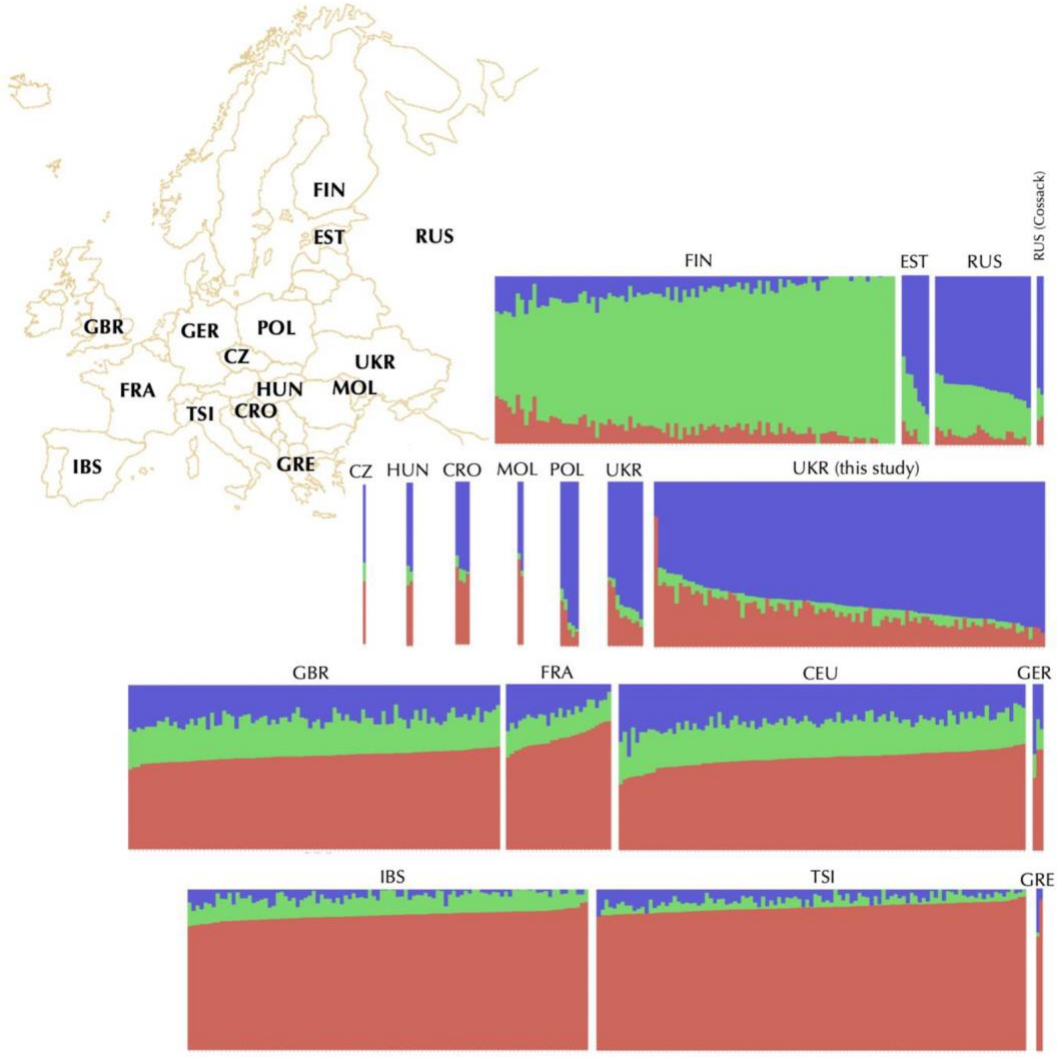

Figure 2-4 Genetic structure of Ukrainian population in comparison to other European populations.

Ukrainian (UKR) from the Estonian Biocentre Human Genome Diversity Panel (EGDP) (139)as well as Simons Genome Diversity Project (140). For identification of the optimal K parameter, we evaluated a range from 2 to 8, with K = 3 resulting in the lowest error. Structure plot constructed using ADMIXTURE package(144) at K = 3 illustrates similarity and differences between genomes from this study as well as samples from the 1KG (Utah residents [CEU] with Northern and Western European ancestry, Toscani in Italy [TSI], Finnish in Finland [FIN], British in England and Scotland [GBR], and Iberian population in Spain [IBS]) (24,135)and French (FRA) and Russians (RUS) from HGDP (136), as well as the relevant high-coverage human genomes Croatian (CRO), Czech (CZ), Estonian (EST), German (GER), Greek (GRE), Hungarian (HUN), Moldovan (MOL), Polish (POL), Russian Cossack (RUS), and Ukrainian (UKR) from the Estonian Biocentre Human Genome Diversity Panel (EGDP) as well as Simons Genome Diversity Project. For identification of the optimal K parameter, we evaluated a range from 2 to 8, with K = 3 resulting in the lowest error.

Table 2-1 The list of the samples in this study, their characteristics and geographical locations, and sources of genomic data for each

(BGISEQ-500 sequencing [BGI Inc., Shenzhen, China], Illumina Global Screening Array genotyping, and Illumina NovaSeq sequencing array [Illumina Inc., San Diego, USA]).

EG6000#	Sex	BGISeq 500	Illumina Global Array	Illumina HiSeq	Latitude and Longitude	Geographical coordinates	Sigle source origin local ancestry	History of diabetes	History of hypertension	Family History of Diabetes (in parents)	Family History of Hypertension in parents	History of other diseases	Multiple sources of local ancestry - parents from different regions	Other self-reported family history
31	F	yes	yes	no	45°57'26.8"N 33°47'52.4"E	45.957438, 33.797891	no	No	No	Yes	No			
13	F	yes	yes	no	46°45'09.7"N 36°50'16.0"E	46.752702, 36.837771	no	No	No	No	No	Allergy		
1	F	yes	no	no	48°00'40.2"N 24°06'51.9"E	48.011173, 24.114424	yes	No	No	Yes	Yes			
29	F	yes	yes	no	48°03'37.0"N 37°57'37.6"E	48.060289, 37.960431	no	No	No	Yes	No			
11	F	yes	yes	no	48°04'09.8"N 23°44'17.0"E	48.069389, 23.738057	yes	Yes	No	No	Yes	Vertebral disc hernia		
35	M	yes	yes	no	48°09'09.9"N 23°01'53.7"E	48.152747, 23.031586	no	No	No	Yes	No			
27	F	yes	yes	no	48°09'22.3"N 23°08'05.9"E	48.156197, 23.134983	no	No	Yes	No	Yes			
4	F	yes	yes	no	48°32'35.2"N 22°59'27.2"E	48.186729, 23.572066	no	No	No	No	No			
26	M	yes	yes	no	48°18'48.2"N 38°04'36.8"E	48.313385, 38.076901	no	No	No	No	Yes			
2	F	yes	yes	no	48°20'54.3"N 22°54'07.1"E	48.353582, 22.901122	no	No	No	No	No			
19	M	yes	yes	no	48°23'09.7"N 25°55'54.6"E	48.386021, 25.931838	no	Yes	Yes	No	Yes	Allergy	mixed	

7	F	yes	yes	no	48°23'29.1"N 22°54'09.2"E	48.391403, 22.902553	yes	Yes	Yes	No	Yes			
32	F	yes	yes	no	48°25'35.0"N 23°42'03.3"E	48.426380, 23.700904	yes	No	No	No	Yes			
88	M	yes	yes	no	48°26'25.2"N 22°43'17.3"E	48.440321, 22.721470	no	No	No	Yes	Yes		mixed	
72	F	yes	no	no	48°28'21.4"N 35°00'32.9"E	48.472600, 35.009137	no	No	No	Yes	Yes			
65	F	yes	yes	no	48°28'24.8"N 35°00'34.8"E	48.473548, 35.009659	no	Yes	Yes	Yes	Yes			
73	F	yes	yes	no	48°28'27.1"N 35°00'25.5"E	48.474200, 35.007077	no	No	No	No	No			
66	F	yes	yes	no	48°28'29.6"N 35°00'24.9"E	48.474874, 35.006919	no	No	No	Yes	Yes			
67	M	yes	yes	no	48.475498"N, 34.920716"E	48.475498, 34.920716	no	No	No	Yes	Yes			
68	F	yes	yes	no	48°28'47.0"N 34°54'55.9"E	48.479730, 34.915523	no	No	Yes	Yes	Yes		mixed	
69	F	yes	yes	no	48°30'42.0"N 35°03'39.7"E	48.511667, 35.061040	no	No	No	Yes	Yes			
70	F	yes	yes	no	48°30'48.6"N 35°03'32.2"E	48.513494, 35.058948	no	No	Yes	No	Yes	Polinosis		
92	M	yes	yes	no	48°30'55.9"N 32°14'56.9"E	48.515526, 32.249134	no	No	No	No	No		mixed	
74	F	yes	yes	no	48°32'28.1"N 34°51'25.8"E	48.541139, 34.857161	no	No	Yes	No	Yes			
3	F	yes	yes	no	48°32'35.2"N 22°59'27.2"E	48.541686, 22.988396	yes	No	No	No	No			
5	F	yes	yes	no	48°33'36.3"N 22°28'45.7"E	48.560091, 22.479349	yes	No	No	Yes	No			
93	M	yes	yes	no	48°33'48.1"N 39°18'55.5"E	48.563362, 39.315403	no	No	No	No	No			
71	F	yes	yes	no	48°34'36.8"N 35°05'51.7"E	48.576882, 35.097688	no	Yes	Yes	Yes	Yes			
76	M	yes	yes	no	48°35'09.9"N 22°28'46.4"E	48.586074, 22.479541	no	No	No	No	No	uro-, cholelitiasis		
94	M	yes	yes	no	48°36'35.8"N 22°17'16.7"E	48.609951, 22.287960	no	No	No	Yes	Yes		mixed	
96	M	yes	yes	no	48°36'44.2"N 22°17'07.8"E	48.612281, 22.285509	no	No	No	Yes	No		mixed	

95	M	yes	yes	no	48°36'44.6"N 22°17'16.6"E	48.612377, 22.287939	no	No	No	Yes	Yes		mixed	
34	M	yes	yes	no	48°36'53.3"N 22°16'04.7"E	48.614810, 22.267959	no	No	Yes	No	No		mixed	
36	F	yes	yes	yes	48°37'28.8"N 22°17'40.5"E	48.624666, 22.294577	no	No	No	No	Yes	Varicosis	mixed	
98	M	yes	yes	no	48°38'02.9"N 22°16'17.7"E	48.634126, 22.271575	no	No	No	No	No	Testicular cancer	mixed	Varicosis
9	M	yes	yes	no	48°39'15.3"N 22°48'42.1"E	48.654261, 22.811705	yes	Yes	No	No	No	Psoriasis		
8	M	yes	yes	no	48°42'37.9"N 22°35'45.0"E	48.710527, 22.595823	yes	Yes	Yes	No	Yes			
97	F	yes	yes	no	48°44'15.3"N 22°28'19.5"E	48.737593, 22.472083	no	No	No	Yes	Yes			
22	F	yes	yes	no	48°46'52.6"N 31°39'10.2"E	48.781277, 31.652828	no	No	No	No	No			
87	M	yes	yes	no	48°47'20.2"N 30°01'38.5"E	48.788948, 30.027352	no	No	Yes	No	Yes			
25	M	yes	yes	no	48°51'04.9"N 37°35'06.9"E	48.851357, 37.585255	no	No	No	No	No	Chronic rhinitis	mixed	
16	F	yes	yes	no	48°51'55.6"N 22°26'40.4"E	48.865431, 22.444547	no	No	No	No	Yes		mixed	
18	F	yes	yes	no	48°52'39.5"N 23°04'23.7"E	48.877627, 23.073250	yes					No info		
20	M	yes	yes	no	48°53'21.1"N 22°27'18.1"E	48.889200, 22.455021	no	No	No	No	Yes			
15	F	yes	yes	no	49°04'56.7"N 33°25'53.9"E	49.082408, 33.431636	no	No	No	No	No			
6	F	yes	yes	no	49°06'30.6"N 23°38'29.0"E	49.108510, 23.641393	yes	No	Yes	No	Yes			
30	F	yes	yes	no	49°08'00.3"N 25°12'06.4"E	49.133426, 25.201764	no	No	No	No	Yes			
89	M	yes	yes	no	49°09'12.2"N 23°01'58.1"E	49.153394, 23.032793	no	No	No	No	No			
24	F	yes	yes	no	49°24'40.1"N 24°36'42.7"E	49.411124, 24.611856	no	No	No	Yes	Yes	Allergy medicaments		
90	M	yes	yes	no	49°28'23.5"N 24°07'45.4"E	49.473181, 24.129274	no	No	No	No	No	Arrythmia		
14	F	yes	yes	no	49°31'02.2"N 23°12'25.7"E	49.517270, 23.207126	no	No	No	No	No			

33	F	yes	yes	no	49°31'12.3"N 23°12'05.5"E	49.520072, 23.201532	no	No	No	No	Yes			
61	M	yes	yes	no	49°31'37.9"N 23°58'38.2"E	49.527191, 23.977288	no	Yes	No	No	Yes	Diabetes mellitus type 1		
12	M	yes	yes	no	49°31'42.6"N 24°46'34.3"E	49.528496, 24.776199	yes	No	No	Yes	No	Oncology		
91	M	yes	yes	no	49°48'01.4"N 30°07'06.8"E	49.800393, 30.118547	no	No	No	No	No		mixed	
62	M	yes	no	no	49°50'20.7"N 24°01'21.8"E	49.839069, 24.022715	no	Yes	No	Yes	Yes	Diabetes mellitus type 1		
63	F	yes	yes	no	49°50'20.7"N 24°01'21.8"E	49.839069, 24.022715	no	No	No	Yes	Yes		mixed	
64	F	yes	no	no	49°50'40.2"N 24°01'05.9"E	49.844511, 24.018296	no	No	No	No	Yes		mixed	
23	M	yes	yes	no	50°00'22.1"N 36°18'09.6"E	50.006127, 36.302679	no	No	No	No	No			
50	F	yes	yes	no	50°15'32.2"N 28°41'41.4"E	50.258950, 28.694822	no	No	No	No	No			
48	F	yes	yes	no	50°15'41.9"N 28°40'40.3"E	50.261637, 28.677854	no	No	No	No	Yes		mixed	
75	M	yes	yes	no	50°15'54.8"N 28°40'26.6"E	50.265216, 28.674057	no	No	Yes	No	Yes			
46	F	yes	yes	no	50°16'04.3"N 28°39'45.8"E	50.267850, 28.662730	no	No	No	No	Yes	Miopia		
47	F	yes	yes	no	50°16'08.5"N 28°39'40.6"E	50.269016, 28.661282	no	No	No	Yes	Yes	Allergy		
80	F	yes	no	no	50°18'18.5"N 34°53'42.5"E	50.305129, 34.895146	no	No	No	No	Yes	Allergy fur, dust		Thyroid cancer
49	F	yes	yes	no	50°20'10.0"N 28°46'04.5"E	50.336109, 28.767902	no	No	No	Yes	Yes			
77	M	yes	yes	no	50°27'31.7"N 34°17'29.4"E	50.458802, 34.291507	no	No	No	No	No			Colon cancer
83	F	yes	no	no	50°33'28.5"N 35°22'14.2"E	50.557910, 35.370609	no	No	No	No	Yes			Urolithiasis, vitiligo
42	M	yes	yes	no	50°37'06.4"N 26°14'14.6"E	50.618438, 26.237382	no	No	No	No	Yes			
43	F	yes	yes	no	50°37'16.3"N 26°13'54.1"E	50.621189, 26.231702	no	No	No	No	Yes			
44	F	yes	yes	no	50°37'23.6"N 26°14'03.9"E	50.623217, 26.234406	no	No	Yes	No	Yes			

40	F	yes	yes	no	50°47'18.8"N 27°12'23.1"E	50.788551, 27.206422	yes	Yes	Yes	Yes	No	DM type 2		
45	F	yes	yes	no	50°51'31.6"N 28°33'39.4"E	50.858774, 28.560945	yes	No	Yes	No	Yes	Allergy		
81	F	yes	yes	no	50°51'51.8"N 35°15'39.1"E	50.864400, 35.260856	no	No	No	Yes	Yes			Lung cancer
84	F	yes	yes	no	50°53'25.6"N 34°49'03.5"E	50.890445, 34.817631	no	No	Yes	Yes	Yes	Allergy		Cirhrosis
85	F	yes	yes	no	50°54'07.7"N 34°48'52.4"E	50.902139, 34.814560	no	No	No	No	Yes			
78	M	yes	yes	no	50°54'14.1"N 34°38'44.8"E	50.903916, 34.645784	no	No	No	No	Yes			
86	F	yes	no	no	50°54'27.2"N 34°47'04.9"E	50.907566, 34.784693	no	No	No	No	No			Stomach cancer
79	F	yes	yes	no	50°54'31.9"N 34°48'19.1"E	50.908857, 34.805298	no	No	No	No	No			Cancer of mammary gland, cancer of ovaries
82	F	yes	yes	no	50°55'50.6"N 34°46'16.1"E	50.930733, 34.771132	no	No	Yes	No	Yes	Allergy pollen		Miocardial infarction
38	M	yes	no	no	50°58'29.8"N 26°43'23.7"E	50.974947, 26.723256	no	Yes	No	No	No	DM type 1		
37	M	yes	yes	no	50°58'33.6"N 26°42'55.8"E	50.975990, 26.715494	no	Yes	No	Yes	Yes	DM type 1		
53	F	yes	yes	no	50°59'24.0"N 31°08'13.3"E	50.989994, 31.137039	yes				Yes			
21	M	yes	yes	no	51°12'11.6"N 24°43'00.8"E	51.203209, 24.716882	no	No	No	No	No			
28	F	yes	yes	no	51°19'42.9"N 28°48'44.6"E	51.328593, 28.812374	no	No	No	No	Yes			
39	M	yes	yes	no	51°20'24.2"N 26°36'39.7"E	51.340042, 26.611015	no	Yes	Yes	No	Yes	DM type 2		
41	M	yes	yes	no	51°23'29.2"N 26°24'03.3"E	51.391456, 26.400927	yes	Yes	Yes	No	Yes	Asthma		
17	F	yes	yes	no	51°25'34.6"N 26°08'16.1"E	51.426284, 26.137811	yes	No	No	No	No			
60	F	yes	yes	no	51°26'32.4"N 31°41'16.9"E	51.442333, 31.688016	no	Yes	Yes			Allergy		
58	F	no	no	no	51°27'36.8"N 31°33'48.0"E	51.460212, 31.563338	no	Yes	Yes	Yes	Yes			

52	F	yes	yes	no	51°27'48.2"N 30°55'41.5"E	51.463387, 30.928186	no	Yes	Yes	No	Yes			
51	F	yes	yes	no	51°30'06.2"N 31°17'24.7"E	51.501728, 31.290201	no	No	No	Yes	Yes		mixed	
56	F	yes	yes	no	51°30'23.9"N 31°16'14.3"E	51.506650, 31.270636	no			Yes	Yes			
59	F	yes	no	no	51°33'40.0"N 31°09'07.8"E	51.561099, 31.152163	no	No	Yes	No	Yes	Hypothyroidism		
55	F	yes	no	no	51°40'27.4"N 31°08'29.3"E	51.674263, 31.141482	yes							
57	F	yes	yes	no	51°48'10.4"N 31°05'13.3"E	51.802883, 31.087032	yes	No	Yes	Yes	No			
10	F	yes	yes	no										
54	F	yes	yes	no	51°22'39.9"N 30°51'16.6"E		yes							

Table 2-2 . Sequencing summary of output from DNBSEQ-G50 and Illumina NovaSeq6000.

₴ Sequencing of 97 samples were attempted on DNBSeq-G50 at BGI sequencing facility (BGI Shenzhen, CHINA), and all 97 were successful.

¥ One sample (EG600036) was sent to Illumina NovaSeq6000 S4 at Psomagen Inc. (Gaithersburg, MD, USA). In addition, 96 samples were genotyped using Illumina Global Screening Array array (Illumina Inc., San Diego, USA), and 87 were successful (86 individual samples and 1 internal QC) remained after filtering.

	DNBSeq-G50 ₴	**Illumina NovaSeq6000** ¥
Samples sequenced	97	1
Read length (bp)	100	150
Reads above Q20 (>99% quality score)	97.85%	96.91 %
Total Reads	99,638,538,182	1,600,898,738
Average reads/sample	1,027,201,425	1,600,898,738
Average GC content	42.05%	41.07

Table 2-3 Filtering summary of the data obtained from 97 whole genomes sequenced with DNBSeq-G50.

₴ Sequencing of 97 samples were attempted on DNBSeq-G50 at BGI sequencing facility (BGI Shenzhen, CHINA), and all 97 were successful.

¥ One sample (EG600036) was sent to Illumina NovaSeq6000 S4 at Psomagen Inc. (Gaithersburg, MD, USA). In addition, 96 samples were genotyped using Illumina Global Screening Array array (Illumina Inc., San Diego, USA), and 87 were successful (86 individual samples and 1 internal QC) remained after filtering.

Sequencing results	**All samples**		
	Total Unique SNPs #	**Filtered Count**	**% Filtered[₴]**
Variation			
SNPs	14,738,063	1,727,084	11.7
Bi-allelic	14,254,070	1,586,787	11.1
Multi-allelic	483,993	140,297	29
Small Indels [¥]	2,808,384	80,780	2.9
Deletions	1,864,698	57,959	3.1
Insertions	1,488,408	42,421	2.9
Structural Variants [$]			
Large Deletions	685,56	52,478	76.5
Large Duplications	3,374	52,478	45.3
Inversions	430	93	21.6
Mobile Element Insertions			
Alu	7550	1790	23.7
L1	3123	2672	85.6
SVA	222	122	55
NUMT	1169	455	38.9

Table 2-4 Summary of variation int he 97 whole genome sequences from Ukraine

[£] Defined as "percent not reported in gnomAD(1000Genomes)"
[¥] Small indels are insertions and deletions < 50bp called by GATK (146).
[$] Large deletions and duplications are those called by *lumpy* (80) which are > 50 bp.

Sequencing results	**All samples**			**On average**	
	Total Unique Variants #	**Novel gnomAD Count**	**% Novel gnomAD (1000Genomes)** [£]	**Average # /sample**	**Average # Novel /sample**
Total sequence reads	99.8 Bn	--	--	1.03 Bn	--
Mean coverage	97 samples at 30X each	--	--	30X	--
Variation					
SNPs	13,010,979	477,564	3.7%(12.6%)	3,488,083	0.1% (0.7%)
Bi-allelic	12,667,283	470,667	3.7%(12.7%)	3,340,557	0.3%(0.6%)
Multi-allelic	343,696	6,897	2.0%(7.4%)	146,340	0.8%(4.7%)
Small Indels [¥]	2,727,604	76,484	2.8%(7.4%)	917,731	0.3% (1.0%)
Deletions	1,805,739	55,599	3.1% (9.0%)	624,919	0.3% (2.4%)
Insertions	1,445,987	30,453	2.1%(6.7%)	571,461	0.2% (2.1%)
Structural Variants [$]					
Large Deletion	16,078	10,914	67.9(48.3%)	3,524	52.6%(19.1%)
Large Duplications	1,845	1,356	73.5%(42.3%)	562	89.4%(35.2%)
Inversions	337	314	93.2% (47.8%)	185	94.1%(48.6%)
Mobile Element Insertions					
Alu	2,316	1805	77.9%(38.1%)	473	68.1%(18.0%)
L1	451	289	64%(50.1%)	79	60.8%(27.8%)
SVA	100	75	75%(52.0%)	20	70%(50%)
NUMT	714	--	--	16	--

Table 2-5 Summary annotation of different genomic elements in the Ukrainian genomes annotated in BGISeq data from 97 Ukrainian samples

[e] Unique alleles represent mutations that were counted only once using the largest transcript, disregarding their frequency in the population

[ε] Some of the mutations listed in the can be classified in more than one category

A. Variants by Location	**# of unique alleles[e]**	**Total allele #**	**Average /sample**
Upstream	2,023,920	6,716,794	69,246
UTR 5 Prime	31,026	122,417	1,263
Exon	320,979	839,045	8,650
UTR 3 Prime	150,302	389,528	4,016
Downstream	2,036,111	6,591,978	67,959
Intergenic	9,844,120	9,844,120	101,486
Intron	9,297,384	42,268,211	435,755
Motif	58,164	58,164	600
B. Functional Variants by Type [ε]			
Splice site acceptor	1,105	3,844	40
Splice site donor	969	3,609	38
Splice site region	19,436	79,853	824
Transcription factor binding site (TFBS) ablation	2,229	2,229	23
Conservative in-frame indels	1544	2,475	26
Gene Fusion	98	1,482	16
Disruptive in-frame indels	978	4,093	43
Missense	61,181	169,454	1,747
Start lost	116	413	5
Stop gained	885	2,442	26
Stop loss	95	324	4
Synonymous	49,731	146,066	1,506
Protein folding	105,436	258,767	2,668

Table 2-6 . Medically-relevant variants in the Ukrainian population included in GWAS and ClinVar databases

[a] Unique variants represent substitutions that were counted only once, disregarding their frequency in the population

Source of Annotation	**# Unique substitutions**[a]	**Total allele #**	**Average /sample**
GWAS catalog	102,551	6,479,953	66,804
ClinVar: pathogenic (or likely pathogenic**)**	189	1,830	19
ClinVar: benign (or likely benign**)**	43,892	1,842,668	18,997
ClinVar: protective (or likely protective**)**	20	1,209	12

Table 2-7 Examples of the functional SNPs with highly differentiating functional markers reported in ClinVar with high differences in the Ukrainian population compared to other neighboring European populations

SNP	**Chr**	**Gene**	**REF/ alt**	**Associated medical condition**	**Non-reference allele frequency**			**Fisher exact test *P*-value**	
					UKR	EUR	RUS	vs EUR	vs RUS
rs2297480	1	*FDPS*	T/G	Efficacy of the bisphosphonate response	0.13	0.26	0.27	0.038	>0.001
rs35329108	5	*SLC6A19*	G/A	Hyperglycinuria iminoglycinuria	0.32	0.23	0.17	0.049	0.004
rs7794745	7	*CNTNAP2*	A/T	Autism	0.48	0.38	0.30	0.032	0.010
rs10151259	14	*RPGRIP1*	G/T	Leber congenital amaurosis, cone-rod dystrophy	0.32	0.24	0.11	0.003	0.014
rs1801320	15	*RAD51*	G/C	Breast cancer susceptibility in *BRCA1* and *BRCA2* carriers	0.19	0.08	0.07	0.047	0

Table 2-8 Examples of the functional markers with the highest non-reference allele frequency differences in the Ukrainian population

Evaluated by the Fisher exact test compared to the frequencies in the neighboring populations: the combined population from Europe (EUR) and Russians from HGDP (RUS)

SNP	**Chr**	**Gene**	**Ref /Alt**	**Function**	**Non-reference allele frequency**			**Fisher exact test *P*-value**	
					UKR	**EUR**	**RUS**	**vs CEU**	**vs RUS**
rs72625995	17	*POM121L8P*	C/T	Exonic, nonsynonymous SNV	0.03	0.62	0.75	2.50E−07	1.86E−06
rs9930886	16	*PTPRN2*	A/G	Exonic, synonymous SNV	0.01	0.33	0.35	2.56E−07	2.19E−06
rs4779816	15	*ZBTB9; BAK1*	A/G	Exonic, nonsynonymous SNV	0.41	0.80	0.83	3.29E−06	7.82E−07
rs585800222	12	*ABCC1*	G/A	Exonic, synonymous SNV	0.03	0.13	0.26	3.06E−04	1.17E−02
rs80150964	11	*SMIM40; KIFC1*	T/C	Exonic, non-synonymous SNV	0.03	0.23	0.19	4.95E−04	

Chapter 3 - Assessing Repetitive Variation in the Genome Through Multi-Platform Discovery

The work presented in this chapter of the dissertation is a working draft of a manuscript in preparation with Dr. Ryan E. Mills. Dr. Mills and I devised the context and scope of the analysis. Dr. Mills provided guidance for the analysis, and I led the implementation of the computational methods and comparison analyses.

3.1 ABSTRACT

Expansions of tandem repeats are known to cause disease but can be difficult to characterize as these repeat expansions are highly variable across individuals. The ability to accurately characterize the length of tandem repeats in the genome is important for the identification of repeat expansions associated with disease, however, due to the repetitive nature of this class, variation has been challenging to detect genome wide with available sequencing technology. Here we perform a systematic analysis of available tandem repeat detection methods for both short and long read sequencing data on a single sample. We find that reference-based methods closely match for both short and long read methods while *de novo* methods differ from each other in both the characteristics of the repeat motifs they interrogate as well as the repeat lengths they predict. Finally, we provide guidelines for which methods are appropriate for different study goals.

3.2 BACKGROUND

Tandem repeats make up approximately 3% of the human genome (4), and expansions of tandem repeats are currently known to be associated with over 50 human diseases, including Huntington's Disease, Fragile X Syndrome, and ALS (159–162). A tandem repeat occurs when a sequence motif of two or more DNA base pairs (bp) is repeated with the repetitions directly adjacent to each other in the genome. These can be categorized into two classes: short tandem repeats (STRs) and variable number tandem repeats (VNTRs) with 2-6 base pair and 7 or more base pair repeat motifs respectively. Expansions of tandem repeats have historically been difficult to assess in the genome given their repetitive nature and the technology available to characterize them (159,163).

The introduction of whole genome sequencing (WGS) has greatly accelerated the discovery of pathogenic repeat expansions with half of known disease associations discovered in the last ten years (159). Prior to WGS, most repeat expansion disorders were discovered using large familial studies (164–166), but these lack the ability to identify rare repeat expansions responsible for disease or to untangle the relative contribution of repeat expansions in complex diseases. Whole genome sequencing allows for the identification of repeat expansions genome wide which has been shown to be important in uncovering additional repeat expansion disorders (92,167,168).

Several bioinformatics approaches have been developed to characterize tandem repeats from whole genome sequencing datasets (89–93,169,170). These can be split into two groups: reference-based and *de novo*. Reference-based detection methods require a list of coordinates in order to characterize tandem repeats. This can include a targeted list of known pathogenic repeat expansions or can include a list of coordinates of all tandem repeats annotated in the reference genome for whole genome analysis. In contrast, *de novo* detection methods aim to determine

repetitive regions from the read sequences for the samples being analyzed and do not require a list of known locations. *De novo* approaches allow for characterization of tandem repeats in a sample that may not be present or annotated in the reference genome, however, read lengths and the algorithms used to detect these regions have influence on the length of repeats and the repeat motifs which can be detected.

Short read WGS enables the characterization of many tandem repeats at once, however, the short length of the reads present challenges for repeat detection. Reads largely made up of repetitive sequence are difficult to align to the reference genome and often map to multiple locations, making it difficult to discern what region of the genome they originated from. For this reason, repetitive reads are often left unmapped. However, several methods have been developed to characterize tandem repeats from short reads. The first set of genome wide tandem repeat profiling methods focused on characterizing repeats fully contained within a single short read (171,172) while subsequent methods used properties of the pair-end reads to include estimates of repeats larger than a single short read. These latter methods include referenced-based methods GangSTR (89), ExpansionHunter (90), and *de novo* method STRling (169). Despite the utility of these bioinformatic approaches, they remain somewhat limited in their ability to predict tandem repeats by the intrinsic nature of short read technology.

The introduction and adoption of long read sequencing is promising for the characterization of repetitive regions in the genome; however, long read sequencing technology has a relatively higher error rate that can complicate the task of mapping and identifying tandem repeat regions. Several approaches have been developed to identify and characterize tandem repeat regions from long,

error-prone reads – namely, reference-based method Tandem-genotypes (92), and *de novo* methods Tricolor (93), and STRaglr (91), perform genome wide characterization of tandem repeats. While these methods have the potential to improve upon tandem repeat detection from short reads, the high insertion and deletion rate of long read sequencing can make the accurate characterization of repeat lengths difficult. Additionally, the higher cost of long read sequencing may be cost prohibitive for some studies (72).

Here we perform an in-depth analysis of tandem repeat detection by applying multiple state-of-the-art methods suitable for whole genome characterization of tandem repeats to a single sample, NA12878, which has been sequenced using several different sequencing platforms (25,173,174). We use this analysis to determine comparability of methods run on the same sample and to identify how many novel repeats are detected using *de novo* approaches compared to reference-based approaches. Finally, we provide guidelines as to when each method or combination of methods is appropriate for different study goals.

3.3 RESULTS

To assess the current capabilities of genome-wide tandem repeat calling in the human genome, we ran six tandem repeat profiling methods on a single sample, NA12878. The six methods are outlined in Table 3-1. Below, we have organized the results in the following manner: first we analyze the tandem repeat characterization methods developed for short read data. Next, we perform the same analysis for long read methods. Finally, we group reference and *de novo* methods to directly compare those techniques across different read lengths.

Detecting Repeat Variation from Short Read Whole Genome Sequencing

We ran three methods on whole genome PCR free Illumina sequencing data (30x coverage) from sample NA12878 (25), to assess the ability of each method to characterize repeat expansions from short reads. These include two reference-based approaches, GangSTR (89) and ExpansionHunter (90), and one *de novo* approach, STRling (169).

Short-read tandem repeat characterization methods

GangSTR takes aligned sequences and a set of coordinates (indicating known repetitive regions in the refence genome) as input and outputs estimated diploid repeat lengths. To estimate repeat lengths, GangSTR characterizes properties from four classes of pair-end reads; 1) the number of repeat copies enclosed in reads which cover the entirety of the repetitive region, 2) the observed fragment length for read pairs where the two reads map on either side of the repetitive region, 3) the distance a non-repetitive mate read of a fully repetitive read maps from the repetitive region, and 4) the number of repeat copies in a read that partially covers the repetitive region. These properties are integrated into a unified model to estimate the maximum likelihood TR length for each repeat loci supplied in the repeat reference panel. Reference repeat panels are supplied for the human reference genome and include both STRs and VNTRs with repeat motifs up to 20bp (89).

ExpansionHunter (v3) (90) also takes aligned sequences as input and requires a list of known repetitive regions in the reference genome and outputs estimated diploid repeat lengths. The largest advancement ExpansionHunter makes compared to its predecessors is that it uses a sequence

graph-based approach in the detection of complex loci containing multiple repeat motifs in close proximity. Reads aligning to the repetitive loci in the reference genome are realigned to the graph-based model representing the locus structure which can include multiple repeat motifs. Identifying alignment paths through the graph structure are used to genotype the repeat loci. This graph structure allows for much more flexibility in the definition of an individual repeat locus and allows for mismatches. Because we are looking to characterize repetitive regions genome wide for this analysis, it was not feasible to curate individually tailored regular expressions for each loci known to be repetitive in the reference genome. Instead, each loci in the reference repeat panel was treated as a simple repeat. The same reference panel supplied by GangSTR was used to generate the input for ExpansionHunter and therefore also includes both STRs and VNTRs with up to 20bp repeat motifs. (89)

STRling (169) takes aligned sequences as input as well but does not require a list of known repetitive loci, allowing the detection of repeats expansions in regions which are not annotated in the reference genome. To identify repeat loci in a sample, STRling utilizes a kmer based approach to identify reads with substantial repetitive content which are either remapped using a well-mapped mate pair as an anchor or designated as an unplaced-pair. Clusters of anchored or soft-clipped reads are reported as putative repeat expansion loci with anchored pairs providing rough boundaries and split-reads providing base pair resolution. The allele length at the putative STR repeat expansion loci is then estimated using properties from three classes of reads aligning to the repetitive region: 1) the length of a repetitive region is observed directly from reads spanning the entirety of the repetitive region, 2) the length of alleles up to the median fragment length are estimated to be proportional to the number of anchored pairs, and 3) the length of alleles beyond

the median fragment length are estimated to be proportional to the number of unplaced pairs. Unlike the previous two approaches, STRling only characterizes STRs – repeat motifs 2-6bps in length – and does not characterize VNTRs, thus we cannot compare STRling to ExpansionHunter or GangSTR in VNTR identification. (169)

STRling estimates longer repeat loci with lower G:C content than reference-based methods in short read data

Reference based methods GangSTR and ExpansionHunter characterized several orders of magnitude more repeat loci (STRs) that differed from the reference genome compared to the *de novo* method STRling (Figure 3-1A). STRling identified repeat motifs of 2-6 base pairs. The majority of STRling results characterized repeat motifs with 2 or 4 base pairs. Congruent with the reference repeat panel, Expansion Hunter and GangSTR report repeats motifs of 2-20 base pairs with 97.5% of repeat motifs between 2 and 6 base-pairs (Figure 3-1B). The two reference-based methods estimate that the majority of repeat motifs contain either 50% G or C nucleotides or consistent entirely of A or T nucleotides (Figure 3-1C). In contrast, the majority of all repeat loci identified by STRling had repeat motifs with less than 50% G or C nucleotides. Indeed, very few repeat loci identified by any of the three short read methods characterized repeat loci that were made of primarily G or C nucleotides.

Expansion Hunter and GangSTR have similar distributions of estimated repeat lengths with most loci being less than 100bp – less than the length of a single read – but also reporting estimated lengths up to 4,075bp for Expansion Hunter and 4,110bp for GangSTR. In contrast, most repeats

characterized by *de novo* method STRling were estimated to be between 100 and 1000bp in length and report loci up to 86,965bp (Figure 3-1D).

ExpansionHunter shows better concordance with Sanger repeat estimates

While ExpansionHunter and GangSTR have abundant regions characterized by both methods, STRling has much less overlap with these approaches (Figure 3-2A). Of the results it reports, only 3% overlap with regions from the reference repeat panel. We observed the concordance for repeat loci where all three methods reported results (Figure 3-2B). Crucially, there are no loci where all three methods report the same results. For some repeat loci, the STRling repeat estimate is close to the estimate of the reference-based methods, but overall it appears to estimate longer repeats than the reference-based methods. STRling frequently estimates repeat lengths longer than the fragment length at these repeat loci while the reference-based methods do not.

We compared our results to a Sanger sequencing dataset (0.3x coverage) for the same sample, NA12878, that was generated as part of a previous study to explore structural variation in humans (173). Sanger sequencing has longer read lengths (~800bp) compared to Illumina (~150 bp) and has a very low base calling error rate, both of which are beneficial for accurately characterizing repetitive sequences, and thus serve as an appropriate resource against which we can benchmark our results. We used Tandem Repeat Finder (175) to identify repetitive sequences within the Sanger reads, which were then aligned to the reference in order to determine overlap with regions profiled by the short read methods. We identified all repeat regions characterized by a short read method that also had a sanger read fully spanning the repeat region.

For the regions where both ExpansionHunter and GangSTR report results, 83.3% of the predicted repeat lengths were the same between these two methods. We quantify the divergence, calculated as a similarity metric, between a pair of repeat count genotypes reported by two different methods (see Methods). A visualization of the similarity metric (see Methods) is shown in Figure 3-2C, comparing ExpansionHunter and GangSTR for regions where they reported differing repeat counts. The result of each method is similarly compared to the overlapping Sanger reads. Repeat regions are sorted based on the similarity metric between ExpansionHunter and GangSTR. Of the 9,056 regions with Sanger reads overlapping repeat loci where Expansion Hunter and GangSTR report different results, Expansion Hunter matches the Sanger results 494 times whereas GangSTR matches the Sanger results 272 times. When calculating the total number of repeat counts difference from the Sanger results over all 9,056 regions, on average, Expansion Hunter differs by 1.35 repeat counts and GangSTR differs by 1.71 repeat counts on average. As Sanger length estimates increase, both ExpansionHunter and GangSTR are quite correlated with one another and underestimate repeat length in comparison to Sanger (Figure 3-3). However, the same pattern does not hold true for STRling estimates which is unbiased when it comes to over or under estimating repeat length in comparison to Sanger (Figure 3-3). Overall, estimated repeat lengths reported by Expansion Hunter showed greater concordance with Sanger than those reported by GangSTR based on the similarity metric detailed in the methods. In particular, when ExpansionHunter and GangSTR significantly disagree on their estimated repeat lengths (bottom right of Figure 3-2C), ExpansionHunter estimates tend to be closer to Sanger reads (bottom left of Figure 3-2C).

Detecting Repeat Variation from Long Read Whole Genome Sequencing

To assess the ability to characterize repeat expansions from long reads, we applied three different methods. These include one reference-based approach, Tandem-genotypes and two *de novo* approaches, Tricolor and STRaglr (Table 1).

Long-read tandem repeat characterization methods

Tandem-genotypes requires that sequencing reads be aligned specifically using the alignment program LAST as input and also requires a list of coordinates specifying repetitive regions in the reference genome (92). The same reference repeat panel supplied to the reference-based short read methods was used with the application Tandem-genotypes. To estimate repeat length, Tandem-genotypes relies on LAST-split which divides each long read into one or more parts and identifies the most probable alignment of each part. From the alignment of the parts which overlap the repeat coordinates, Tandem-genotypes uses the difference in the number of base pairs in the reference and query sequence to estimate the size of an insertion (or deletion) in the repeat regions. Repeat length estimates are reported for each read overlapping repeat regions supplied in the reference panel (89).

Tricolor is a *de novo* method that requires that read alignments be haplotype resolved prior to running the program and does not require a list of repeat coordinates (93). Tricolor's first module, *sensor*, detects repetitive regions within each long read by identifying drops in Shannon entropy across 20bp non-overlapping windows and merging any repetitive regions within 100bp of one another. Tricolor's second module, *refer*, extracts and trims all reads completely overlapping each repetitive region and creates haplotype resolved consensus sequences. These consensus sequences are used to identify the predominate repeat motifs and estimate the number of copies present. By

default, Tricolor only looks for repeat motifs up to six base pairs and requires there be at least five perfect copies of the motif. Additionally, results are only output if the estimated repeat length exceeds 50bp and at least one of the haplotypes differs from the reference sequence.

Like Tricolor, STRaglr is a *de novo* method (91). STRaglr looks only to find repetitive expansions as opposed to characterizing all repeat loci to limit run-time. Repetitive regions are discovered by identifying large insertions (default > 100bp) via the CIGAR string or split reads which are then analyzed for repetitive content using Tandem Repeat Finder to identify both the repeat boundaries and repeat motif. 70% of the insertion sequence must be made up of a single repeat motif. Once repetitive regions have been identified, reads aligning between the repeat coordinates or within 80bps (to rescue potentially missed split reads) are designated as candidate reads. These are then analyzed by Tandem Repeat Finder to determine the number of times the matching repeat motif is present in each read. A gaussian mixed model is used to cluster the repeat counts determined from each read overlapping a repetitive region into two groups to return a diploid genotype for each region.

STRaglr characterizes longer repeat estimates that contain greater G:C content than Tricolor or Tandem-genotypes

Like the short read methods, the *de novo* approaches, Tricolor and STRaglr did not report as many results as the reference-based approach, Tandem-genotypes (Figure 3-4A). Consistent with the reference repeat panel, Tandem-genotypes reports repeat motifs of 2-20 base pairs with most being between 2-6 base pairs. Tricolor only reports repeat motifs between 2-6 base pairs while STRaglr between 2-50 base pairs (Figure 3-4B). The distribution of the percentage of the repeat motifs

made up of C or G nucleotides is very similar between the reference-based method Tandem genotypes and *de novo* method Tricolor. The most repeat motifs were made up of 50% G or C nucleotides followed closely by repeat motifs made entirely of A or T nucleotides (Figure 3-4C). In contrast, the majority of repeat events reported by STRaglr are made up of 50% or more CG content. This represents a key distinction between the two *de novo* methods. A majority of the estimated lengths of the repeats characterized by Tandem-genotypes and Tricolor are between 10 and 100 bps. Overall, STRaglr reports larger repeats than the other two methods with most repeats estimated to be between 100 and 10,000bps (Figure 3-4D).

Of the *de novo* methods, Tricolor had substantial overlap with the repeat reference panel but also contributes several novel regions (Figure 3-5A). STRaglr reported a mix of novel repeat loci and repeat loci included in the reference repeat panel. With Tricolor and Tandem-genotypes reporting the most results for the same loci, we wanted to determine if one approach was performing more accurately than the other using the Sanger reads again as a truth set. We show concordance for repeat loci where all three methods reported results (Figure 3-5B). Note, while STRaglr estimates are relatively close to the estimates of Tandem Genotypes and Tricolor for half of the repeat loci analyzed, STRaglr overestimates repeat length of the other half of repeat loci compared to Tandem Genotypes and Tricolor. Concordance for repeat loci where STRaglr and at least one of Tandem genotypes or Tricolor report results is shown in Appendix Figure B 1.

A visualization of the similarity metric (see Methods) is shown in Figure 3-5C, comparing Tandem Genotypes and Tricolor for regions where they reported differing repeat counts. The result of each method is similarly compared to the overlapping Sanger reads. Repeat regions are sorted based on

the similarity metric between Tandem Genotypes and Tricolor. Of the 9,934 regions characterized by both Tandem Genotypes and Tricolor that were also covered by Sanger reads, 1,061 Tricolor results matched Sanger repeat length estimates while 1,039 Tandem Genotypes repeat results matched Sanger repeat length estimates. When calculating the total number of repeat counts difference from the Sanger results over all 9,934 regions, on average, Tricolor differs from Sanger by 0.95 repeat counts and Tandem genotypes differs by 1.02 repeat counts on average. While not as striking as in the short read data, it appears that Tricolor results match more closely to Sanger estimates, especially when Tricolor and Tandem-genotypes significantly diverge from one another (bottom of heat map, Figure 3-5C). As Sanger length estimates increase, both Tricolor and Tandem Genotypes underestimate repeat length in comparison to Sanger (Figure 3-6). However, the same pattern does not hold true for STRaglr estimates, which tend to overestimate repeat length compared to Sanger across all Sanger estimated lengths (Figure 3-6).

Short-read reference-based methods provide similarly accurate repeat length estimations compared to long read referenced based methods.

The majority of coordinates supplied in the reference repeat panel are characterized by all three reference-based methods (Figure 3-7A). Interestingly, the long-read method, Tandem Genotypes reports the fewest unique repeats (Figure 3-7A). Unsurprisingly, all three reference-based methods report similar repeat motif lengths (Figure 3-7B) and G:C content of repeat motifs (Figure 3-7C). All three reference-based methods additionally report similar distributions of repeat estimate lengths, with most estimations falling between 10 and 100 base pairs. Regions that differed between the methods were compared to Sanger sequences to determine which methods were the most accurate (Figures 3C and 3D). Expansion Hunter matched better with repeats estimated from

Sanger reads compared to Tandem Genotypes, though Tandem Genotypes matched more closely with Sanger than GangSTR (Appendix Figure B 2).

de novo approaches provide diverse estimates of repeat length

While the reference-based approaches were all supplied the same set of reference coordinates to focus repeat characterization, each of the *de novo* methods applied different approaches to detect repetitive loci directly from the read sequences. Compared to the reference-based methods, we see much less overlap between the results of *de novo* approaches (Figure 3-7E). Tricolor identifies substantially more repetitive regions within the genome of NA12878. 11,477 novel repeat loci - repeat loci not included in the repeat reference panel - were characterized by *de novo* methods. 685 repetitive loci were identified by more than one method with only 59 identified by all three methods. Of the loci identified by all three methods, the length estimates for the two long read methods were closer to each other than either were to the short read method, STRling (Figure 3-7H). Two of these methods only detect STRs, while STRaglr is the only *de novo* method with the capacity to identify VNTRs that are not annotated in the reference genome (Figure 3-7F).

Performance on Known Pathogenic Expansion Loci

We looked at the tandem repeat count estimates across all six methods for a set of ten repeat loci known to be pathogenic when expanded (see Table 3-2). Six repeats showed exact concordance between all methods able to characterize them while three showed variation in the estimated repeat length across methods. The remaining repeat was only characterized by a single method (Figure 3-8).

Two *de novo* methods, STRaglr and STRling, were unable to characterize any of the ten pathogenic repeats, likely because they are not in an expanded form in this sample and therefore do not exceed the threshold for detection by these methods. STRaglr focuses on repetitive regions with insertions larger than 100 bps relative to the reference genome, which likely do not exist in this sample as it is not known to have any expansions at these loci. STRling does not report any results for these loci because they are likely too small to be detected. The largest repeat length based on the estimations of the other methods is 90 bp (CAG repeat on Chr6) which would not exceed the 80% threshold of a 150bp read that is required for a read to be deemed informative by STRling.

Nine of the ten pathogenic loci were included in the reference repeat panel and were characterized by all three reference-based methods except the GGGGCC repeat which Tandem-genotypes failed to categorize. Tricolor, a *de novo* method, was able to identify five of the nine regions included in the reference repeat panel. Additionally, it was able to characterize one repeat not identified by any of the reference-based methods, showing the utility of *de novo* methods in disease contexts.

3.4 DISCUSSION

Tandem repeats are an understudied class of variation in the genome known to cause disease when expanded. The advancements in whole genome sequencing technology have allowed genome wide characterization of tandem repeats, greatly improving the potential to detect novel expansions associated with disease. Here, we conducted an in-depth analysis of six bioinformatic methods for genome wide characterization of tandem repeats for the sample NA12878, three designed to be used on short-read data and three on long read data. Additionally, these methods include a mix of reference-based methods and *de novo* methods.

In characterizing three short read methods we found that two referenced-based methods, Expansion Hunter and GangSTR, reported results for many of the same loci. However, when their estimated repeat lengths were compared to overlapping Sanger reads from the same sample, Expansion Hunter outperformed GangSTR. While Expansion Hunter and GangSTR have the capacity to detect VNTRs, the vast majority of the repeat loci characterized by these methods had repeat motif lengths less than 6bp. *De novo* method STRling has little overlap with the reference-based methods and generally reported repeat loci with higher repeat length estimates compared to the other two short read detection methods. Interestingly, the repeat loci identified by STRling consisted primarily of repeat motifs with less than 50% G:C content.

Our analysis of three long read tandem repeat detection methods found significant overlap in repeat loci characterized by reference-based method Tandem Genotypes and *de novo* method Tricolor. Both methods had similar distributions of predicted repeat length and G:C content for the repeat loci characterized. In contrast, the *de novo* method STRaglr had little overlap with the other two long read methods and had the highest percentage of repeat motifs with greater than 50% GC content. STRaglr is also the only long-read method to characterize VNTRs.

When considering all three reference-based methods, regardless of read length, the characteristics of the repeat loci interrogated are very similar, given they were supplied the same set of repeat loci as input. Overall, Expansion Hunter using short reads provided repeat length estimates closest to long read method Tandem genotypes. The repeats identified by the *de novo* methods differed from one another in several ways. Tricolor had similar estimated lengths and G:C content to the

reference-based methods while STRling and STRaglr generally reported most repeat estimates between 100 and 1000bp in length. There was very little overlap in the repeat loci identified by the methods. Together they identified over 11,000 novel repeat loci that were not included in the repeat reference panel; however, only 685 were identified by more than one method leading to the conclusion that these may contain many false positives. Since two of the *de novo* based methods often predicted repeat lengths greater than 800 bps, Sanger sequencing is not the best method for validation and other techniques such as PCR may be more appropriate for the validation of these novel repeat expansions.

Sanger sequencing reads were used as a "truth set" in several analyses in this paper. There are several limitations of this approach. The Sanger sequencing dataset was low coverage and often only a single read spanned a repeat locus, which would enable the measurement of the concordance of only a single allele. Additionally, using Sanger as a "truth set" has limited utility to validate repeat expansions greater than 800bp – the average length of a Sanger read. Additional validation through PCR amplification or Southern blot would help elucidate the more accurate method for larger repeats, although this will have lower resolution. Additionally, TRF was used to determine the length of repeats in the Sanger reads. TRF has been shown to be conservative in identification of repeat loci in a sequence, specifically with very short motif lengths (176,177). While a low false positive rate is beneficial for validation, this may have limited our ability to use Sanger reads for validating STRs.

This analysis was also limited in that each method was only run using the default parameters on a single sample. Running each program multiple times while varying the input parameters may result

in improved (or worse) performance. Performing this analysis on additional samples including samples with varying sequencing coverage could also provide additional insight on when it is appropriate to use each method

From these results, we have developed a set of guidelines for researchers exploring STRs and VNTRs with either short-read or long-read datasets:

Short-read guidelines

While there are some unique repeats regions analyzed by ExpansionHunter and GangSTR, it likely isn't enough to warrant running both methods on a short read dataset. However, STRling provides an orthogonal set of repeats, specifically ones that are not annotated in the reference sequence. While the method is theoretically able to analyze regions that are annotated in the reference sequence, in practice, we do not see a lot of overlap between STRling results and those from reference-based methods. For this reason, we recommend running both ExpansionHunter and STRling on short read data when looking to characterize genome wide repetitive repeats. However, if only looking for large STRs and are not interested in small variations in repeat length across many regions, STRling has a much faster runtime than ExpansionHunter or GangSTR.

Long-read guidelines

Guidelines for analysis of long-read datasets is not as straightforward. STRaglr, one of the two *de novo* methods, characterizes longer repeat estimates that contain greater G:C content than Tricolor or Tandem-genotypes. Additionally, STRaglr can characterize VNTRs, unlike Tricolor or Tandem-genotypes. For this reason, we recommend running either Tricolor or Tandem-genotypes

together with STRaglr to provide a diverse set of characterized repeat expansions. Of the three methods, Tandem-genotypes provides the greatest number of unique repeat loci. That said, Tricolor offers utility in pathogenic studies, in that it can identify small novel repeat expansions, that can be characteristic of pre-mutations, that the other two *de novo* methods will miss.

3.5 METHODS

Data

We obtained the publicly available alignment of Illumina reads for NA12878 (25) to GRCh38. We obtained previously generated FASTQ files generated by PacBio sequencing of NA12878 (174). FASTA and qscore files were download for previously generated Sanger sequencing of NA12878 (173) and were converted in to FASTQ files using a custom python script.

Repeat reference panel

The reference repeat panel was downloaded from (https://github.com/gymreklab/GangSTR, version 13). In brief, this panel was created by running Tandem Repeat Finder on the human reference genome and several filters were used to refine the repeat set. Some filters which were used included removing repeats with repeat motifs greater than 20 base pairs, removing repeats with homopolymer runs, and repeats which could be fully represented as a single sub-motif. See (89) for more details.

Alignment

PacBio reads were aligned using minimap2 (77) (v2.17-r974-dirty; map-pb preset option) and LAST (v1256; as specified https://github.com/mcfrith/last-rna/blob/master/last-long-reads.md -

without repeat masking) (178). Sanger reads were aligned using minimap2 (77) (v2.17-r974-dirty; map-sr preset option). Illumina reads were already aligned when downloaded (25) (BWA-MEM) (179).

Tandem Repeat Characterization

Default parameters were used run the tandem repeat characterization methods unless otherwise stated.

GangSTR was run using the Illumina alignment and the repeat reference panel mentioned above. To run ExpansionHunter, a JSON input file was created starting from repeat reference panel. Each repeat was made into a JSON entry using the repeat motif as the locus structure in the format "([RU])*" and the variant type “Repeat”. Repeat loci within 1,000 bp of a gap in the reference were removed, as these would produce errors when running ExpansionHunter. Non-autosomal chromosomes were excluded from the analysis. Tricolor was run using the minimap alignment of PacBio reads. The minimum length of tandem repeat option was set to 10 (default is 30) to better match list of known repeat regions from the reference genome. Tandem-genotypes was run using the LAST alignment of PacBio reads and the repeat reference panel. STRaglr was run using the minimap alignment of PacBio reads. STRling was run using the short read alignment. Tandem repeat counts characterization for each sanger read aligning to a known repetitive region was performed using Tandem Repeat Finder and the recommended parameters `trf - 2 7 7 80 10 50 500 -h -ngs`.

Cross-method repeat region identification comparisons

The repetitive regions from the results of all methods were merged using the merge functionality of bedtools (180) to create a master list of all repetitive regions identified across methods. Repeats reported by reference-based methods were only included in the analysis if the estimated repeat length differed from the reference genome. The results from each method were intersected with the "master list" to determine which regions the method characterized (50bp were added on either side of STRling repeat regions as regions were often reported as a single base pair). Each region from the master list was only counted once, even if the method reported multiple repeats within the same region from the master list.

Cross-method repeat length concordance comparisons

Given the variety in the output format across the different methods, the following steps were performed to allow for the comparison among the methods:

(1) Identify matching repeat calls across methods – method results were considered the same repeat if they had a) overlapping coordinates (50bp were added on either side of STRling repeat regions as regions were often reported as a single base pair) b) matching repeat motifs (specified in the method output; all rotations of the repeat motif were allowed, i.e. AT and TA). If multiple results with the same repeat motif were within the repeat boundaries, the results were combined (Tricolor often had several smaller repeats whereas other methods would characterize it as a single larger repeat)

(2) Convert method outputs to a uniform format (repeat motif counts) – GangSTR, ExpansionHunter, and STRaglr all report results as absolute repeat motif counts and were used as is. Tricolor also reports absolute repeat counts but multiple repeat motifs can be given in

the same VCF entry – different repeat motifs were separated into different results or counts were summed if the same repeat motif was reported twice in one VCF entry. Tandem-genotypes and STRling reported repeat motif count relative to the reference sequence. These were converted to absolute repeat counts by adding or subtracting count change from the number of repeat motifs seen in the reference sequence $[ceil((stop - start)/repeat\ length)]$.

(3) Determine the genotype per sample – Several of the long-read callers report the repeat motif count per read as opposed to one count per haplotype. To estimate the genotype for the individual, we performed k-means clustering with k=2 to group the estimated repeat lengths. The mean value of the repeat lengths contained in each cluster were rounded to the nearest integer and were reported as the repeat loci genotype.

(4) Calculate a similarity metric - following was used as a measure for similarity between two methods results for a given loci. (GT_{A1}, GT_{A2}) is the genotype reported by method A and (GT_{B1}, GT_{B2}) is the genotype reported by method B:

$$\min(abs(GT_{A1} - GT_{B1}) + abs(GT_{A2} - GT_{B2}), abs(GT_{A1} - GT_{B2}) + abs(GT_{A2} - GT_{B1}))$$

3.6 FIGURES AND TABLES

Table 3-1 The reference-based and de novo bioinformatic methods for short-read and long-read technologies.

Method	Citation	Ref- Based	De novo	Pre-steps	Results Format
Short Reads					
GangSTR	Mousavi et al. (2019)	X			Absolute counts, per sample
ExpansionHunter	Dolzhenko et al (2019)	X		JSON ref repeat input	Absolute counts, per sample
STRling	Dashnow et al. (PrePrint)		X		Relative counts, per sample
Long Reads					
Tandem-genotypes	Mitsuhashi et al. (2019)	X		LAST alignment required	Relative counts, per read
Tricolor	Bolognini et al. (2020)		X	Haplotype resolved BAM	Absolute counts, per sample
STRaglr	Chiu et al. (2021)		X		Absolute counts, per read/sample
Sanger					
TRF	Benson (1999)			Reads overlapping repeats	per read

A

Method	Reported Results	Method Type
GangSTR	89,030	Reference-based
ExpansionHunter	93,130	Reference-based
STRling	955	De novo

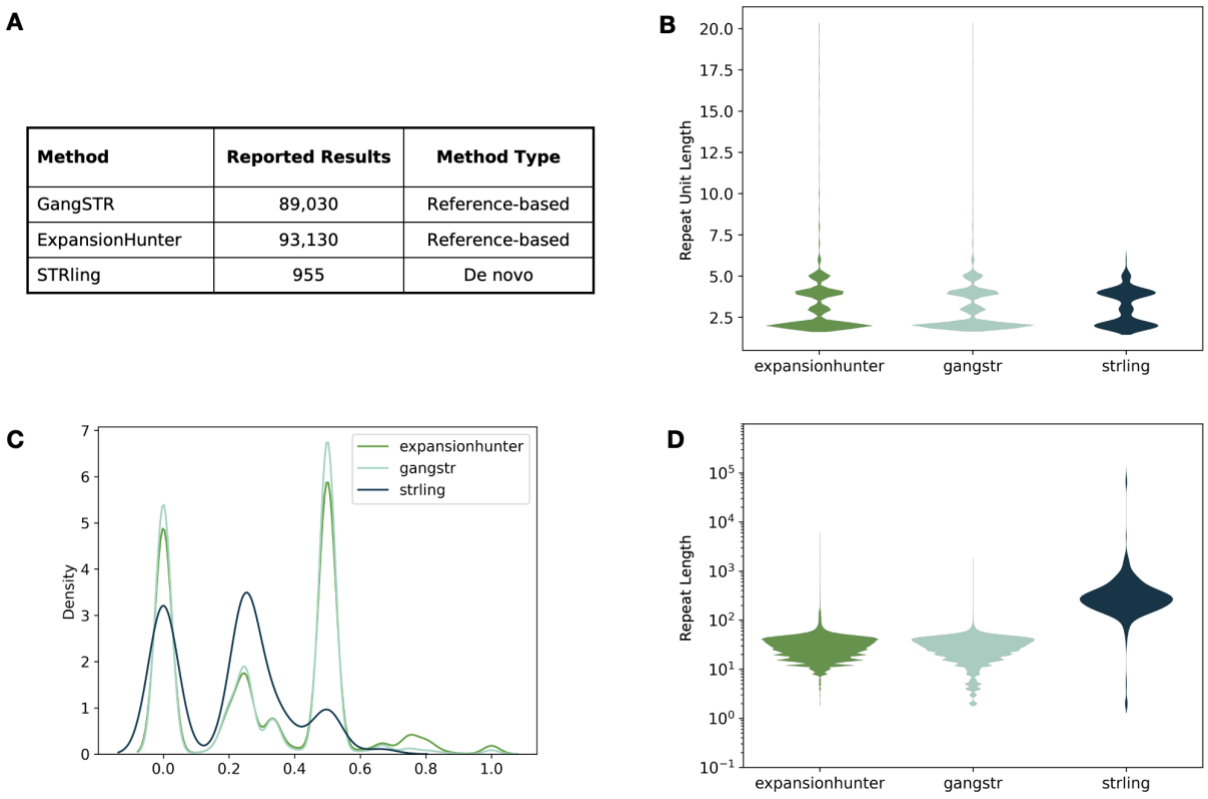

Figure 3-1 Comparison of methods using short read data.

A) Table reporting the number of repeats characterized by each method as well as the method type. B) Violin plot that displays the distribution of repeat motif lengths characterized by each method. C) Density plot showing the % of G:C content of repeat motifs characterized by each method. Repeat motifs containing only A or T nucleotides align with 0.0 along the x axis, motifs containing only G or C nucleotides align with 1.0. D) Violin plot displaying the distribution of estimated repeat lengths for each method.

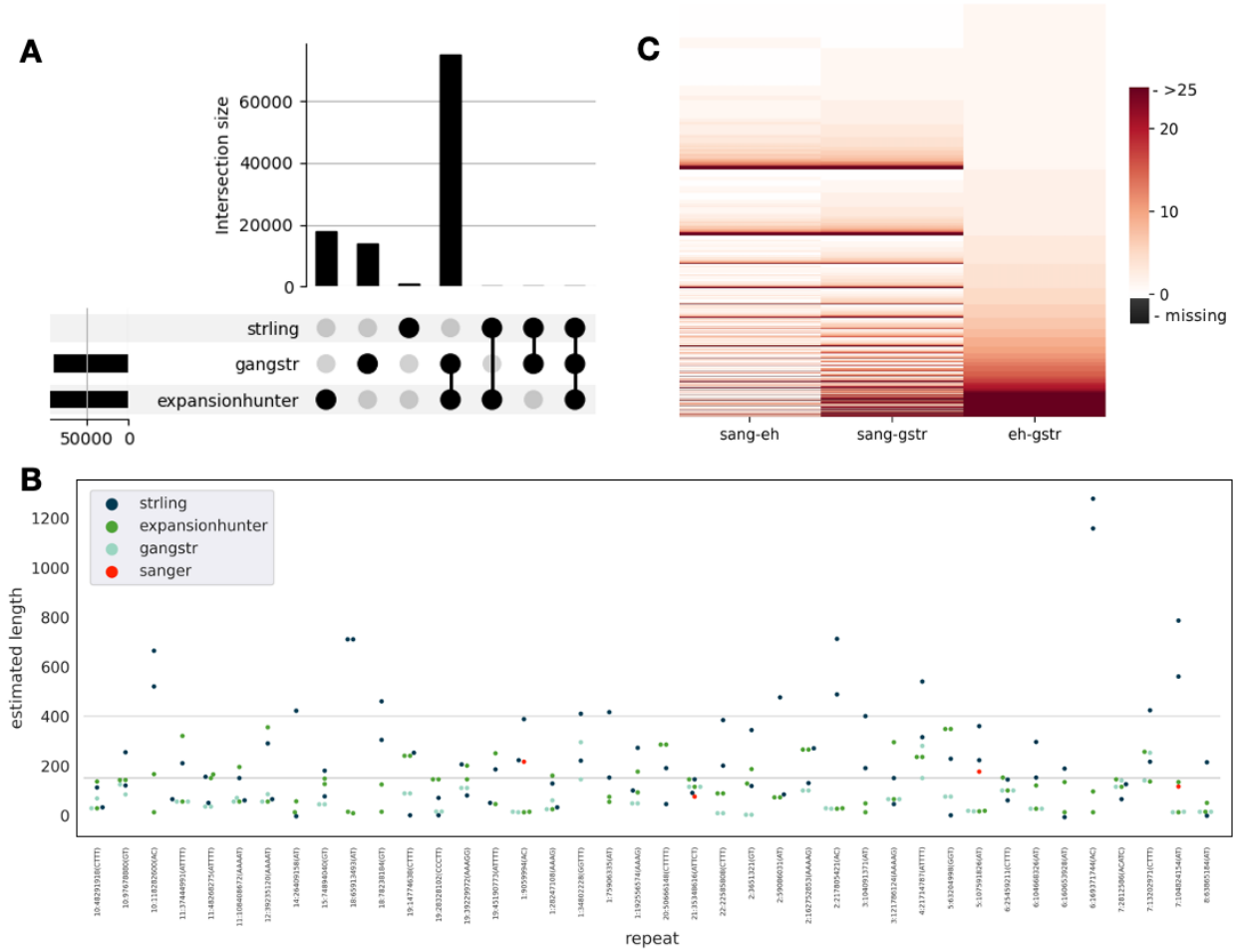

Figure 3-2 Repeat length concordance across short-read methods.

A) Upset plot of overlapping repeat loci characterized by each method. B) Dot plot of estimated repeat lengths for repeat loci characterized by STRling (dark blue) and reference based methods (ExpansionHunter-green, GangSTR – light blue). Sanger dots (red) are included for repeat loci spanned by Sanger reads. C) Heatmap of similarity metric comparisons between reference based methods and Sanger.

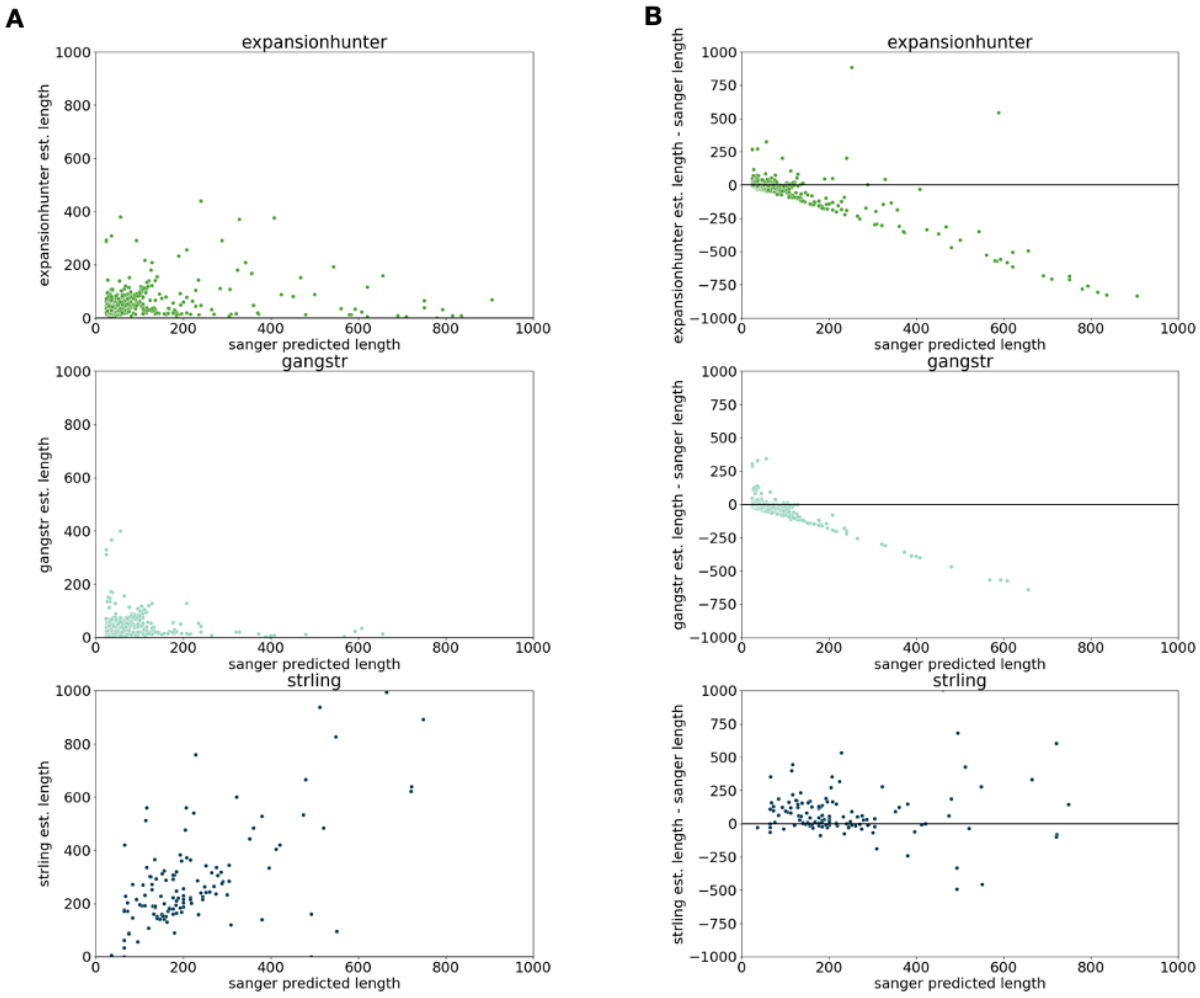

Figure 3-3 Comparisons between Sanger Sequencing and each short-read method.

A) Scatter plots displaying direct comparison of each methods repeat length estimate along the y axis and the Sanger repeat estimate along the x axis. B) Scatter plots displaying the difference in repeat estimates by the two methods along the y axis across the predicted Sanger lengths along the x axis. Each scatter plot represents a separate comparison between Sanger estimates and either ExpansionHunter, GangSTR or STRling estimates.

A

Method	Reported Results	Method Type
Tandem Genotypes	77,912	Reference-based
Tricolor	34,972	De novo
STRaglr	4,612	De novo

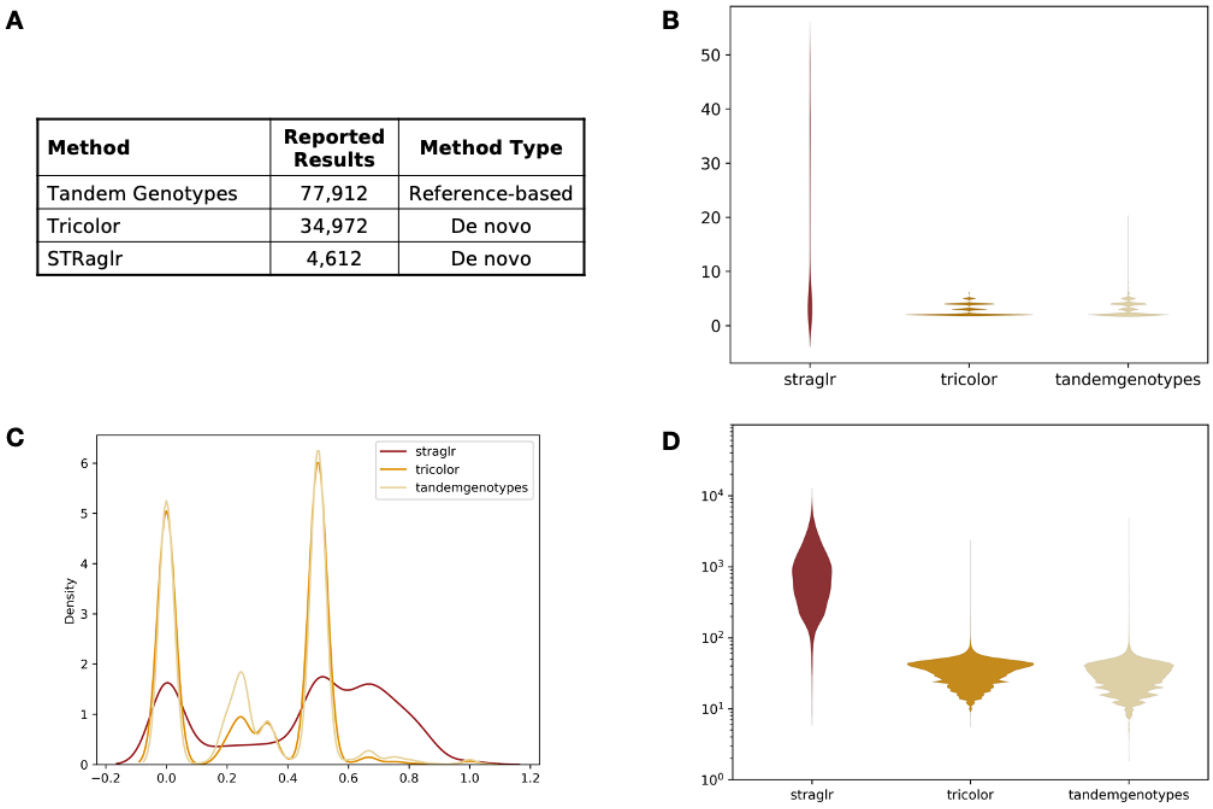

Figure 3-4 Comparison of methods using long read data.

A) Table reporting the number of repeats characterized by each method as well as the method type. B) Violin plot that displays the distribution of repeat motif lengths characterized by each method. C) Density plot showing the % of G:C content of repeat motifs characterized by each method. Repeat motifs containing only A or T nucleotides align with 0.0 along the x axis, motifs containing only G or C nucleotides align with 1.0. D) Violin plot displaying the distribution of estimated repeat lengths for each method.

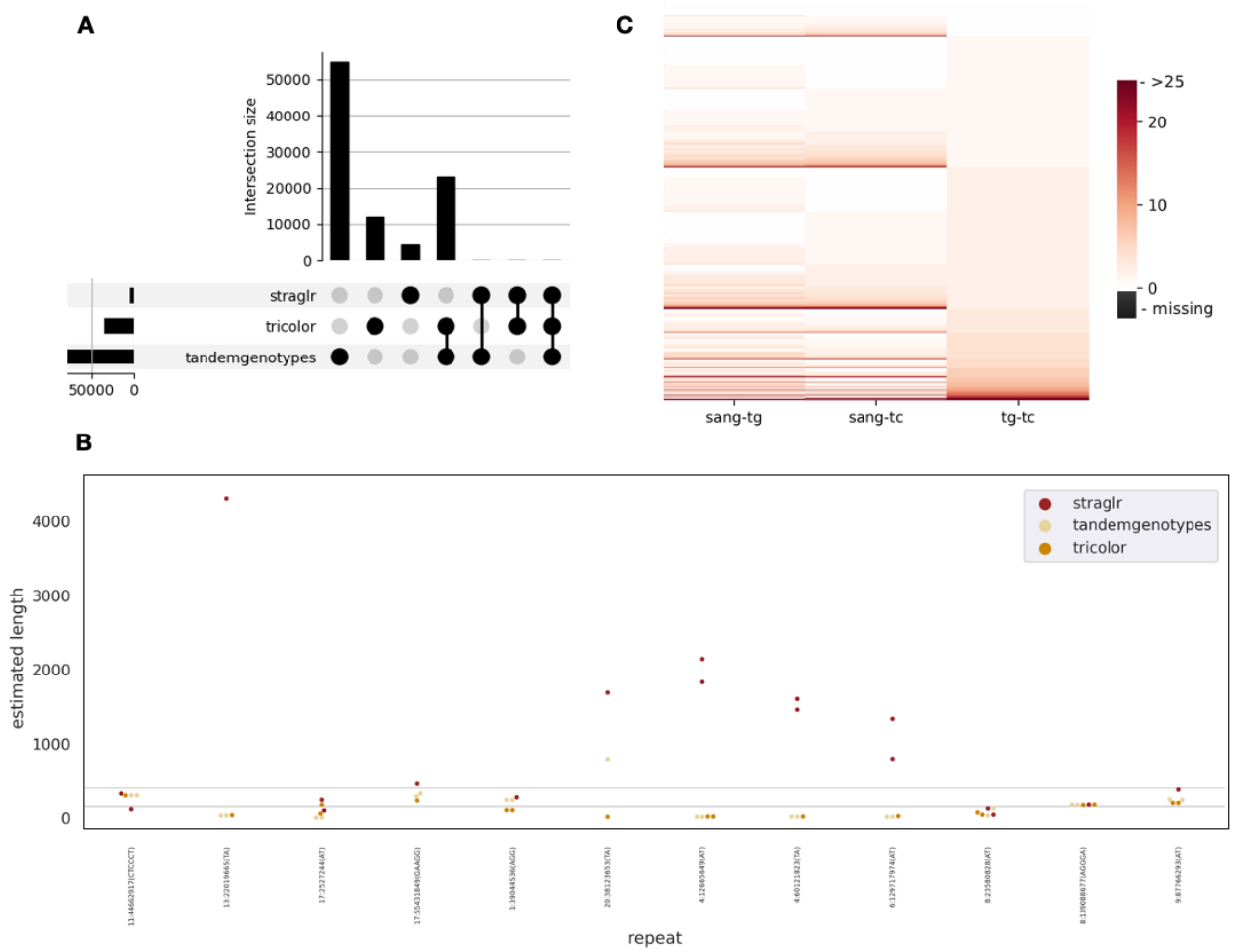

Figure 3-5 Repeat length concordance across long read methods.

A) Upset plot of overlapping repeat loci characterized by each method. B) Dot plot of estimated repeat lengths for repeat loci characterized by STRaglr (dark red) and Tandem Genotypes - yellow, Tricolor – orange). C) Heatmap of similarity metric comparisons between reference based methods and Sanger.

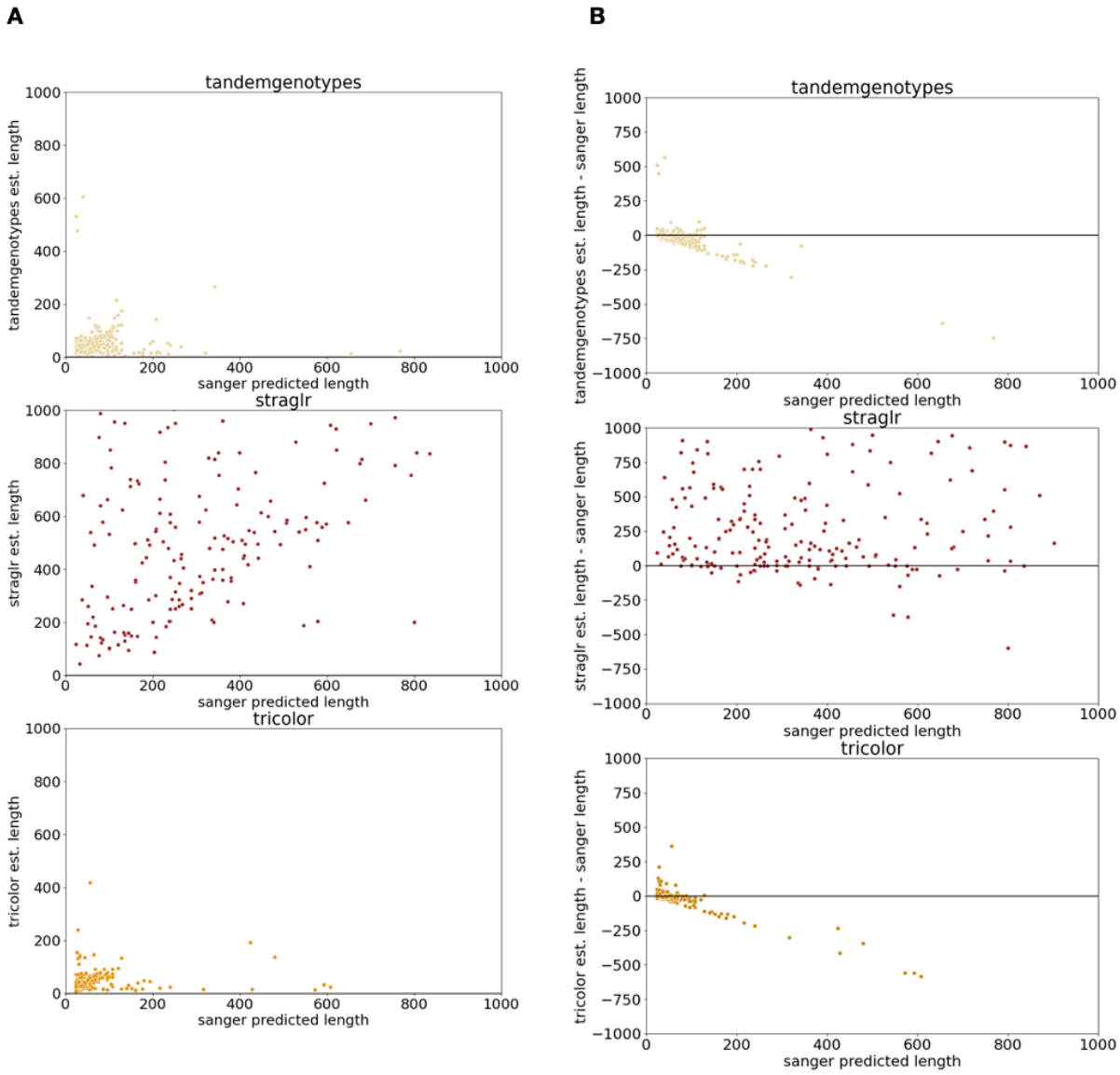

Figure 3-6 Comparisons between Sanger Sequencing and each long read method.

A) Scatter plots displaying direct comparison of each methods repeat length estimate along the y axis and the Sanger repeat estimate along the x axis. B) Scatter plots displaying the difference in repeat estimates by the two methods along the y axis across the predicted Sanger lengths along the x axis. Each scatter plot represents a separate comparison between Sanger estimates and either Tandem Genotypes, STRaglr, or Tricolor estimates.

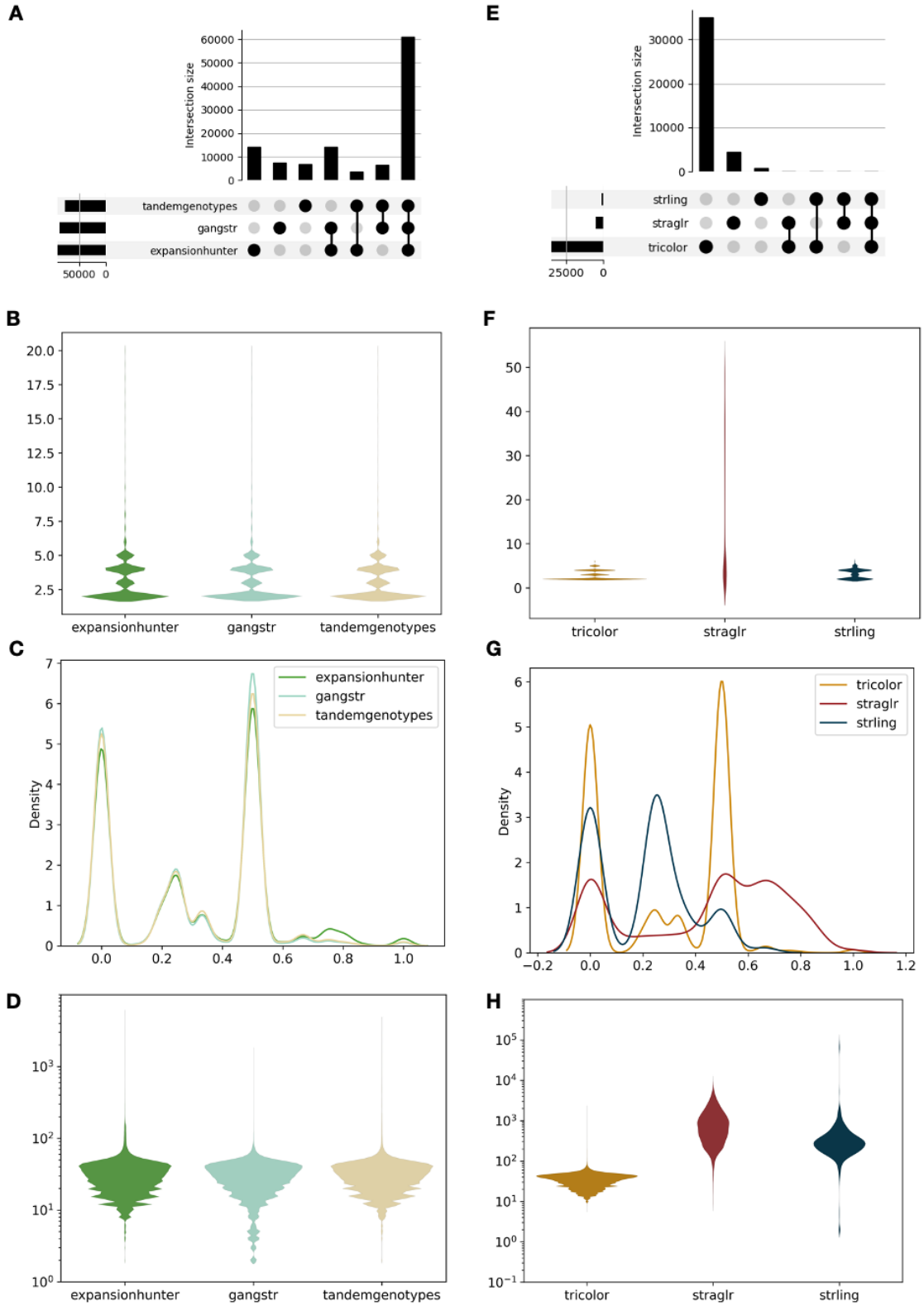
A
Intersection size
60000
50000
40000
30000
20000
10000
0
tandemgenotypes
gangstr
expansionhunter
50000 0
E
Intersection size
30000
20000
10000
0
strling
straglr
tricolor
25000 0
B
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
expansionhunter
gangstr
tandemgenotypes
F
50
40
30
20
10
0
tricolor
straglr
strling
C
Density
expansionhunter
gangstr
tandemgenotypes
0.0
0.2
0.4
0.6
0.8
1.0
G
Density
tricolor
straglr
strling
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
D
10^3
10^2
10^1
10^0
expansionhunter
gangstr
tandemgenotypes
H
10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
tricolor
straglr
strling

Figure 3-7 Characteristics of repeat loci detected by reference and *de novo* based methods.

(A-D correspond to reference based methods while E-H correspond to de novo methods) A,E) Upset plot of overlapping repeat loci characterized by each method. B,F) Violin plot that displays the distribution of repeat motif lengths characterized by each method. C,G) Density plot showing the % of G:C content of repeat motifs characterized by each method. Repeat motifs containing only A or T nucleotides align with 0.0 along the x axis, motifs containing only G or C nucleotides align with 1.0. ,HD) Violin plot displaying the distribution of estimated repeat lengths for each method

Table 3-2 Table of known pathogenic repeat expansions

Disease	**Gene**	**Repeat Motif**	**Location**	**Chr**	**GRC38 Co-ordinates**	**Citation**
Cerebellar ataxia, neuropathy, vestibular areflexia syndrome	RFC1	AAGGG (AAAAG)	Intron	4	39366381-39287456	(181)
Spinocerebellar ataxia type 1	ATXN1	CAG	CDS	6	16761490-16299112	(182)
Spinocerebellar ataxia type 2	ATXN2	CAG	CDS	12	111599673-111452214	(183)
Spinocerebellar ataxia type 3	ATXN3	CAG	CDS	14	92106582-92058552	(184)
Spinocerebellar ataxia type 6	CACNA1A	CAG	CDS	19	13506479-13206442	(185)
Huntington's disease	HTT	CAG	CDS	4	3074681-3243960	(186)
Myotonic dystrophy type I	DMPK	CTG	3' UTR	19	45782490-45769709	(187)
Myotonic dystrophy type 2	CNBP	CCTG	Intron	3	129183896-129167827	(188)
Fridreich's ataxia	FXN	GAA	Intron 1	9	69035752-69079076	(189)
C9ORF72 amyotrophic lateral sclerosis frontotemporal dementia	C9ORF72	GGGGCC	Intron	9	27573866-27546546	(190)

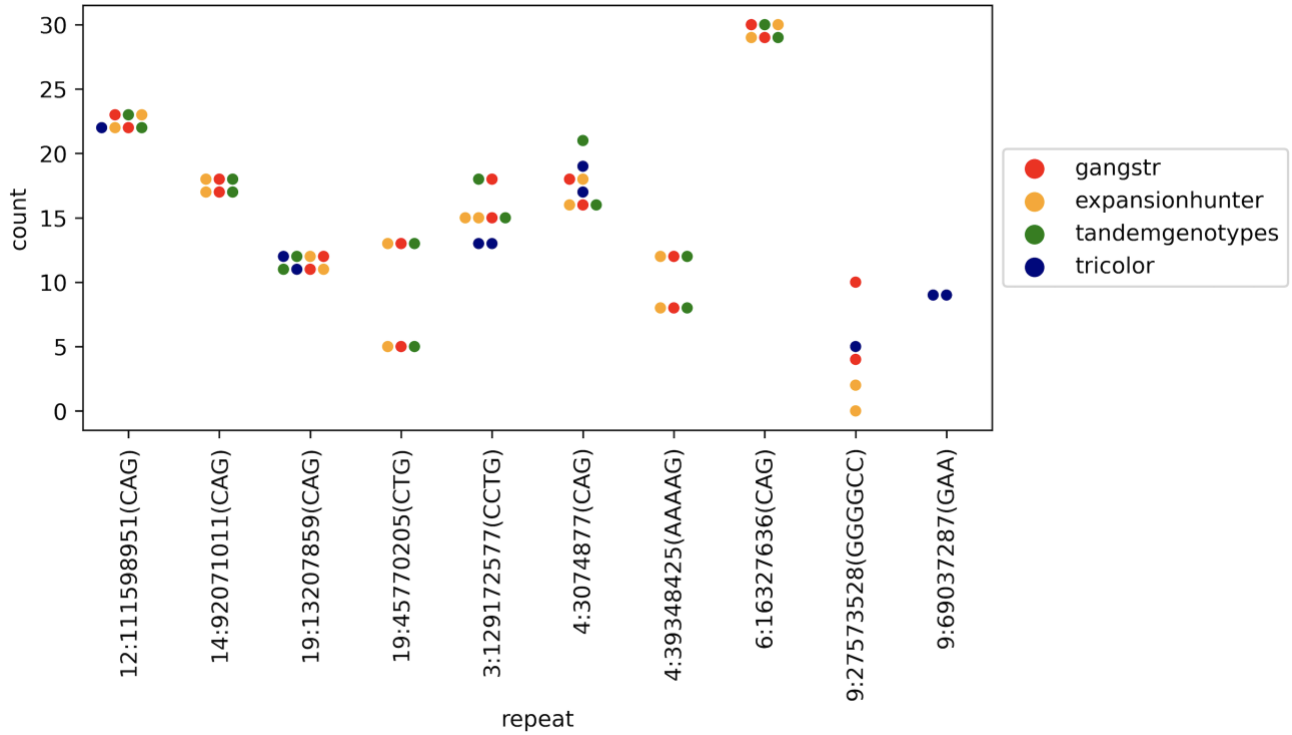

Figure 3-8 Pathogenic Variant concordance across the tandem repeat characterization methods.

Repeat count estimations for each method for ten repeats where expansions are known to be pathogenic. Each method should have two results for each repeat region, one from each chromosome.

Chapter 4 - Integrating Multiple Sequencing Technologies to Identify Repeat Expansions in a Disease Cohort

The material presented in this chapter is derived from preliminary work in progress as part of a collaborative project and is not being pursued as a manuscript at this time. The computational portion of this project aimed to identify potential repeat expansions in participants of the Michigan Genomics Initiative diagnosed with neurodegenerative diseases. These potential expansions can be targeted with long-read Oxford Nanopore sequencing for more accurate characterization of repeat expansions which can be used to further study their association with disease. Dr. Ryan Mills, Dr. Alan Boyle, and Dr. Peter Todd devised the context and scope of the project. Dr. Mills provided guidance for computational approach, and I led the implementation of the computational analysis.

4.1 ABSTRACT

Amyotrophic lateral sclerosis (ALS), or Lou Gehrig's disease, is the most common motor neuron disease in adults. Although 5-10% of all cases have a familial connection, the vast majority of incidents have no known genetic factor. Recent genome-wide association studies have identified several SNPs that correlate with ALS phenotypes. However, the strongest connections have been linked to tandem repeat expansions which have been critically understudied due to technological limitations. Here, we developed and applied an approach to link repeat expansions with nearby SNPs and report potentially relevant repeat expansions in two cohorts of ALS patients, one with short-read whole genome sequencing (WGS) (n=24) and one with SNP microarray data (n=31). We estimate repeat lengths in the WGS dataset using Expansion Hunter and compare two methods for estimating repeat lengths using surrounding SNP haplotypes for the SNP microarray datasets.

We compare the estimated repeat lengths in the ALS samples to a control population to identify candidate repeat expansions which could be confirmed using targeted long read sequencing.

4.2 BACKGROUND

Amyotrophic lateral sclerosis (ALS) is the most common motor neuron disease in adults with a prevalence of approximately 6 cases per 100,000 people (191). ALS is a neurodegenerative disease affecting the nerve cells in the brain and in the spinal cord. The age of onset is typically between 40 and 70 with patients typically surviving 2-4 years after symptoms present (192). As the world population continues to age and diagnostic capabilities grow, the prevalence of ALS is projected to increase by 69% by 2040, underscoring the need to better understand the genetic underpinnings of this disease (193).

In approximately 5-10% of ALS cases, there is a known family history of the disease (194). About half of familial cases have variants in genes known to be linked with ALS. Some of the most common variants include SOD1, C9ORF72, FUS and TARDBP (195–199). The remaining 90-95% of ALS cases are considered sporadic cases and do not have a known family history of the disease; however, it is estimated that heritability of ALS is around 60%. This suggests there are additional variants to be uncovered. Several large association studies have identified loci associated with ALS but oftentimes these studies are unable to determine the causal variant (200–202). Other neurological diseases such as Huntington's Disease and Fragile X syndrome are caused by repeat expansions (203) and offer a reasonable target variant class for investigation.

Repeat expansions in the C9ORF72 gene are the most common known ALS variant (204); however, tandem repeat variations but have been harder to include in large association studies due to limitations in the ability to characterize them with currently available technology. While short read sequencing methods can be used to characterize tandem repeats, this technology still has several limitations when assessing repetitive sequences. For example, issues arise when aligning a short read derived from a repetitive region of the genome that consists of a large proportion of repetitive sequences. In this case, alignment methods often produce multiple alignments or fail to align the read altogether. Because of this, repetitive regions are often excluded in variant analysis pipelines. However, several methods have been developed attempting to overcome these challenges and detect repeat expansions in the genome which can aid in the detection of additional repeat expansions associated with ALS (89,90).

Several methods have been developed for repeat expansion characterization from whole genome sequencing datasets as discussed in depth in the previous chapter. While long read sequencing has advantages over short read sequencing for detecting complex variation like tandem repeat expansions, generating new whole genome, long read sequencing datasets for many samples is often cost prohibitive (93). Comparatively, short-read sequencing is currently more practical for generating whole genome sequencing data for large numbers of samples, such as those that may be required for the discovery of more repeat expansions linked with ALS. However, there is an abundance of genomic data already generated for individuals with various diseases, including ALS, in large biobanks or databases such as the Michigan Genomics Initiative (205).

Despite the availability of these datasets, they often only contain SNP microarray data instead of whole genome sequencing given that microarrays are significantly less expensive, especially at large scale. While microarrays are not straightforward for characterizing repeat expansions, it has previously been shown that repeat lengths can be estimated from surrounding SNP variations in microarray datasets (206–208). The ability to characterize repeat lengths in the samples included in these databases would significantly increase the overall number of disease samples that could be included in large association studies to identify potential repeat expansion loci linked to any given disease. Because there are limitations in profiling repeats using short reads and SNP arrays (209), expansions identified using these approaches are great candidates for targeted long read sequencing, which can provide more accurate repeat length estimates while avoiding the cost of whole genome long read sequencing (167,210).

Here we identify potential repeat expansions in individuals with ALS from available short read and SNP array datasets with the goal of prioritizing regions of the genome for targeted long read sequencing. The ALS samples come from two cohorts. One cohort includes Illumina whole genome sequencing (WGS) samples generated by the Mayo Clinic (211), and the other cohort contains SNP genotyping array data from the Michigan Genomics Initiative (MGI) at the University of Michigan (205). We estimate the length of tandem repeats near SNPs associated with ALS in samples with WGS using Expansion Hunter (90), a method for characterizing tandem repeats in short read data. We utilize these tandem repeat callsets to develop and assess a method for estimating tandem repeat lengths from SNP array data. We compare this method to the existing imputation method, Beagle (100). Finally, we compare the repeat length estimates for the ALS

samples to a healthy control population to identify potential expansions which can be targeted with long read sequencing for more accurate characterization.

4.3 METHODS AND MATERIALS

To alleviate any issues with nomenclature, I have provided a definition of terms used throughout the methods and results sections (see Box 1).

Box 1: Terminology used in Chapter 4

ALS Cohort – refers to all samples diagnosed with ALS (both WGS and Microarray)

WGS ALS Cohort – refers to the 24 samples with whole genome sequencing data obtained from the Mayo Clinic

Microarray ALS Cohort – refers to the 31 samples with SNP microarray data obtained from MGI

1KG Cohort – refers to the 2,504 samples with whole genome sequencing from the 1000 Genomes Project

Repeat loci – refers to a region in the reference genome containing a repetitive sequence

Repeat motif – refers to the base pair sequence that is repeated at the repeat loci

Repeat length – refers to the number of copies of the repeat motif for a given repeat loci

Repeat Expansion – refers to the increase in the number of copies of the repeat motif in comparison to either the reference genome or to a control population

Candidate Repeat Expansions – refers to repeat loci which are estimated to be repeat expansions and are potential candidates for further analysis with targeted long read sequencing

Datasets

ALS Cohort

WGS ALS Cohort

We obtained BAM files for whole genome Illumina sequencing data of 24 individuals diagnosed with ALS from the Mayo Clinic. DNA was extracted from four different brain tissues (211). We pooled reads generated from the four tissues for each sample for SNP calling. GATK (146) was used to call SNP variation using the germline short variant detection best practice pipeline. SHAPEIT2 (212) was used to phase SNPs.

Microarray ALS Cohort

We obtained SNP genotype data from the Michigan Genomics Initiative (MGI) (205) for 31 individuals diagnosed with ALS who also have biospecimens available. Genotyping was performed by MGI using a custom genotyping array based on the Illumina Infinium CoreExome-24 bead array. Phasing (Eagle (213)) and imputation (TOPMed imputation server (214)) were previously performed by MGI.

1000 Genomes Control Samples (1KG Cohort)

We obtained high coverage whole genome Illumina sequencing BAM files and previously phased SNP calls for 2,504 presumably healthy samples from the 1000 genomes project (1KGP) as control samples (25). This includes samples from 26 diverse populations around the world. These samples have an approximate coverage of ~30X.

Methods for estimating tandem repeat length

Expansion Hunter

Expansion Hunter (90) is a method for characterizing the length of tandem repeats from short read sequencing data. Expansion Hunter requires a list of repeat coordinates in the reference genome

known to contain tandem repeats in order to assess each region. This method was discussed and benchmarked in the previous chapter.

Beagle 5.1

Beagle (100,215) is a well-known method for the imputation of ungenotyped variation. While not strictly used for estimating tandem repeat lengths, tandem repeat lengths can be treated as multi-allelic variants which can be phased along with SNP variants to generate reference panels. In turn, these reference panels are used to estimate the repeat lengths in SNP Microarray datasets - a technology that, when used alone, does not genotype tandem repeats.

ForecaSTR

We created ForecaSTR as a potential alternative to Beagle for estimating tandem repeat lengths in SNP Microarray datasets. For each repeat loci of interest, we generated a training set using individuals with whole genome sequencing data (in this study, we used 2,504 presumably healthy samples and 31 samples diagnosed with ALS) that have homozygous repeat lengths for the loci as estimated by Expansion Hunter. For all individuals included in the training set, the two haplotypes containing phased SNPs (that are also present on the SNP microarrays) within 100,000 bp of the repeat boundaries were extracted creating a matrix with dimensions 2S x N where S is number of homozygous samples and N is the number of SNPs within 100,000bp of the repeat loci. Each entry in the matrix is coded as either 0 or 1 indicating the presence or absence of the reference allele in the individuals for the SNP. The repeat length estimates for the homozygous samples were extracted to create a matrix with dimensions 2S x 1 where the values of the matrix indicate the

number of repeat copies present in the individuals. (Only homozygous samples were used to ensure the correct SNP haplotypes were paired with the correct repeat estimate). (Figure 4-1).

We used the phased SNP haplotypes matrix paired with the estimated repeat lengths matrix to train an ordinal linear model for each repeat loci. The input variables of the ordinal linear model included the presence or absence of the surrounding SNPs in each reference sample and the output variable of the model was the predicted repeat count for each of those samples. Using an ordinal model treats the repeat counts as categorical variables but also accounts for the sequential nature of the increasing repeat count sizes. The model only outputs repeat lengths present in the training dataset. The number of homozygous samples and SNPs used for prediction, and the number of unique repeat lengths present in the training datasets are recorded for each repeat loci. To estimate the length of repeat loci in samples genotyped using SNP arrays, SNPs within 100,000 base pairs of a repeat loci are input into the ordinal linear model trained for the repeat.

Identifying reference repeat loci near ALS GWAS SNPs

To narrow the scope of the project, we focused our analysis on repeats loci that may be driving the SNP associations of ALS GWAS hits. All associations for ALS included in the GWAS catalog (126) were downloaded (download date: 04-26-22). All associations with genomic coordinates were kept (Appendix Table C 1). We downloaded the set of simple repeats present in the reference genome which was generated by running Tandem Repeat Finder (175) on the reference genome GRCh38 from the UCSC table browser (216) and identified repeats within 40,000 bp on either side of the ALS GWAS hits to create a set of repeats loci for analysis. We removed repeats with

repeat motif lengths of one or greater than 15bp as these can be difficult to characterize using short reads and the threshold is consistent with other thresholds used in the field (89).

Estimating tandem repeat lengths in WGS samples

To estimate the repeat length of the repeat loci near ALS GWAS hits in the two whole genome sequencing datasets (WGS ALS samples and 1KG control samples), we ran Expansion Hunter (default parameters). For the ALS samples, Expansion Hunter was run twice, once by pooling the different brain tissues together and once on each brain tissue separately. We created the input variant catalog to include the reference repeat loci near ALS GWAS hits. Each repeat was made into a JSON entry using the repeat motif as the locus structure in the format "([RU])*" and "Repeat" as the variant type. Repeat loci within 1,000 bp of a gap in the reference were removed, as these caused the program to error out.

Estimating repeat lengths in the MGI ALS cohort

We applied two approaches for estimating repeat length for the repeat loci near ALS GWAS SNPs in the Microarray ALS samples.

Beagle 5.1

We built a reference panel using the two WGS datasets (WGS ALS samples [pooled tissues] and 1KG Cohort) by combining the SNP and STR callsets generated by GATK and Expansion Hunter respectively. These variants were phased using the phasing method implemented in Beagle (215) with default parameters. The reference panel was then used to impute STRs onto the SNP

haplotypes of the Microarray ALS cohort using the imputation method implemented in Beagle (100) (default parameters).

ForecaSTR

We used the WGS ALS (pooled tissues) and 1KG Cohort datasets to train an ordinal linear model for each reference repeat loci near ALS GWAS hits. We extracted the phased haplotypes containing SNPs within 100,000 bp of the repeat boundaries from the samples in the Microarray ALS cohort creating a matrix with dimensions 2A x N where A is number of Microarray ALS samples and N is the number of SNPs within 100,000bp of the repeat. The SNP haplotypes matrix for each repeat loci was input into the trained model for the repeat to estimate the repeat lengths for the samples in the Microarray ALS cohort.

Benchmarking repeat length estimations from SNP profiles

We generated simulated array data for the 1KG sample NA12878 which has whole genome short read sequencing as a test dataset to assess how close the repeat length estimations derived from SNP profiles are to the repeat length estimations from short reads. The simulated array was created by sub-setting the SNP calls based on the WGS for sample NA12878. The SNP calls were subset to include only the SNPs present in the genotyping array used to genotype the Microarray ALS cohort.

For Beagle, we removed NA12878 from the generated reference panel and imputed the STRs into the simulated array data using the imputation method implemented in BEAGLE. For ForecaSTR, we trained the ordinal linear model using the whole genome sequencing datasets, 1KG cohort

excluding NA12878 and WGS ALS cohort, and estimated the repeat lengths in NA12878 using only the simulated array SNPs as input for ForecaSTR.

Null Models

We created several null models to compare against Beagle and ForecaSTR. For the Beagle random model, we randomly selected two alleles from all repeat alleles present for a repeat in the reference panel for each repeat loci. For the Beagle naïve model, we selected the most common allele present in the reference panel for each candidate repeat. For the ForecaSTR random model, we randomly selected repeat lengths from those present in the ForecaSTR training set which includes only the repeat lengths from the samples homozygous for repeat length from the whole genome sequencing datasets.

Identifying candidate repeat expansions in the ALS cohort

To identify potential repeat expansions, we compared the estimated repeat lengths in the ALS samples (WGS ALS cohort [pooled tissue] repeat lengths estimated by Expansion Hunter; Microarray ALS cohort repeat lengths estimated by Beagle 5.1) to the estimated repeat lengths in the, presumably healthy, 1KG cohort samples (estimated by Expansion Hunter). For each repeat loci in the reference genome near known ALS GWAS hits, the mean and standard deviation of the estimated repeat lengths in the 1KG control cohort were calculated. Each sample contributed two repeat lengths, one for each allele.

Repeat length outliers present in each ALS sample were identified by comparing each repeat length allele present at each locus to the control distribution using a z-test. Multiple test correction was

applied using the Benjamini-Hochberg procedure (217) and results with an adjusted p-value less than 0.05 were considered potential repeat expansions. Repeat loci were reported as candidate repeat expansions if a repeat length outlier was identified in more than one ALS sample.

Candidate repeat expansions in genes

The genome annotation for GRCh38 was downloaded from GENCODE (V39) (218). Repeat loci in genes were identified by intersecting the coordinates of repeat loci with coordinates of "gene" features from the GENCODE V39 annotation using the intersect module of bedtools2 (180). The same procedure was used to identify repeat loci present in exons using the "exon" features from the GENCODE V39 annotation.

Candidate repeat expansions expressed in brain

Gene expression levels across different tissues were obtained from GTEx v8 (219). The median expression scores were extracted for all genes containing candidate repeat expansions. The expression scores were put in two groups, one for tissues in the brain and one for the remaining tissues in the dataset.

4.4 RESULTS

Development of an approach to link SNPs with tandem repeat expansion length

We created a method, ForecaSTR, for estimating repeat lengths from surrounding SNP haplotypes as described in the Methods. In brief, we create a reference panel using SNP and tandem repeat variation callsets generated from whole genome sequencing datasets. We trained ordinal linear

models for each input repeat loci using phased SNP haplotypes and the length of repeat loci as explanatory and response variables respectively. We used samples homozygous for the repeat length to ensure the SNP haplotypes were paired with the correct repeat length value. To estimate the length of repeat loci in samples genotyped with SNP microarrays, phased SNP haplotypes containing SNPs surrounding repeat loci coordinates were input into the corresponding trained linear model. (Figure 4-1)

Identifying reference repeat loci near ALS GWAS SNPs

To identify potential repeat expansions associated with ALS, we focused our analysis on repeat loci within 40,000bp of ALS GWAS hits. Of the over one million repeat loci present in the reference genome (as characterized by Tandem Repeat Finder), we identified 2,612 falling within 40,000 bp of 217 ALS GWAS hits. Each GWAS SNP had between 4 and 94 repeats with an average of 26.7 repeats within 40,000 bps per hit. (Figure 4-2 A). For this analysis we capped repeat unit lengths at 15 because repeat unit lengths greater than 15 are difficult to characterize using short read technologies. Figure 4-2 B displays a histogram of the repeat unit lengths. Repeat motifs with one, two, and four base pairs were the most common, but larger VNTR repeats motif lengths are also included. This set of 2,612 repeat loci were used for all following analyses. Note, identifying reference repeat loci is not an outcome of running ForecaSTR (or Beagle), but is a input to run the algorithms.

Assessing the performance of tandem repeat length estimation from SNP haplotypes

To assess and compare the accuracy of the two methods, Beagle and ForecaSTR, for estimating tandem repeat lengths from SNP haplotype data, we simulated array data for repeat loci near ALS

GWAS hits for the 1KG sample NA12878. We imputed repeat lengths for the reference repeat loci near ALS GWAS hits using the simulated data as input for two methods: 1) Beagle and 2) ForecaSTR (see Methods). We then compared the repeat lengths estimated from each of the imputation methods to those estimated from short read data for NA12878 using Expansion Hunter. Repeat length estimates from Beagle and ForecaSTR had similar concordance with Expansion Hunter, with Beagle reporting the same repeat length estimate for 81.2% of the alleles and ForecaSTR reporting the same repeat length estimate for 78.6% of the alleles (Figure 4-3). These are both significantly better than the random null model which matched the Expansion Hunter estimated length for 21.4% of alleles. 91.9% of alleles predicted by Beagle were within 2 counts of the Expansion Hunter estimate while 87.3% of alleles predicted by ForecaSTR were within 2 counts of the Expansion Hunter estimate. While not nearly as striking as the random null model, both Beagle and ForecaSTR predict repeat length alleles closer to Expansion Hunter than the naïve model (Figure 4-3).

We next examined potential variables that might influence the accuracy of ForecaSTR predictions. We investigated whether the number of training samples, surrounding input SNPs, or unique repeat length alleles influenced how well the ordinal linear model (ForecaSTR) estimated repeat lengths and matched the short read estimated repeat lengths by Expansion Hunter. There was no clear linear relationship between any of these variables and how closely the estimated lengths of the ordinal model estimation and Expansion Hunter were to each other (Figure 4-4).

Identifying candidate repeat expansions in the ALS cohort

We estimated the length of the 2,612 repeat loci near ALS GWAS SNPs for samples in the ALS and 1KG cohorts. Expansions Hunter was applied to the 2504 1KG cohort samples and the 24 WGS ALS cohort samples. Beagle was applied to the 31 Microarray ALS cohort samples. We identified repeat lengths in the ALS cohort samples that were outliers compared to the distribution of repeat lengths present in the 1KG control population using a z-test. 107 repeat loci had at least one repeat length allele outlier in at least two ALS samples (Table 4-1).

Repeat loci present within gene boundaries are indicated in Table 4-1. 62 repeat loci fell within gene boundaries with 7 of those loci falling within exons. Expression data across the 54 tissues included in the GTEx database are plotted in Figure 4-5 for each of the genes containing a repeat length outlier in the ALS sample. Several genes have high expression in brain tissues with some genes being almost exclusively expressed in the brain. One such gene is the CTNND2 gene containing a CA repeat which has expanded allele lengths estimated in 9 WGS ALS samples and 14 Microarray ALS samples (Figure 4-5) with the ALS samples having 11 or 12 copies more than the mean count seen in the 1KG datasets (standard deviation in 1KG: 2.8 copies).

Additionally, we looked at whether different repeat lengths were identified in different brain tissues for the WGS ALS samples. Some repeat loci were variable between individuals, but the same length was observed across all four tissue types within a sample (Figure 4-6 A). Other repeat loci were highly variable across tissues within the same sample (Figure 4-6 B). For a few repeat loci, including those in the *PTPRN2* and *ABCG1* genes, repeat expansion generally seemed to occur in tissues 2-4 but repeat expansions were not seen in tissue 1 (Figure 4-6 C,D).

4.5 DISCUSSION

ALS is a disease with a genetic component which is not yet fully understood. While several GWAS have identified significant loci associated with ALS, many of the causal variants have not yet been identified (200–202). Tandem repeats have historically been difficult to characterize in the genome and may contribute to the missing heritability of ALS. While long read sequencing is advantageous for characterizing repeat expansion, it can be cost prohibitive to perform at the level of whole genome sequencing (93,167,210). We sought use previously generated genomic datasets to prioritize repetitive regions for further interrogation in subsequent studies using targeted long read sequencing.

We developed and presented a method to estimate the length of repeat loci from microarray genotype data to expand our repeat characterization analysis to include additional ALS samples from the MGI database. When testing the method on a sample with WGS, we found over 87% of the estimations to be within 2 repeat counts of the estimation determined by Expansion Hunter; however, the method was outperformed by the established imputation method Beagle.

We identified expanded repeats in 24 ALS patients compared to a healthy population including a particularly interesting repeat locus in the intron of the *KCNG2* gene where an expanded repeat was found in 41 of the 48 alleles from the 24 samples. This gene is primarily expressed in the brain and is a member of a family of potassium voltage-gated channel modifier genes (220). As potassium channels are an important part of neuronal action potential propagation – the opening of voltage-gated potassium channels follows the flux of sodium into a cell during the action

potential, causing hyperpolarization (221) - a mutation in KCNG2 could potentially impact action potential signaling in nerve cells and pathologically alter their behavior in ALS (222).

A CA repeat expansion in the PTPRN2 gene was estimated to be expanded in 18 of the 48 alleles from the WGS ALS samples. This gene is also primarily expressed in the brain. The repeat length was highly variable within the four brain tissues sequenced; however, it was never estimated to be expanded in tissue 1. While a repeat expansion in this gene has not been previously associated with ALS, this gene was found to have high levels of DNA methylation in ALS individuals compared to monozygotic twins who did not have the disease (223). While this is not directly related to repeat expansions, hypermethylation has previously been shown to be correlated with tandem repeat expansion status (224–226). Together, these provide support for further research into repeat expansions in this gene to determine a potential link to ALS.

When preforming outlier detection on the repeat lengths estimated by ForecaSTR for the MGI ALS samples, we observed that by requiring samples to be homozygous for a locus in order to be included in the training set, we are potentially unable to estimate the most informative repeat loci. For example, we were unable to reliably estimate the repeat length in the MGI ALS dataset at the TG repeat locus on chromosome 18 that appeared to be expanded in many samples in the WGS ALS cohort. The repeat lengths were highly polymorphic in the 1KG control population at this locus and no samples were observed to be homozygous at this locus in this cohort. Only three repeat lengths were included in the training dataset, all deriving from WGS ALS samples with expanded repeats, which would result in all predicted repeat lengths in the Microarray ALS dataset to appear as outliers. We were able to apply Beagle to impute different repeat lengths into the

Microarray ALS samples for this variant. Another limitation of only using homozygous samples in the training set can occur if a repeat expansion is very rare in the general population. While it is possible the rare allele is present in a homozygous state, it is likely there are many instances where a rare expansion is only present on one allele and will be excluded, resulting in only the common allele lengths included in the training set. Using beagle instead of ForecaSTR for imputing repeat expansions does not rely on only using homozygous samples from the training set and an improvement was seen when using this method. However, the limitation of only estimating repeat lengths present in the training set/reference panel, still persists. This lends to the importance of including disease samples in the refence panel when applying this approach. Only 24 WGS ALS samples were used in this analysis. Future work can incorporate additional samples, like those present in publicly available datasets such as Answer ALS (27).

Expansion Hunter was used to estimate repeat lengths in WGS data but this method has limitations which are important to acknowledge when assessing the results of all analyses for this chapter. 1) Expansion Hunter requires a list of coordinates and repeat motifs to characterize a given repeat loci and therefore, no novel repeat expansions can be discovered. 2) Expansion Hunter is a method for detecting repeat expansions from short reads and has several limitations for characterizing complex or repetitive sequences in the genome. These limitations extend to the repeat lengths estimated via imputation for samples with SNP array data due to the models being entirely based on repeat lengths estimated using Expansion Hunter in the training datasets. All benchmarking efforts in this paper were only to determine if the imputation methods matched Expansion Hunter and did not attempt to assess if the Expansion Hunter results accurately represent what was in the sample as was presented in the previous chapter.

Our goal was not to estimate the length of a repeat to the exact count but instead to prioritize any repeat regions in the Microarray ALS samples that are potentially expanded compared to a healthy population. In this work, we have identified repeat loci that appear to be expanded in several samples with ALS. As these repeat loci are potentially encouraging targets, next steps include performing targeted long read sequencing on the MGI ALS samples with biospecimens available to obtain more accurate estimates of repeat size in these individuals prior to including them in any large association studies. Examining the repeat expansions in additional ALS samples will increase the sample size and thereby provide increased statistical power necessary to perform larger association studies.

4.6 FIGURES AND TABLES

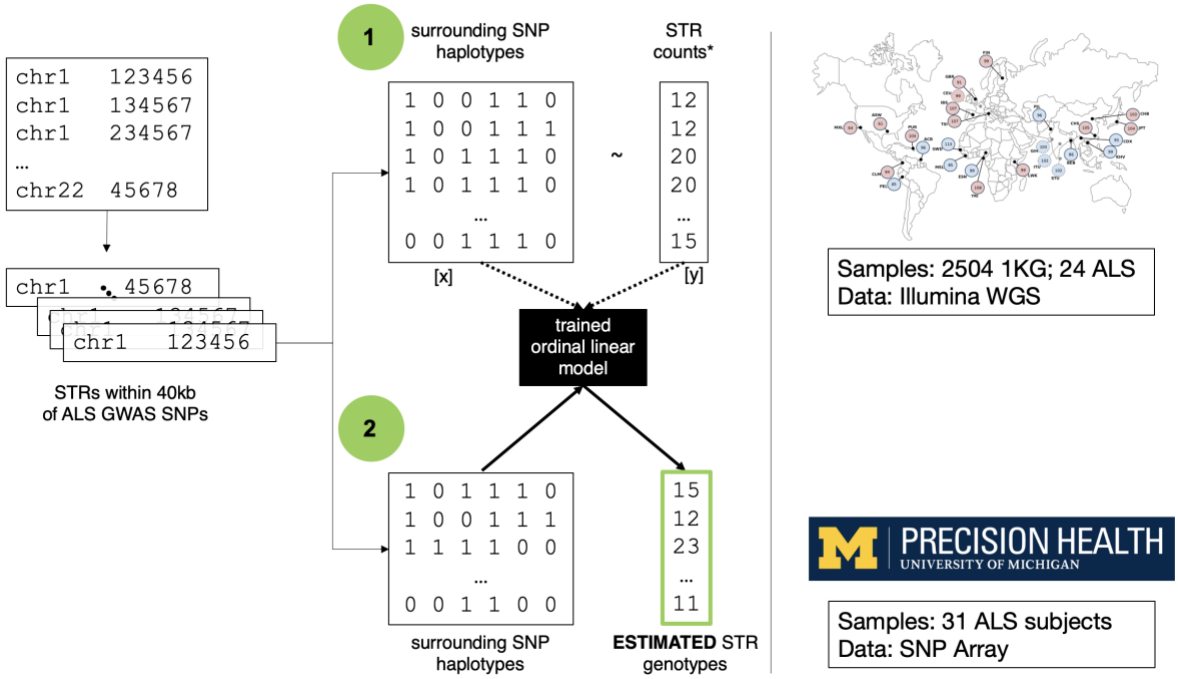

Figure 4-1 Schematic of the ForecaSTR method

1) For each repeat supplied to ForecaSTR, a training set is created using individuals with whole genome sequencing data that have homozygous repeat lengths for the loci as estimated by Expansion Hunter (or any other WGS tandem repeat characterization method). For all individuals included in the training set, the two haplotypes containing phased SNPs also present on the SNP microarrays within 100,000 bp of the repeat boundaries are extracted. The repeat length estimates for the homozygous samples are paired to the SNP haplotypes and are used to train an ordinal linear model. 2) To estimate the length of repeat loci in samples genotyped using SNP arrays, SNPs within 100,000 base pairs of a repeat loci are input into the ordinal linear model trained for the repeat.

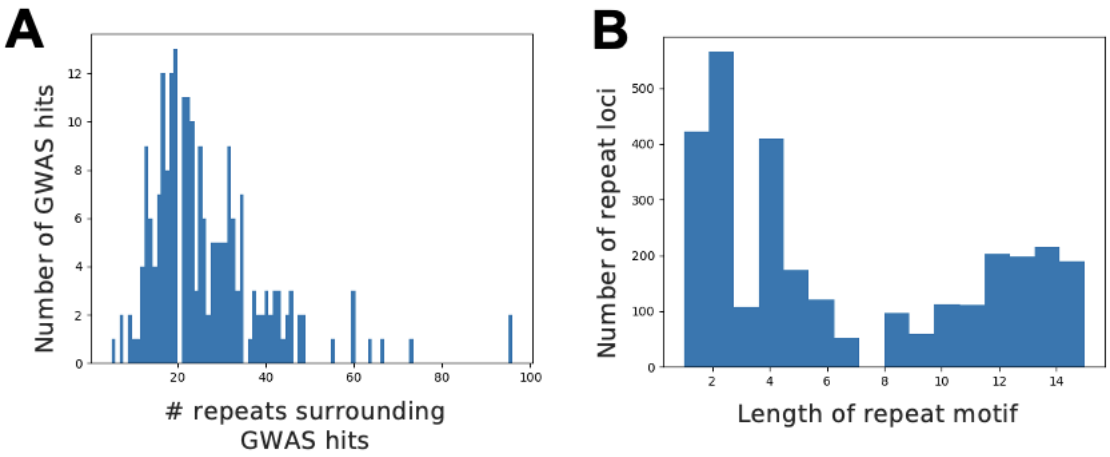

Figure 4-2 Characteristics of repeat loci surrounding ALS GWAS hits

A) Histogram showing the number of repeat loci within 40kb for each ALS GWAS hit B) Histogram showing the distribution or repeat motif lengths for the repeat loci included in the outlier analysis

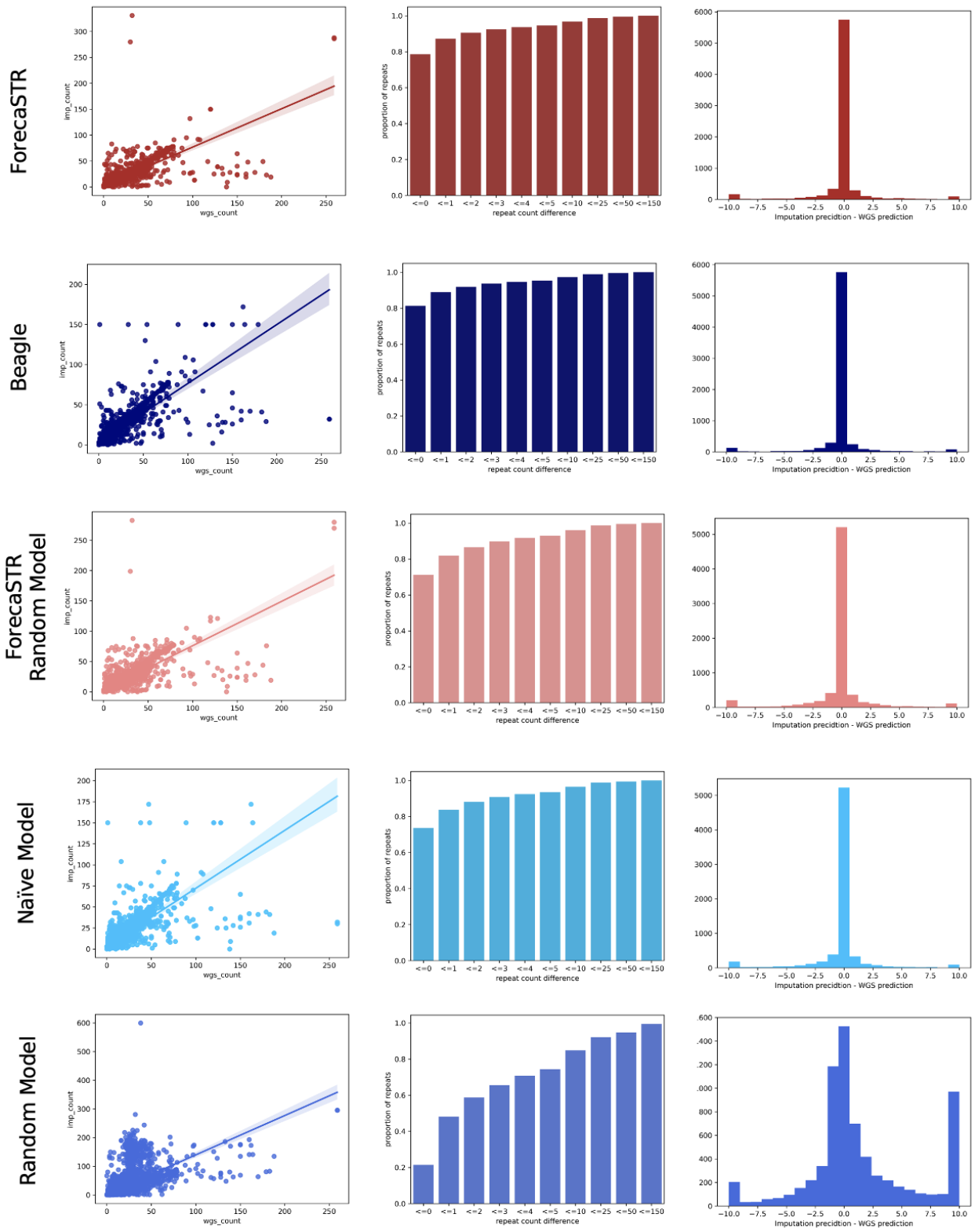

Figure 4-3 Benchmarking tandem repeat imputation methods in sample NA12878

Each row compared the results of either an imputation method or a null model to Expansion Hunter. Rows one and two are imputation methods and rows three through five are null models.

Column one compares the repeat lengths predicted by short-read tandem repeat detection method Expansion Hunter to the repeat lengths predicted by the imputation methods or null models. Column two shows the cumulative number of repeats where the repeat length estimates by the imputation methods or null models are off by various number of counts compared to repeat length estimate determined by Expansion Hunter. Column three contains histograms showing the difference in the number of repat motif counts estimated by Expansion Hunter and the imputation methods or null models. Negative values indicate the imputation method or null model estimated repeat lengths less than Expansion Hunter while positive numbers indicate the imputation method or null model estimated repeat lengths greater than Expansion Hunter. (All repeat loci off by more than ten repeat counts were placed in the -10 and 10 bins)

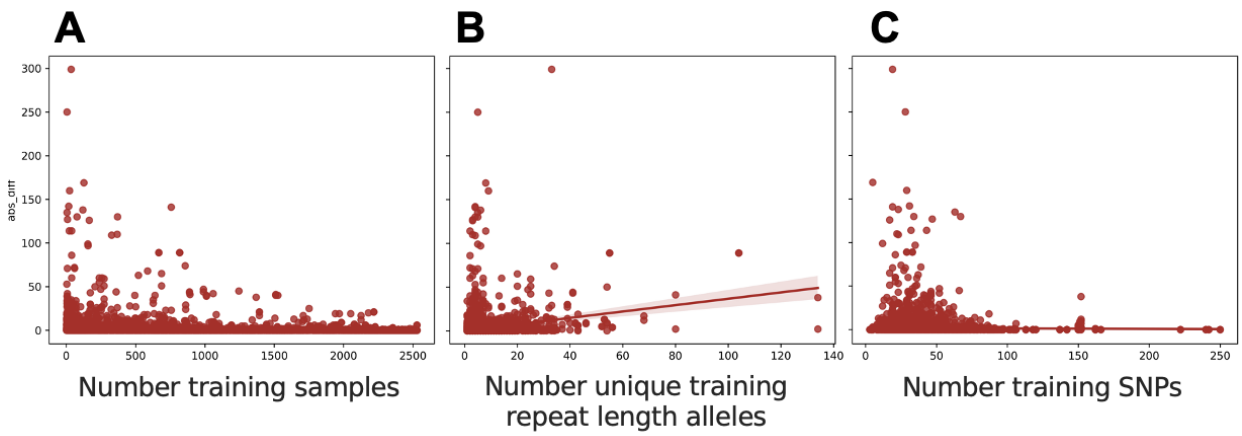

Figure 4-4 Effect of features of ForecaSTR training set on accuracy of repeat length prediction in sample NA12878

Each scatter plot shows the relationship between a feature of the training dataset and the difference in the in the number of repeat counts predicted by Expansion Hunter and ForecaSTR in sample NA12878. The features include A) the number of samples included in the training dataset for each repeat. This is based on the number of samples homozygous for each repeat loci. B) the number of unique repeat counts present in the training dataset for each repeat and C) the number of training SNPs within 100,000 bps of each repeat loci that were including in the training set.

Table 4-1 Repeat loci predicted to have outlier lengths in ALS samples

Locus	# Mayo outlier alleles (N=48)	# MGI outlier alleles (N=62)	Gene	Exon	1KG Mean	1KG SD	Mayo Repeat Counts (Repeat count x Number of samples)	MGI Repeat Counts (Repeat count x Number of samples)
chr18-79829186-79829354-TG	41	6	KCNG2	.	11.00	4.45	73x1,71x1,69x1,65x2,64x8,63x2,57x11,43x14,32x1	57x2,43x3,32x1
chr2-48615466-48615559-GGAGAG	33	17	STON1-GTF2A1L	.	8.00	1.48	27x1,26x2,25x4,24x5,23x6,22x4,21x2,20x1,18x7,17x1	25x1,19x1,18x14,15x1
chr8-142013108-142013194-ATCACC	25	14	.	.	7.00	1.48	16x16,15x9	16x2,15x12
chr8-2541227-2541333-AT	23	29	ENSG00000282142	.	40.00	2.97	69x1,53x21,51x1	56x1,54x3,53x25
chr13-22019636-22021009-TA	21	10	.	.	12.00	5.93	72x2,70x3,69x2,68x2,63x1,53x2,50x2,49x1,47x2,39x4	39x10
chr13-53020110-53020154-TG	20	22	.	.	17.00	1.48	24x13,23x7	24x16,23x6
chr1-26595384-26595451-CCCTGC	19	8	.	.	11.00	1.48	25x1,24x1,23x4,22x2,21x3,20x3,19x1,18x1,17x3	24x2,22x3,21x1,20x1,19x1
chr18-58138329-58138398-CCT	18	22	NEDD4L	.	27.00	1.48	41x9,40x1,39x1,38x1,37x1,36x1,34x3,33x1	41x20,38x1,34x1
chr7-157656810-157658038-ACACACACACAC	18	13	PTPRN2	.	17.00	7.41	92x1,88x1,86x1,81x1,80x1,77x1,76x3,75x5,74x1,71x1,59x1,57x1	90x1,82x1,75x7,73x3,72x1
chr3-10471364-10471406-TG	18	12	ATP2B2	.	16.00	1.48	28x4,27x9,26x5	27x12
chr6-32709842-32709890-CTTC	17	21	.	.	4.00	1.48	17x1,12x16	17x1,16x1,13x1,12x17,11x1
chr16-85595802-85595831-CCCA	17	.	GSE1	.	8.00	1.48	37x1,32x3,31x1,30x2,29x2,28x2,27x2,26x1,24x1,22x1,20x1	.
chr10-5576277-5576441-CCTCTCTCTCTGTCT	16	1	.	.	3.00	1.48	9x2,8x14	8x1
chr10-112419102-112419131-TG	16	.	ENSG00000232934	.	14.00	1.48	21x1,20x5,19x10	.
chr10-112419102-112419131-TG	16	.	ACSL5	.	14.00	1.48	21x1,20x5,19x10	.
chr6-32709795-32709897-TCTCTTCCT	15	17	.	.	5.00	1.48	15x1,12x10,10x4	15x1,14x1,12x15
chr1-188084965-188085119-TATATATATAAAATTTA	14	17	ENSG00000285894	.	9.00	1.48	89x1,80x2,79x1,76x1,73x1,70x2,67x1,65x1,61x1,26x3	79x1,78x3,76x2,75x2,74x1,72x2,71x1,69x2,50x1,49x1,26x1

chr11-127874911-127874958-TG	13	12	.	.	14.00	1.48	24x4,21x1,20x3,19x5	24x3,23x6,22x2,20x1
chr10-3760818-3760881-TTCC	12	1	ENSG00000229672	.	17.00	1.48	24x1,23x5,22x6	22x1
chr2-238627901-238627968-TG	11	25	LINC01937	.	35.00	1.48	49x1,48x1,47x4,46x1,45x1,44x1,43x1,40x1	62x1,56x1,55x1,48x4,47x9,46x3,45x3,44x1,43x1,41x1
chr2-55771310-55771341-AC	11	7	.	.	17.00	1.48	25x4,23x2,22x5	25x1,24x1,23x5
chr20-15233969-15234010-TA	11	.	MACROD2	.	21.44	4.20	59x1,39x10	.
chr18-31709553-31709604-TG	10	1	.	.	22.00	1.48	28x1,27x9	27x1
chr8-2191106-2191425-TCCT	10	.	.	.	7.00	5.93	39x1,37x2,35x2,34x2,33x1,32x2	.
chr5-11081072-11081122-CA	9	14	CTNND2	.	24.01	2.80	35x9	36x1,35x13
chr4-36043832-36043866-GGGGAGGGGAGG	9	3	ARAP2	.	4.00	1.48	12x1,11x1,9x7	10x2,9x1
chr21-31546556-31546587-AAAG	7	4	TIAM1	.	8.33	1.86	19x1,18x2,16x2,15x2	18x2,17x1,15x1
chr3-127045510-127045646-AGGGAGAG	7	.	.	.	12.64	4.91	29x7	.
chr21-42264391-42264552-TCCATCCATCCA	7	.	ABCG1	.	12.38	2.55	23x2,22x1,21x4	.
chr9-18634858-18634921-AT	6	9	ADAMTSL1	.	32.40	1.04	50x6	70x1,50x6,49x1,48x1
chr20-52119533-52119568-AT	6	9	ZFP64	.	19.04	1.60	30x1,29x5	30x2,29x7
chr8-2178538-2178605-CTTC	6	6	ENSG00000289036	.	16.25	3.45	40x1,39x1,38x1,37x1,36x1,32x1	41x1,38x1,37x3,35x1
chr8-2171578-2171604-TA	6	4	ENSG00000289036	.	11.36	1.21	24x1,21x1,19x1,17x1,16x2	17x3,16x1
chr2-207645301-207645327-AT	6	2	.	.	12.91	1.01	34x2,33x1,28x1,21x1,17x1	42x1,18x1
chr11-127874871-127874912-TA	5	12	.	.	11.00	1.48	21x4,20x1	22x2,21x4,20x3,19x2,18x1
chr2-55762428-55762536-CTTTCTCT	5	7	.	.	13.21	1.99	26x1,24x1,21x2,20x1	25x2,24x3,23x1,21x1
chr18-41645522-41645567-TG	5	5	.	.	22.59	1.08	27x5	27x5
chr6-32715478-32715542-AAAG	5	3	.	.	18.56	1.25	23x5	29x1,23x2
chr18-41613092-41613154-CA	5	1	KC6	.	32.20	1.14	36x5	37x1
chr12-26509958-26510004-TG	5	1	ITPR2	.	25.00	5.41	64x1,56x1,52x1,49x1,48x1	49x1
chr6-10439856-10439891-AC	4	10	MIR5689HG	.	17.40	1.13	25x1,24x3	25x2,24x8
chr11-18464966-18465011-TGTTT	4	4	LDHAL6A	.	9.82	0.91	19x1,18x2,13x1	17x2,16x1,13x1
chr6-32702379-32702408-GT	4	3	.	.	12.00	1.48	25x4	25x3

chr6-32715466-32715542-AAAAAGAAAG	4	2	.	.	8.60	0.86	12x4	12x2
chr20-49834185-49834215-TA	4	2	SLC9A8	.	14.59	3.22	80x1,69x1,51x1,37x1	59x1,55x1
chr14-72465550-72465697-TGGA	4	2	RGS6	.	38.71	3.97	59x1,55x1,54x2	62x1,59x1
chr13-71477080-71477105-TA	4	.	DACH1	.	12.76	1.26	52x1,45x2,31x1	.
chr13-22022288-22022527-TATATAATATGTTAA	4	.	.	.	2.31	0.94	8x1,7x3	.
chr13-22021330-22022787-TATATAT	4	.	.	.	21.04	4.74	65x4	.
chr2-34479778-34479831-TTTC	3	21	LINC01320	.	14.00	1.48	32x1,31x1,29x1	38x2,35x1,34x1,33x3,32x3,31x4,30x2,29x3,28x1,27x1
chr4-119022662-119022793-TTTAT	3	13	SYNPO2	.	27.92	1.69	35x2,34x1	40x2,39x1,38x5,36x2,35x2,34x1
chr4-13958960-13959001-AC	3	8	LINC01182	.	18.09	1.25	26x1,24x1,23x1	24x2,23x6
chr11-134868572-134868612-AC	3	3	.	.	20.25	2.67	31x2,30x1	32x2,30x1
chr5-172893868-172893906-AT	3	2	ERGIC1	.	26.25	10.34	66x1,65x1,61x1	70x2
chr14-67757640-67757722-TCTT	3	2	ZFYVE26	.	19.00	1.48	27x1,24x2	25x1,24x1
chr1-39695417-39695485-AAAGAAAGAAG	3	1	PPIE	.	6.39	0.75	11x3	11x1
chr8-2191869-2191927-TCCT	3	.	.	.	15.00	2.97	34x2,25x1	.
chr6-16761484-16761544-CTC	3	.	ATXN1, ATXN1-AS1	X	22.53	1.41	33x2,30x1	.
chr6-129446500-129446580-AGGGG	3	.	LAMA2	.	16.00	2.97	35x1,30x1,26x1	.
chr6-129446500-129446580-AGGGG	3	.	ENSG00000226149	.	16.00	2.97	35x1,30x1,26x1	.
chr4-189169845-189169874-TA	3	.	.	.	16.16	4.07	125x1,53x2	.
chr20-49834318-49834443-TATATATATACTG	3	.	SLC9A8	.	3.69	2.03	14x2,12x1	.
chr14-30596129-30596165-GT	3	.	G2E3	.	18.08	1.38	59x1,33x1,28x1	.
chr1-28990333-28990405-TCCC	3	.	EPB41	.	18.00	1.48	25x2,23x1	.
chr10-5573325-5573377-AAGG	3	.	.	.	13.04	1.48	19x2,18x1	.
chr1-43633012-43633046-AT	2	14	.	.	10.00	1.48	16x1,15x1	20x1,18x2,17x1,16x10

chr2-34479775-34479844-TTTTTTCTTTTT	2	12	LINC01320	.	6.00	1.48	12x2	42x1,28x2,27x1,22x2,21x1,16x1,15x1,14x1,12x2
chr12-26451365-26451408-CA	2	12	ITPR2	.	19.00	1.48	29x1,25x1	29x2,28x1,26x2,25x7
chr11-127852052-127852106-AAAAAAAAACA	2	8	.	.	5.00	1.48	12x1,10x1	31x2,30x3,28x1,26x2
chr6-32708874-32708913-TTGTTTTTTTTTT	2	6	.	.	2.80	2.49	12x2	12x6
chr6-32708879-32708911-TTTTTTTTTG	2	3	.	.	2.72	2.66	14x1,13x1	13x3
chr8-2531536-2531574-GT	2	1	ENSG00000282142	.	18.55	0.71	23x1,22x1	22x1
chr12-3226744-3226795-TA	2	1	TSPAN9	.	16.26	4.69	37x2	44x1
chr1-155927832-155927874-CA	2	1	KHDC4	.	14.04	3.59	28x1,27x1	27x1
chr2-207609772-207609825-CTCCCT	2	.	METTL21A	X	4.86	4.55	27x2	.
chr2-166543390-166543518-AAGAAAGAAAGAGAGAAGAGA	2	.	SCN7A	.	7.06	1.15	11x2	.
chr20-49834139-49834186-TC	2	.	SLC9A8	.	22.09	3.39	45x1,34x1	.
chr20-35226773-35226798-GCC	2	.	MMP24OS,MMP24	X	8.31	2.56	30x2	.
chr20-35223330-35223358-ATA	2	.	MMP24OS	.	10.00	1.48	16x1,15x1	.
chr19-17631154-17631205-CTCC	2	.	UNC13A	.	13.64	6.68	51x1,47x1	.
chr17-47509589-47509632-AAT	2	.	.	.	15.16	1.37	20x2	.
chr14-95624167-95624304-GGAA	2	.	ENSG00000258927	.	15.72	6.15	36x2	.
chr14-92042946-92042977-AT	2	.	.	.	13.49	3.14	47x2	.
chr1-43501209-43501307-CTCTCC	2	.	.	.	18.92	1.60	26x2	.
chr11-127863772-127863805-TG	2	.	.	.	20.00	1.48	26x2	.
chr10-131901020-131901063-CGG	1	7	PPP2R2D	X	14.68	0.81	24x1	21x1,19x6
chr21-42264391-42264552-TCCATCCATCCATCCATCCA	1	2	ABCG1	.	9.10	2.64	19x1	23x1,22x1
chr1-232004883-232005061-CCTT	1	2	DISC1	.	35.00	1.48	42x1	43x1,41x1
chr4-108674629-108674751-GAAG	1	1	ENSG00000286136	.	29.00	1.48	35x1	37x1
chr18-41577593-41577618-ACAT	1	1	KC6	.	6.40	0.66	33x1	21x1
chr1-155945280-155945323-GT	1	1	.	.	23.31	1.43	28x1	29x1

chr10-5554770-5554808-AC	1	1	.	.	19.00	1.48	24x1	25x1
chr13-53008511-53008563-TTTCTTTTC	.	13	.	.	5.84	1.72	.	15x3,14x7,13x3
chr10-86429023-86429152-AT	.	11	.	.	70.84	1.08	.	85x1,80x4,79x2,78x3,77x1
chr10-60014077-60014118-AT	.	10	.	.	17.00	1.48	.	30x1,28x1,27x1,26x1,25x2,24x3,23x1
chr2-212015041-212015112-AT	.	8	ERBB4	.	20.00	1.48	.	37x1,34x1,31x1,30x4,26x1
chr14-79504641-79504675-TA	.	6	NRXN3	.	17.00	1.48	.	26x4,24x2
chr21-41618016-41618043-AC	.	5	.	.	13.04	1.56	.	75x1,30x4
chr22-49764913-49764941-AT	.	4	.	.	12.78	2.06	.	21x4
chr13-53008511-53008572-TTTCTTTTTCTTTC	.	4	.	.	5.13	1.68	.	14x1,13x2,11x1
chr4-94675900-94675943-TTTTTTTTCTTTTTTTT	.	3	ENSG00000249951	.	3.07	1.28	.	46x1,45x2
chr4-108691795-108692021-ATATATACACACATAT	.	3	.	.	3.92	2.13	.	14x3
chr3-124439200-124439237-CA	.	3	KALRN	.	18.47	2.49	.	45x1,32x2
chr5-123747207-123747236-GT	.	2	.	.	14.25	1.92	.	25x2
chr18-41577304-41577584-TATATATATATT	.	2	KC6	.	20.96	3.35	.	40x1,34x1
chr14-79569602-79569631-GT	.	2	NRXN3	.	15.18	0.93	.	21x2

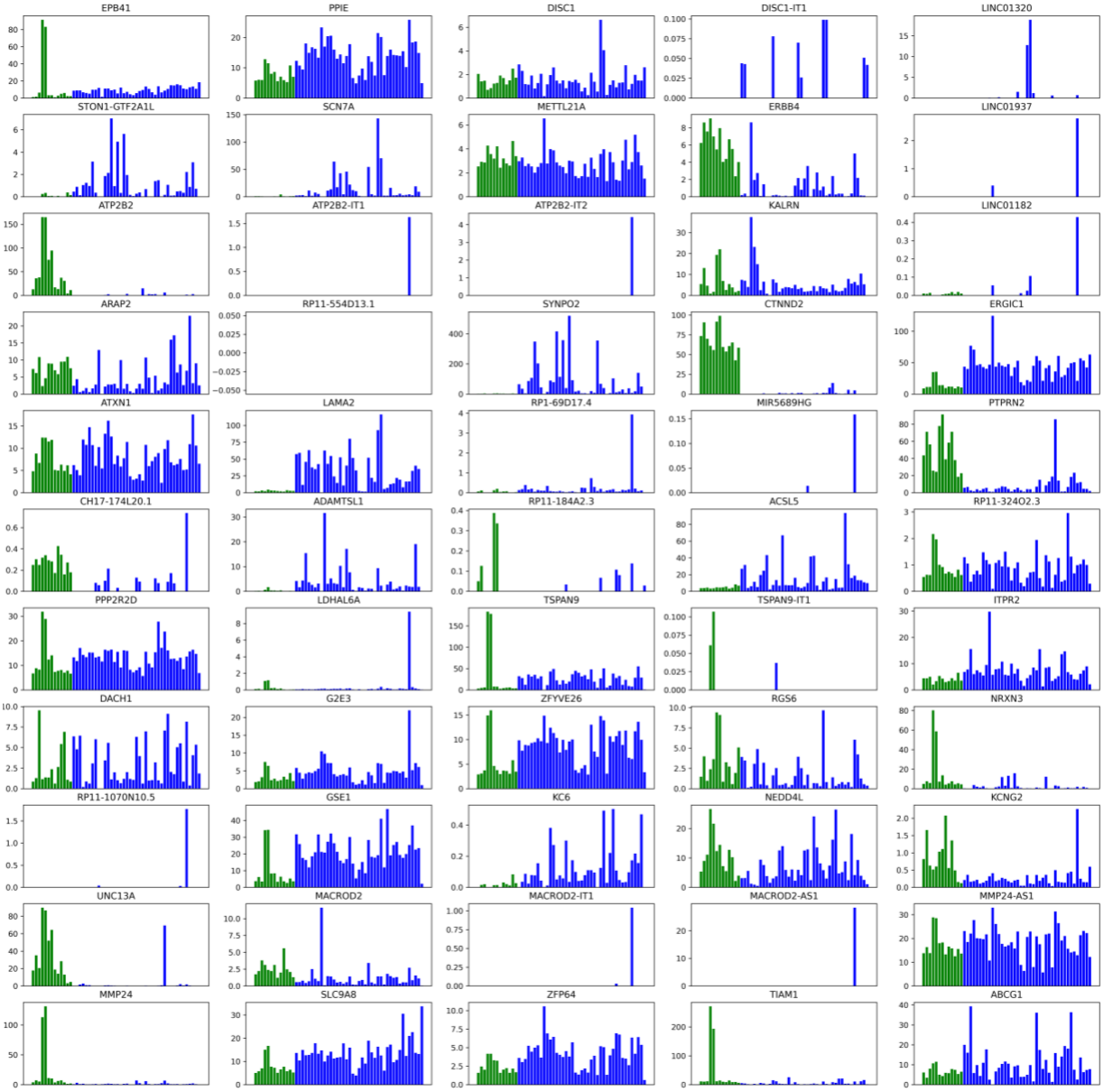

Figure 4-5 Gene expression levels (from GTEx) for genes with potential repeat expansions

Each histogram shows the gene expression levels in RPKM for different tissues. Bars showing gene expression levels for brain tissues are shown in green while bars for all other tissues are shown in blue.

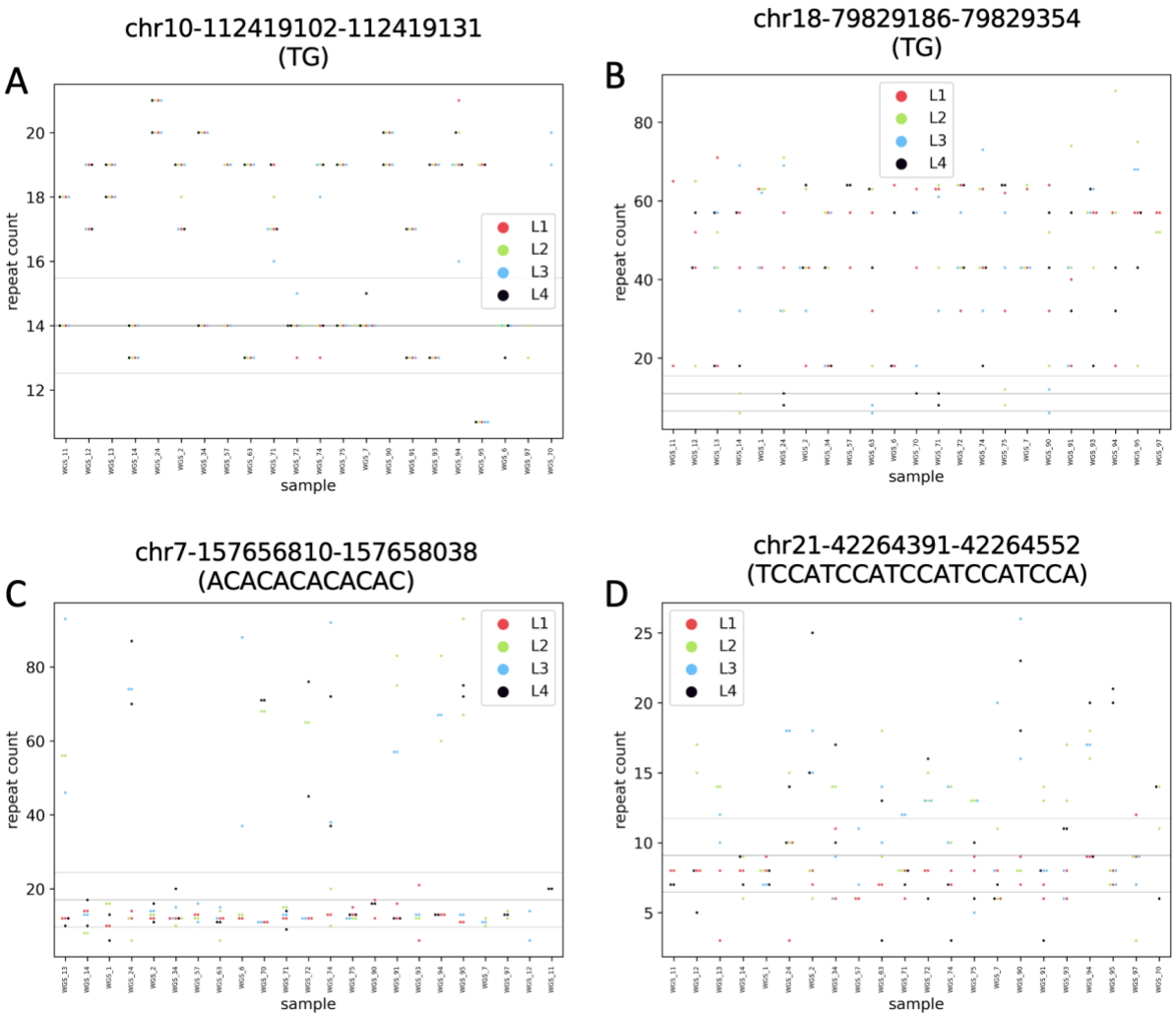

Figure 4-6 Representative dot plots of repeat estimates across different brain tissues in Mayo ALS samples.

Each dot plot shows the different repeat length estimates across the four brain tissues in each sample for a given repeat loci. The three lines on each plot show the mean (middle line) and standard deviation (outside lines) of the repeat lengths seen in the 1KG samples for the repeat loci in the plot. A) This dot plot is an example of a repeat loci that is variable between samples but is relatively stable within an individual. B) This dot plot is an example of a repeat loci that shows high variability between tissues. C) and D) The repeats in these two dot plots show variable repeat expansions in three tissues no repeat expansions are estimated in the fourth tissue.

Chapter 5 - Conclusion

5.1 SUMMARY

Studying genomic variation enables us to investigate differences across populations of individuals and uncover the causes of disease (2). Across the last 50 years, researchers have developed increasingly better technology and techniques to sequence the human genome and characterize the variation present across the genome. Despite these incredible strides, current methods do struggle to capture complex variations and highly repetitive regions of the genome. Additionally, with every technological advance in sequencing, there is an equal need for new bioinformatic approaches to be developed in order to gain insight from the data generated by these technologies.

Until we can sequence the entire genome end-to-end, we will only be able to assess genomic variation at the locations accessible given the technology. Further, researchers are often faced with a key choice, for there is a trade-off between the resolution of the genomic data generated and the cost of time and resources it takes to generate this data. Combining sequencing technologies can offer a balance by being cost efficient while also taking advantage of the benefits of higher resolution genomic data.

In this thesis, I have outlined three distinct ways multiple sequencing technologies can be combined to better understand the underlying variation in the human genome. Additionally, I employ multiple bioinformatic approaches to analyze the datasets produced by multiple

sequencing technologies. When viewed comprehensively, this thesis offers a series of case studies to illustrate how researchers can integrate multiple sequencing technologies and employ multiple bioinformatic approaches together in a multi-modal approach to detect genomic variation across both disease and diversity studies. The major conclusions of this dissertation are listed below, organized by main findings.

The use of multiple sequencing platforms provides a useful way to detect and characterize variation in a population

In Chapter 2, I present a genome wide survey of the genomic variation present in 97 individuals from Ukraine sequenced using BGI nanoball sequencing. BGI nanoball sequencing technology is less established than other short-read sequencing methods, and an outstanding question prior to this work was how variant detection using this technology might differ from other, more established technologies. Therefore, the variants detected from BGI sequencing data were compared to variation detected using SNP microarrays for 86 of the individuals and variation detected using Illumina WGS for one of the individuals. We found the variation detected using BGI WGS was comparable to the more established technologies. The genome wide variation dataset was used to establish the evolutionary differences of this population – Ukrainian individuals - compared to neighboring regions across Europe. Further, this work and the genome wide variation dataset can be used to identify variation specific to this population that may be relevant for identifying cause of diseases in Ukrainians.

Short and long read sequencing technologies are appropriate for detecting different classes of tandem repeat variation

In Chapter 3, I used both short and long read sequencing methods to detect tandem repeat variation in the human genome. These variations have been historically difficult to detect because of their repetitive nature. Many repetitive regions have not been included in previous versions of the reference genome, or if they have, are often removed from analysis pipelines because of the alignment of reads to these regions can often be unreliable. In particular, when using short read sequencing data alignment of reads to repetitive regions of the genome is fraught with errors and misalignments (6,227,228). The introduction of long read sequencing data has potential advantages for detecting tandem repeat variation specifically. In this chapter, I performed an in-depth analysis of different bioinformatic approaches for both short and long read sequencing data to evaluate their ability to detect tandem repeat variation genome wide. I found, somewhat unsurprisingly, short read methods have difficulty detecting repeat variation beyond the fragment length of 400 bp. This was especially apparent for reference-based methods while *de novo* detection methods could estimate and characterize larger repeat events even from short-read sequencing datasets.

Providing guidance on bioinformatic programs when using long- or short-read sequencing datasets

In Chapter 3, I additionally provide guidance about which bioinformatic program should be used when investigating variation in short or long read datasets. Specifically, *en lieu* of running both programs separately, bioinformaticians should run Expansion Hunter over GangSTR on short read data due to the large overlap in variation detected with these methods and the better concordance of EH with Sanger sequencing (our 'truth set'). For long read sequencing, the advice is not so clear-cut. Each bioinformatic method identified unique genomic variations; however, STRaglr

identified much larger repeat expansion lengths than the other two methods, and Tricolor and Tandem Genotypes methods had substantial overlap in variation identified. Perhaps, an ensemble approach is best when utilizing long-read sequencing data.

Short read sequencing data can inform tandem repeat estimation in SNP array datasets to enhance disease associated variation detection

Identifying causal genetic variations associated with ALS remains an important line of research as the majority of cases have no known genetic etiology (229). In Chapter 4, I used both short read WGS datasets and SNP microarrays to identify potential tandem repeat expansions associated with ALS. We hypothesized that a repeat expansion, not the associated SNP from a database of ALS GWAS hits, may be the causal variant. Thus, I focused our analysis on repeats around ALS GWAS hits. Because repetitive variation has been difficult to assess in the past, disease associated repeat expansions likely have gone undetected compared to other, more easily accessible variations. Additionally, I increased the number of disease samples beyond the available WGS data samples by imputing tandem repeat lengths for ALS samples included in the MGI biobank that have SNP microarray data. Importantly, we identified several repeat loci with lengths in multiple ALS samples that were outliers when compared to 1KG control population. Short reads, as discussed previously in Chapter 3, and SNP arrays are not entirely reliable for accurately determining the length of repetitive regions, therefore we suggest these repeat loci be candidates for targeted long read sequencing to obtain more accurate repeat length characterization. Only then can further research be performed to investigate their potential association with ALS.

5.2 FUTURE DIRECTIONS

Imputing tandem repeats for SNP array datasets

The analysis performed in Chapter 4 focused on identifying repeat expansions that may be associated with the disease ALS; however, the MGI database contains genetic and medical record data for many other diseases. The imputation of tandem repeat lengths can be performed for these individuals, specifically for neurodegenerative diseases as tandem repeat expansions have been implicated in many of these disease phenotypes (168,230–233). However, repeat lengths must be present in a reference panel for them to be imputed into a given sample. Thus, imputation of tandem repeats is greatly benefited by including disease samples in the reference panel. This requires access to WGS for the disease of interest, which can be difficult, depending on the disease of study. There are publicly available datasets such as the Answer ALS dataset of ALS or the Simon Simplex collection for Autism that could be used for this purpose. As discussed in Chapter 4, there are significant limitations to estimating tandem repeat lengths from SNPs and follow-up studies should be done to obtain more accurate repeat sizes for candidate repeat expansions. New methods show the utility of sequencing multiple known repeat regions at once using Oxford nanopore sequencing technology (234). The combination of SNP array data, WGS datasets and targeted long read sequencing can provide an avenue for a cost-efficient way to identify repeat expansions associated with disease.

Advancements of long read sequencing and tandem repeat detection

As shown in Chapter 3, the use of long read sequencing provides the ability to characterize repeat variation genome wide better than what could previously be done using short read sequencing.

However, WGS long read sequencing is currently less common than short read sequencing due to cost. Undoubtedly, as the technology is optimized, long read sequencing will likely become more accessible for researchers across different institutes and will become relatively more accurate compared to today's technology. These eventual advancements will no doubt enable a broader characterization of large complex variations, including tandem repeats. As the technology progresses, long read sequencing will allow complex types of variation to be integrated into large disease association and evolutionary studies in the future.

As an example of its utility, the use of long read sequencing has recently been used to create the first complete, gapless genome (235). Before the release of this complete genome earlier this year, there have been regions of the genome where variation has been impossible to characterize because they were not included in the reference genome. These regions were mostly made-up of repetitive sequences in the telomeric and centromeric regions (236).

Integrating these new sequences into the reference genome is a large advancement for the field. The inclusion of these sequences allows for the characterization of variation in these regions across individuals. It is highly likely that there is extensive structural variation hidden within these previously dark regions of the genome and studies can now, for the first time, shed light on how variations in these regions might impact our diversity and our health. Both long- and short-read sequencing datasets can be aligned to the new regions to comprehensively evaluate potential associations between disease and variations within these newly identified regions of the genome. Further development of bioinformatic tools may be needed to compare and characterize newly identified variations in the context of older association studies.

Detecting variation within an individual in addition to across individuals

Most of the analysis performed for this thesis involved characterizing germline variation by analyzing bulk sequencing for samples from a single tissue. While this type of analysis has resulted in abundant new knowledge about the variation present in the human species, many new avenues of research involve looking at the somatic variation present within an individual. The tissue from which DNA is extracted can affect the ability to identify disease associated variation. Examining and analyzing the genetic differences between tissues in the same individual can help elucidate mechanisms and progression of a disease. Characterizing the relative variation of tandem repeats within an individual could be particularly interesting - given the high mutation rate of this type of variation (58,237–239), it is likely that repetitive sequences will vary more across cells than other types of genetic variation, such as SNPs.

On a more granular level, single-cell sequencing can uncover genetic variation within the same tissue. Since its introduction, single cell sequencing has revolutionized multiple biomedical and basic biological fields by introducing a resolution that had previously been unseen and embracing heterogeneity across samples (240,241). Characterizing SNPs in cells of the same phenotype has already shown that some cells may exhibit the SNP but others will not. Single cell sequencing will additionally support better understanding of tandem repeats as well. Our work in Chapter 3, analyzing the output from the single molecule long read sequencing, provides an opportunity for further analysis, using single cell sequencing approaches. Many of the bioinformatic methods we employed (Tri-Color, Tandem Genotypes, STRaglr) reported the estimated repeat length for each read covering the repeat loci. Often, the bioinformatic tools reported more than two repeat lengths,

the maximum number expected. In our analysis, we treated this as sequencing error and clustered them into two alleles. However, in some cases, this may actually be indicative of somatic variation within an individual. This type of variation can be explored and further characterized by using single cell sequencing approaches.

Classifying variation across multiple tissues and multiple cells will also require the development of accompanying bioinformatics approaches to be built on top of the current detection methods. The amplification necessary for single cell analysis introduces errors that will present challenges to variant calling, such as handling missing data or allelic dropout (242–246). As this thesis has shown the utility of integrating various technologies and methods for analysis, single cell analysis can also be integrated with other sequencing technologies to provide additional insight when detecting different types of variation across multiple tissues and cells. Finally, as with any developing technology, optimizing compute times for single-cell analysis represents an open challenge.

Integrating diverse genomes in medical research

As an umbrella over all the future work that has been discussed, research identifying genomic variation must be performed with diversity in mind. As the use of genomics by both scientists and clinicians to achieve personalized medicine becomes more and more prevalent, it is imperative that diverse genomes are included in the research that supports this practice to ensure these practices are applicable, safe, and effective for everyone. Initiatives such as the 1000 Genomes Project (24) and gnomAD (247) have greatly increased the diversity of publicly available genomes

with the purpose of providing a resource for the variation present in the human species. However, there are still populations which are under-represented in these databases.

Further, there is a stark lack of representation within the reference genome. While its well-known that the reference genome should not be thought of as a 'universal genome', approximately 70% of the reference genome came from a single sample (248). Further, the reference genome has its own biases, demonstrating high risk of disorders like type 1 diabetes, among others (249). As such, any comparisons of variation to the reference genome in a disease context must account for reference bias. For example, since rare alleles could be present in the reference genome, potential pathogenic variants could be incorrectly unidentified (250).

Further complicating lack of diversity issues within the reference genome, GWAS also have poor representation from specific populations. In 2018, approximately 78 percent of individuals included in GWAS comes from people of European descent, but only 16 percent of the global population is of European descent (35). (While this disparity is unacceptable, it should be noted that prior to 2009, 96% of individuals that participated in GWAS were of European descent (251). This remarkable and persistent sampling bias of large-scale GWAS is perhaps a critical reason why GWAS hits do not always translate to clinically relevant treatment options for the world. Undoubtedly, by sampling from primarily a population of individuals of shared descent, researchers have perpetuated the health disparities that are already present in global medicine. For example, the strength of association differed in non-European populations for 25% of the variants GWAS identified as being associated with type 2 diabetes (252).

That said, projects like the 1000 genomes project (24) and gnomAD (30) are excellent first steps to providing reference genomes across various populations worldwide. In gnomAD's paper on structural variation, their work included samples where the majority (54%) were of non-European ancestry. They found that samples of African ancestry exhibited the greatest genetic diversity and that East Asian samples featured the highest levels of homozygosity across samples (247). This finding underscores the importance of further characterizing all populations in the world, but certainly major efforts need to be made to understand and identify variation across African populations.

As diversity becomes better represented within and across these databases and studies, that diversity and variation can be better captured in the form of graph reference genomes (as opposed to a linear reference genome) for various populations which allow for alternate loci to be included in the reference genome. These graph reference genomes will provide a much-needed update to the field's gold standard for comparison and variant identification, the reference genome.

While further investigation of European ancestry may not be the most pressing need within genomic variation studies, our work in Chapter 2 was still the first characterization of the Ukrainian population, representing approximately 50 million individuals whose ancestry could be better understood. While 97 samples are by no means representative of the full genomic diversity present in the Ukrainian population, the genomic variant dataset generated in Chapter 2 of this thesis is one that can be integrated into future medical and evolutionary studies. Increasing the number of individuals sequenced from Ukraine or other populations which have not been included in previous

large sequencing initiatives (such as those from African or Latin American ancestry) will provide important data for future studies.

In sum, greater diversity within GWAS and other genomic variation studies have been shown to provide medical benefits and greater biomedical insight. In fact, a recent study quantified the importance of diversity in just such studies and provided evidence that increasing diversity rather than studying additional individuals of European ancestry results in substantial improvements in mapping variants (253). Greater efforts to incorporate diversity in studies of genome variation will be key to identifying the next set of causal variations that lead to clinically relevant findings and eventual downstream medical discoveries.

Appendices
Appendix A Supporting Information for Chapter 2

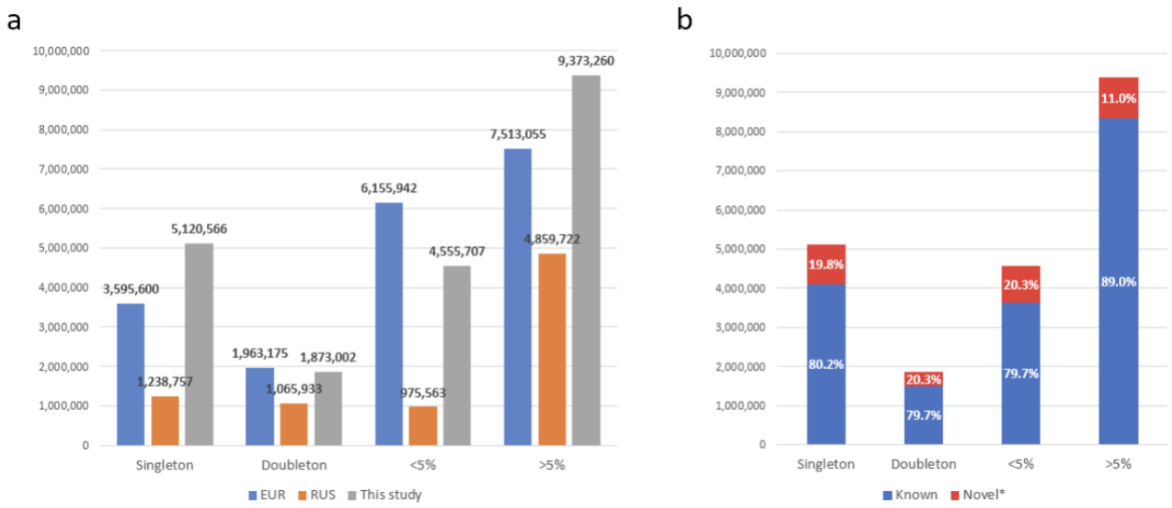

Figure A-1 Frequencies of various classes of SNPs in the Ukrainian genome variation database.

Definitions are as follows: Singleton (passed the GATK QC once), Doubleton, Rare (3-10 counts roughly equivalent to 1%< x < 5%) and Common (>5%) to make it closer to the 1KGP definitions. B. Percent novel mutations in various classes of SNPs

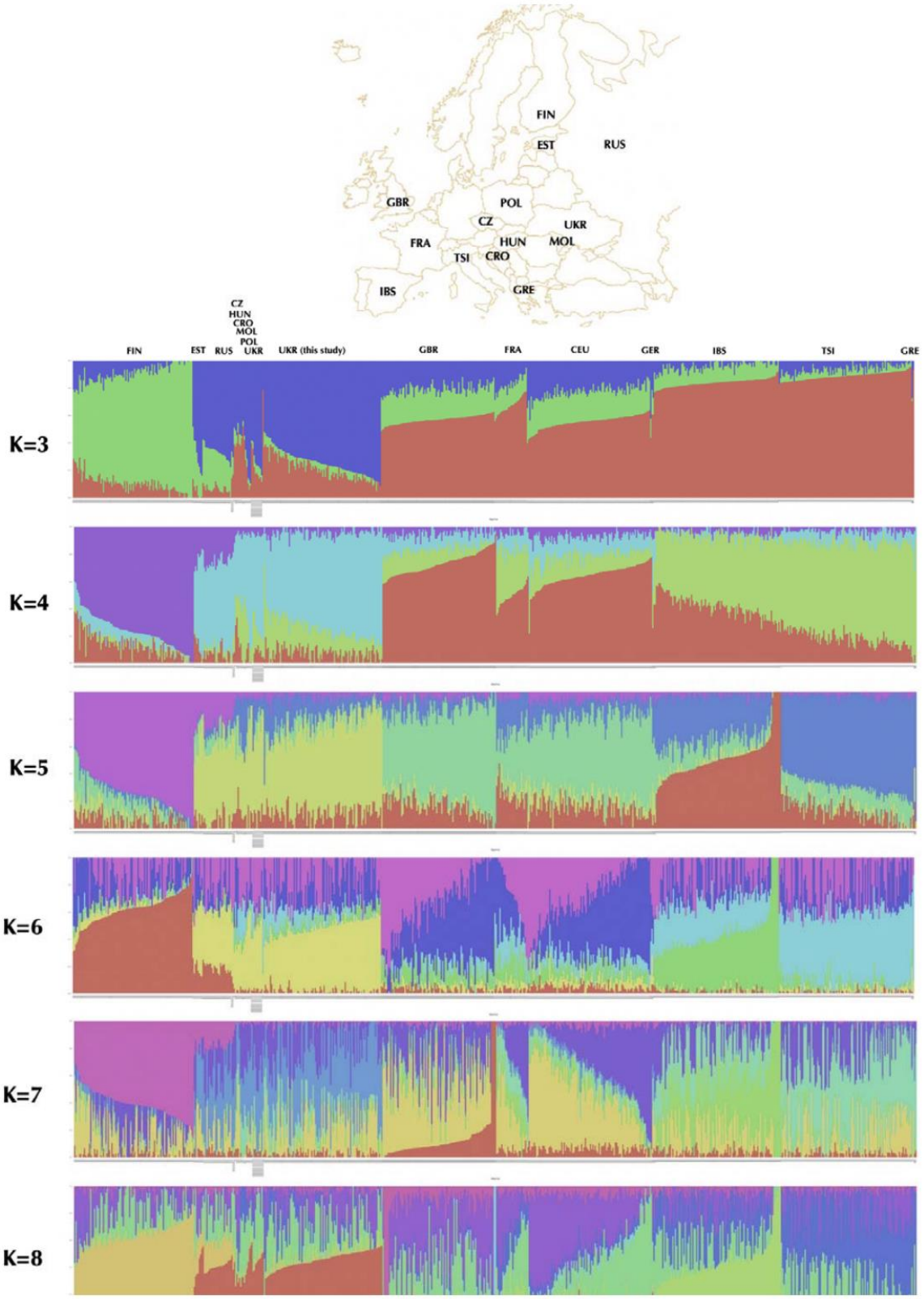

Figure A-2 Genetic structure of Ukrainian population in comparison to other European populations.

For identification of the optimal K parameter, we used the 10-fold cross-validation function of ADMIXTURE in range from 2 to 8, with K=3 resulting in the lowest error. This analysis included genomes from this study as well as samples from the 1000Genomes Project (Utah Residents (CEU) with Northern and Western European Ancestry, Toscani in Italy (TSI), Finnish in Finland (FIN), British in England and Scotland (GBR), and Iberian Population in Spain (IBS)

French(FRA) and Russians (RUS) from HGDP [39], as well as the relevant high-coverage human genomes Croatia (CRO), Czech (CZ), Estonian (EST), German (GER), Greek (GRE), Hungarian (HUN), Moldovan (MOL), Polish (POL), Russian Cossack (RUS) and Ukrainian (UKR) from the Estonian Biocentre Human Genome Diversity Panel (EGDP) as well as Simmons Genome Diversity project.

Appendix B Supporting Information for Chapter 3

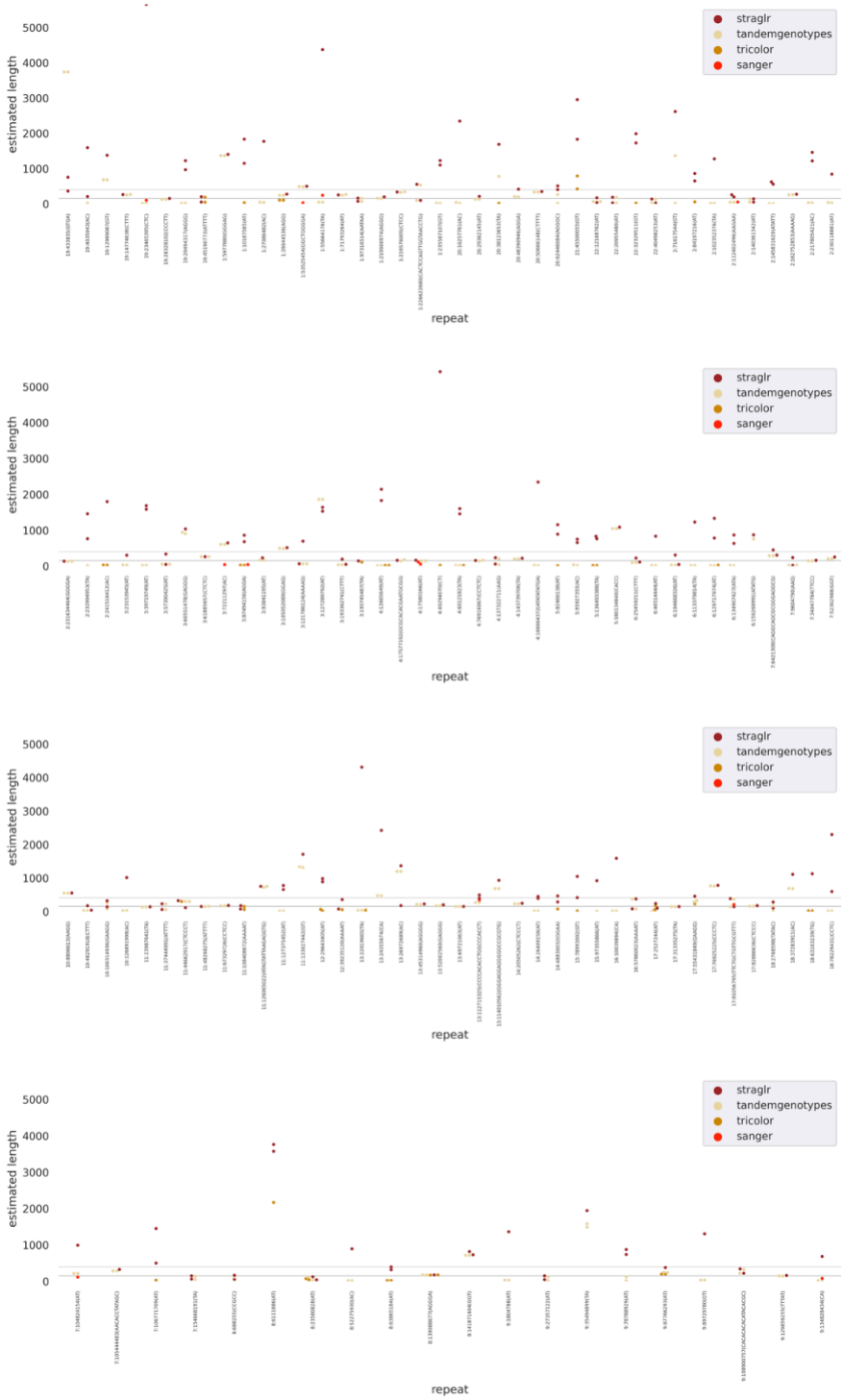

Figure B-1 Dot plots of estimated repeat lengths for repeat loci characterized by long read methods STRaglr (dark red) and Tandem Genotypes - yellow, Tricolor – orange)

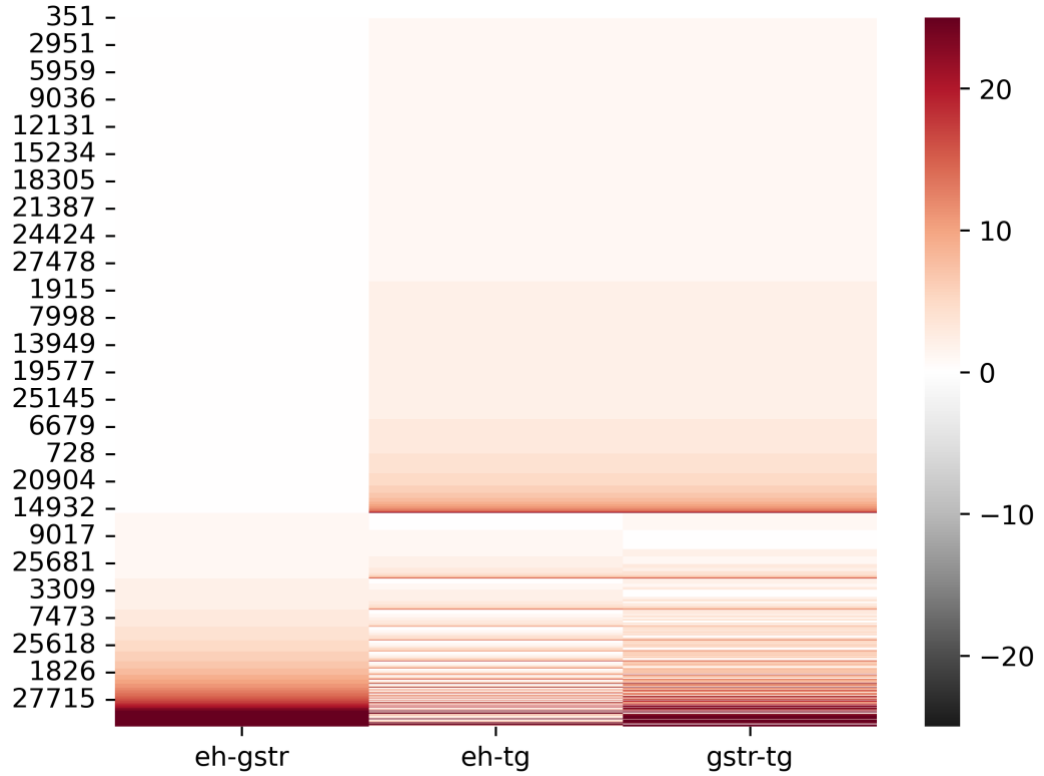

Figure B-2 Heatmap of similarity metric comparisons between GangSTR (gstr), Expansion Hunter (eh) and Tandem Genotypes (tg).

Appendix C Supporting Information for Chapter 4

Study and Citation		**SNP IDs**
(Chen CJ,2015)	www.ncbi.nlm.nih.gov/pubmed/26580837	rs2185341,rs3772760,rs4879628,rs11224052,rs2785946,rs9953769
(Cronin S, 2007)	www.ncbi.nlm.nih.gov/pubmed/18057069	rs10260404
(Cronin S, 2008)	www.ncbi.nlm.nih.gov/pubmed/18987618	rs10260404
(Laaksovirta H, 2010)	www.ncbi.nlm.nih.gov/pubmed/20801718	rs3849942,rs13048019
(Kwee LC, 2012)	www.ncbi.nlm.nih.gov/pubmed/22470424	rs2278170,rs3113494
(McLaughlin RL, 2014)	www.ncbi.nlm.nih.gov/pubmed/25442119	rs7019351
(Xie T, 2014)	www.ncbi.nlm.nih.gov/pubmed/24529757	rs982274,rs4824093,rs2972219,rs2457174,rs1981626,rs2492937,rs4424056,rs8192851,rs10501765,rs4761659,rs7999075,rs1605070,rs26850 56,rs11744876,rs9327881,rs9329300,rs7117082,rs9977018,rs1198775 8,rs11062578,rs7899101,rs3798696,rs17162257,rs1572511,rs1075428 3,rs6531209,rs3749146,rs1400816,rs6722486,rs9599848,rs11590421,r s7601234,rs1199333,rs4148112,rs4482178,rs12891047,rs10131300,rs1 0145110,rs16945894,rs730547,rs16975050,rs529445,rs6137726,rs480 9847,rs1464443,rs1559473,rs11921451,rs13100616,rs3852053,rs4128 705,rs6851442,rs2667100,rs10495822,rs4234080,rs1987842,rs767921 8,rs13133845,rs10029851,rs7698598,rs7148498,rs4240810,rs320637,r s7224488,rs7267421,rs11167260,rs9825420,rs7755729,rs4719220,rs13 256095,rs7830371,rs16938145,rs17684824,rs10508264,rs17603886,rs 551585,rs2985334,rs10489764,rs2247208,rs16902328,rs5029317,rs19 29412,rs776776,rs10982990,rs1002442,rs10458771,rs3740713,rs2322 978,rs3782455,rs1147246,rs11061269,rs9568797
(Wei L, 2019)	www.ncbi.nlm.nih.gov/pubmed/31872054	rs12145183,rs1419311,rs1483023,rs9610216,rs1722923,rs6703183,rs8 141797
(Schymick JC, 2007)	www.ncbi.nlm.nih.gov/pubmed/17362836	rs16984239,rs11099864,rs12680546,rs6013382,rs2782931
(van Es MA, 2007)	www.ncbi.nlm.nih.gov/pubmed/17827064	rs2306677
(Dekker AM, 2019)	www.ncbi.nlm.nih.gov/pubmed/30976013	rs3849942
(van Es MA, 2007)	www.ncbi.nlm.nih.gov/pubmed/18084291	rs10260404,rs3825776,rs7580332
(Shatunov A, 2010)	www.ncbi.nlm.nih.gov/pubmed/20801717	rs4799088,rs1488902,rs10122902; rs3849942,rs10122902; rs3849942
(Goris A, 2013)	www.ncbi.nlm.nih.gov/pubmed/24234648	rs2935183
(Landers JE, 2009)	www.ncbi.nlm.nih.gov/pubmed/19451621	rs16856202,rs873917,rs7577894,rs10438933,rs8066857,rs697739,rs31 77980,rs2823962,rs1541160,rs3099950,rs11241713,rs855913,rs77020 57,rs13015447,rs10192369
(Diekstra FP, 2014)	www.ncbi.nlm.nih.gov/pubmed/24931836	rs3849943,rs12608932,rs13268726,rs7477,rs7638688,rs10233425,rs12 546767
(Nakamura R, 2020)	www.ncbi.nlm.nih.gov/pubmed/32968195	rs3736947

(Ahmeti KB, 2012)	www.ncbi.nlm.nih.gov/pubmed/22959728	rs2364403,rs2819332,rs669446,rs1421746,rs10503672,rs1491818,rs2904524,rs7147705,rs3011225,rs12651329,rs6956741,rs7047865,rs2036225,rs2303565,rs7607369,rs1320900,rs7665939,rs6918777,rs4917300,rs3849942,rs4529888,rs9533799,rs1971791,rs8056742,rs7477,rs2006933,rs11082762,rs12608932,rs1923626,rs7729723,rs4913250,rs2838568,rs524675,rs42714,rs11096913,rs1320900,rs12608932,rs4877387,rs1510510,rs10870270,rs2199351
(van Es MA, 2009)	www.ncbi.nlm.nih.gov/pubmed/19734901	rs2405657,rs3849942,rs9971637,rs774359,rs5916687,rs5937496,rs2814707,rs12608932
(Deng M, 2013)	www.ncbi.nlm.nih.gov/pubmed/23624525	rs6703183,rs8141797
(Fogh I, 2016)	www.ncbi.nlm.nih.gov/pubmed/27244217	rs139550538,rs2412208,rs969599,rs72911847,rs75285952,rs115134572
(Fogh I, 2013)	www.ncbi.nlm.nih.gov/pubmed/24256812	rs12608932,rs3849942,rs34517613,rs1788776
(Benyamin B, 2017)	www.ncbi.nlm.nih.gov/pubmed/28931804	rs12608932,rs3849943,rs35714695,rs616147
(Nicolas A, 2018)	www.ncbi.nlm.nih.gov/pubmed/29566793	rs10463311,rs3849943,rs74654358,rs12973192,rs75087725,rs113247976,rs17070492,rs10139154,rs10143310,rs9901522
(van Rheenen W, 2016)	www.ncbi.nlm.nih.gov/pubmed/27455348	rs75087725,rs616147,rs10139154,rs74654358,rs3849943,rs12608932,rs35714695
(van Rheenen W, 2021)	www.ncbi.nlm.nih.gov/pubmed/34873335	rs62333164,rs2985994,rs10280711,rs12608932,rs80265967,rs229195,rs229194,rs631312,rs9275477,rs113247976,rs75087725,rs10463311,rs17785991,rs4075094,rs517339

Table C-1 Table of ALS GWAS SNPs used for detection potential repeat expansions associated with ALS

Bibliography

1. Collins FS, Mansoura MK. The Human Genome Project. Cancer. 2001;91(S1):221–5.

2. Benton ML, Abraham A, LaBella AL, Abbot P, Rokas A, Capra JA. The influence of evolutionary history on human health and disease. Nat Rev Genet. 2021 May;22(5):269–83.

3. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016 Jan;107(1):1–8.

4. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature [Internet]. 2001 Feb 15 [cited 2021 Sep 23]; Available from: http://deepblue.lib.umich.edu/handle/2027.42/62798

5. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. 2020 Mar;21(3):171–89.

6. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012 Jan;13(1):36–46.

7. Bansal V, Boucher C. Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going? iScience. 2019 Aug 30;18:37–41.

8. Osman N, Shawky AEM, Brylinski M. Exploring the effects of genetic variation on gene regulation in cancer in the context of 3D genome structure. BMC Genomic Data. 2022 Feb 17;23(1):13.

9. Frequently Asked Questions - Genome Reference Consortium [Internet]. [cited 2022 Jul 28]. Available from: https://www.ncbi.nlm.nih.gov/grc/help/faq/

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

11. Gibbs RA. The Human Genome Project changed everything. Nat Rev Genet. 2020 Oct;21(10):575–6.

12. Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet. 2003 Mar;33(3):266–75.

13. Eichler EE. Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. N Engl J Med. 2019 Jul 4;381(1):64–74.

14. Kay MA. State-of-the-art gene-based therapies: the road ahead. Nat Rev Genet. 2011 May;12(5):316–28.

15. Cooke Bailey JN, Bush WS, Crawford DC. Editorial: The Importance of Diversity in Precision Medicine Research. Front Genet [Internet]. 2020 [cited 2022 Apr 25];11. Available from: https://www.frontiersin.org/article/10.3389/fgene.2020.00875

16. International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789–96.

17. Christensen K, Murray JC. What genome-wide association studies can do for medicine. N Engl J Med. 2007 Mar 15;356(11):1094–7.

18. Ke X, Cardon LR. Efficient selective screening of haplotype tag SNPs. Bioinformatics. 2003 Jan 22;19(2):287–8.

19. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep 2;467(7311):52–8.

20. Deloukas P, Bentley D. The HapMap project and its application to genetic studies of drug response. Pharmacogenomics J. 2004 Apr;4(2):88–90.

21. Hua F, Guo Y, Sun Q, Yang L, Gao F. HapMap-based study: CYP2A13 may be a potential key metabolic enzyme gene in the carcinogenesis of lung cancer in non-smokers. Thorac Cancer. 2019;10(4):601–6.

22. McVean G, Spencer CCA, Chaix R. Perspectives on Human Genetic Variation from the HapMap Project. PLOS Genet. 2005 Oct 28;1(4):e54.

23. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, et al. The 1000 Genomes Project: data management and community access. Nat Methods. 2012 May;9(5):459–62.

24. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

25. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios [Internet]. bioRxiv; 2021 [cited 2022 May 7]. p. 2021.02.06.430068. Available from: https://www.biorxiv.org/content/10.1101/2021.02.06.430068v2

26. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021 Feb;590(7845):290–9.

27. Baxi EG, Thompson T, Li J, Kaye JA, Lim RG, Wu J, et al. Answer ALS, a large-scale resource for sporadic and familial ALS combining clinical and multi-omics data from induced pluripotent cell lines. Nat Neurosci. 2022 Feb;25(2):226–37.

28. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. Mol Psychiatry. 2020 Aug;25(8):1859–75.

29. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The genetic architecture of type 2 diabetes. Nature. 2016 Aug;536(7614):41–7.

30. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.

31. Sherry ST, Ward M, Sirotkin K. dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. Genome Res. 1999 Aug 1;9(8):677–9.

32. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, et al. DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. 2013 Jan;41(Database issue):D936-941.

33. 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. 2018 Jan 4;46(D1):D1062–7.

34. Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, et al. On the Use of General Control Samples for Genome-wide Association Studies: Genetic Matching Highlights Causal Variants. Am J Hum Genet. 2008 Feb 8;82(2):453–63.

35. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019 Mar 21;177(1):26–31.

36. Suzuki Y, Morishita S. The time is ripe to investigate human centromeres by long-read sequencing†. DNA Res Int J Rapid Publ Rep Genes Genomes. 2021 Oct 11;28(6):dsab021.

37. Shastry BS. SNP alleles in human disease and evolution. J Hum Genet. 2002 Nov;47(11):561–6.

38. Harris Kelley. Evidence for recent, population-specific evolution of the human mutation rate. Proc Natl Acad Sci. 2015 Mar 17;112(11):3439–44.

39. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012 Aug;488(7412):471–5.

40. Durbin RM, Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct;467(7319):1061–73.

41. Petersen GM, Parmigiani G, Thomas D. Missense Mutations in Disease Genes: A Bayesian Approach to Evaluate Causality. Am J Hum Genet. 1998 Jun 1;62(6):1516–24.

42. Pauling L, Itano HA, Singer SJ, Wells IC. Sickle Cell Anemia, a Molecular Disease. Science. 1949 Nov 25;110(2865):543–8.

43. Hunt JA, Ingram VM. Allelomorphism and the Chemical Differences of the Human Hæmoglobins A, S and C. Nature. 1958 Apr;181(4615):1062–3.

44. Rodriguez-Murillo L, Salem RM. Insertion/Deletion Polymorphism. In: Gellman MD, Turner JR, editors. Encyclopedia of Behavioral Medicine [Internet]. New York, NY: Springer; 2013 [cited 2022 May 8]. p. 1076–1076. Available from: https://doi.org/10.1007/978-1-4419-1005-9_706

45. Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, et al. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res. 2011 Jun 1;21(6):830–9.

46. Collins FS, Drumm ML, Cole JL, Lockwood WK, Vande Woude GF, Iannuzzi MC. Construction of a general human chromosome jumping library, with application to cystic fibrosis. Science. 1987 Feb 27;235(4792):1046–9.

47. Vogels A, Fryns JP. The Prader-Willi syndrome and the Angelman syndrome. Genet Couns Geneva Switz. 2002;13(4):385–96.

48. Lakich D, Kazazian HH, Antonarakis SE, Gitschier J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. Nat Genet. 1993 Nov;5(3):236–41.

49. Bunting SF, Nussenzweig A. End-joining, translocations and cancer. Nat Rev Cancer. 2013 Jul;13(7):443–54.

50. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer. 2007 Apr;7(4):233–45.

51. Rabbitts TH. Commonality but Diversity in Cancer Gene Fusions. Cell. 2009 May 1;137(3):391–5.

52. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005 Oct 28;310(5748):644–8.

53. Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012 Jul;487(7407):330–7.

54. Hazkani-Covo E, Zeller RM, Martin W. Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. PLOS Genet. 2010 Feb 12;6(2):e1000834.

55. Frequency of recent retrotransposition events in the human factor IX gene - Li - 2001 - Human Mutation - Wiley Online Library [Internet]. [cited 2022 Apr 25]. Available from: https://onlinelibrary.wiley.com/doi/10.1002/humu.1134

56. Replication slippage involves DNA polymerase pausing and dissociation. EMBO J. 2001 May 15;20(10):2587–95.

57. Trang H, Stanley SY, Thorner P, Faghfoury H, Schulze A, Hawkins C, et al. Massive CAG Repeat Expansion and Somatic Instability in Maternally Transmitted Infantile Spinocerebellar Ataxia Type 7. JAMA Neurol. 2015 Feb 1;72(2):219–23.

58. Fan H, Chu JY. A Brief Review of Short Tandem Repeat Mutation. Genomics Proteomics Bioinformatics. 2007 Jan 1;5(1):7–14.

59. Lee JM, Ramos EM, Lee JH, Gillis T, Mysore JS, Hayden MR, et al. CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. Neurology. 2012 Mar 6;78(10):690–5.

60. Crawford DC, Acuña JM, Sherman SL. FMR1 and the fragile X syndrome: Human genome epidemiology review. Genet Med. 2001 Sep 1;3(5):359–71.

61. Syvänen AC. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet. 2001 Dec;2(12):930–42.

62. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977 Dec;74(12):5463–7.

63. Rohland N, Reich D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res. 2012 May 1;22(5):939–46.

64. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017 Oct;550(7676):345–53.

65. Pfeifer SP. From next-generation resequencing reads to a high-quality variant data set. Heredity. 2017 Feb;118(2):111–24.

66. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 2009 Nov;6(11):S13–20.

67. Beck TF, Mullikin JC, Biesecker LG. Systematic Evaluation of Sanger Validation of NextGen Sequencing Variants. Clin Chem. 2016 Apr;62(4):647–54.

68. Goldfeder RL, Wall DP, Khoury MJ, Ioannidis JPA, Ashley EA. Human Genome Sequencing at the Population Scale: A Primer on High-Throughput DNA Sequencing and Analysis. Am J Epidemiol. 2017 Oct 15;186(8):1000–9.

69. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016 Jun;17(6):333–51.

70. Huang J, Liang X, Xuan Y, Geng C, Li Y, Lu H, et al. A reference human genome dataset of the BGISEQ-500 sequencer. GigaScience. 2017 Apr 1;6(5):1–9.

71. Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, et al. Targeted long-read sequencing identifies missing disease-causing variation. Am J Hum Genet. 2021 Aug 5;108(8):1436–49.

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

73. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. Hum Mol Genet. 2018 Aug 1;27(R2):R234–41.

74. Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H. Benchmarking of long-read correction methods. NAR Genomics Bioinforma. 2020 Jun 1;2(2):lqaa037.

75. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015 Oct 1;13(5):278–89.

76. Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, et al. Sequencing of human genomes with nanopore technology. Nat Commun. 2019 Apr 23;10(1):1869.

77. Minimap2: pairwise alignment for nucleotide sequences | Bioinformatics | Oxford Academic [Internet]. [cited 2022 Apr 25]. Available from: https://academic.oup.com/bioinformatics/article/34/18/3094/4994778

78. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

79. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012 Sep 15;28(18):i333–9.

80. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014 Jun 26;15(6):R84.

81. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinforma Oxf Engl. 2016 Apr 15;32(8):1220–2.

82. Ahsan MU, Liu Q, Fang L, Wang K. NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks. Genome Biol. 2021 Sep 6;22(1):261.

83. English AC, Salerno WJ, Reid JG. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics. 2014 Jun 10;15(1):180.

84. PacificBiosciences/pbsv [Internet]. PacBio; 2022 [cited 2022 Apr 25]. Available from: https://github.com/PacificBiosciences/pbsv

85. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single molecule sequencing. Nat Methods. 2018 Jun;15(6):461–8.

86. Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Pittard WS, et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. 2017 Nov 1;27(11):1916–29.

87. Dayama G, Emery SB, Kidd JM, Mills RE. The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res. 2014 Nov 10;42(20):12640–9.

88. Zhou W, Emery SB, Flasch DA, Wang Y, Kwan KY, Kidd JM, et al. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. Nucleic Acids Res. 2020 Feb 20;48(3):1146–63.

89. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. 2019 Sep 5;47(15):e90.

90. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019 Nov 1;35(22):4754–6.

91. Chiu R, Rajan-Babu IS, Friedman JM, Birol I. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. Genome Biol. 2021 Aug 13;22(1):224.

92. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. Genome Biol. 2019 Dec;20(1):58.

93. Bolognini D, Magi A, Benes V, Korbel JO, Rausch T. TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data. GigaScience. 2020 Oct 7;9(10):giaa101.

94. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019 Apr 16;10(1):1784.

95. Teo SM, Pawitan Y, Kumar V, Thalamuthu A, Seielstad M, Chia KS, et al. Multi-platform segmentation for joint detection of copy number variants. Bioinformatics. 2011 Jun 1;27(11):1555–61.

96. van Belzen IAEM, Schönhuth A, Kemmeren P, Hehir-Kwa JY. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. Npj Precis Oncol. 2021 Mar 2;5(1):1–11.

97. Moldován N, Tombácz D, Szűcs A, Csabai Z, Snyder M, Boldogkői Z. Multi-Platform Sequencing Approach Reveals a Novel Transcriptome Profile in Pseudorabies Virus. Front Microbiol [Internet]. 2018 [cited 2022 May 9];8. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2017.02708

98. Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, et al. Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet. 2016 Jan;24(1):2–5.

99. Zhao X, Weber AM, Mills RE. A recurrence-based approach for validating structural variation using long-read sequencing technology. GigaScience. 2017 Aug 1;6(8):1–9.

100. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018 Sep 6;103(3):338–48.

101. Goldstein DB, Weale ME. Population genomics: Linkage disequilibrium holds the key. Curr Biol. 2001 Jul 24;11(14):R576–9.

102. Palmer LJ, Cardon LR. Shaking the tree: mapping complex disease genes with linkage disequilibrium. The Lancet. 2005 Oct 1;366(9492):1223–34.

103. Subtelny O. Ukraine: A History, Fourth Edition. University of Toronto Press; 2009. 805 p.

104. Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, et al. The genomic history of southeastern Europe. Nature. 2018 Mar;555(7695):197–203.

105. Warmuth V, Eriksson A, Bower MA, Barker G, Barrett E, Hanks BK, et al. Reconstructing the origin and spread of horse domestication in the Eurasian steppe. Proc Natl Acad Sci. 2012 May 22;109(21):8202–6.

106. Schubert M, Jónsson H, Chang D, Der Sarkissian C, Ermini L, Ginolhac A, et al. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. Proc Natl Acad Sci. 2014 Dec 30;111(52):E5661–9.

107. Gaunitz C, Fages A, Hanghøj K, Albrechtsen A, Khan N, Schubert M, et al. Ancient genomes revisit the ancestry of domestic and Przewalski's horses. Science. 2018 Apr 6;360(6384):111–4.

108. Librado P, Fages A, Gaunitz C, Leonardi M, Wagner S, Khan N, et al. The Evolutionary Origin and Genetic Makeup of Domestic Horses. Genetics. 2016 Oct 1;204(2):423–34.

109. Demay L, Péan S, Patou-Mathis M. Mammoths used as food and building resources by Neanderthals: Zooarchaeological study applied to layer 4, Molodova I (Ukraine). Quat Int. 2012 Oct 25;276–277:212–26.

110. Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas AS, Manica A, Moltke I, et al. Genomic structure in Europeans dating back at least 36,200 years. Science. 2014 Nov 28;346(6213):1113–8.

111. Eberhardt P, Owsinski J. Ethnic Groups and Population Changes in Twentieth-Century Central-Eastern Europe: History, Data, and Analysis. New York: Routledge; 2016.

112. Oleksyk TK, Brukhin V, O'Brien SJ. The Genome Russia project: closing the largest remaining omission on the world Genome map. GigaScience [Internet]. 2015 Dec 1 [cited 2021 Sep 23];4(1). Available from: https://doi.org/10.1186/s13742-015-0095-0

113. Oleksyk TK, Wolfsberger WW, Weber AM, Shchubelka K, Oleksyk OT, Levchuk O, et al. Genome diversity in Ukraine. GigaScience [Internet]. 2021 Jan 29 [cited 2021 Sep 23];10(1). Available from: https://doi.org/10.1093/gigascience/giaa159

114. Zhernakova DV, Brukhin V, Malov S, Oleksyk TK, Koepfli KP, Zhuk A, et al. Genome-wide sequence analyses of ethnic populations across Russia. Genomics. 2020 Jan 1;112(1):442–58.

115. Kim J, Weber JA, Jho S, Jang J, Jun J, Cho YS, et al. KoVariome: Korean National Standard Reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses. Sci Rep. 2018 Apr 4;8(1):5677.

116. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017 May 1;27(5):849–64.

117. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018 Nov 1;28(11):1747–56.

118. Campbell IM, Gambin T, Jhangiani SN, Grove ML, Veeraraghavan N, Muzny DM, et al. Multiallelic Positions in the Human Genome: Challenges for Genetic Analyses. Hum Mutat. 2016;37(3):231–4.

119. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 2019 Jun 3;20(1):117.

120. K Y, G H. Structural Variation Detection from Next Generation Sequencing. J Gener Seq Appl [Internet]. 2015 [cited 2021 Sep 23];01(S1). Available from: https://www.omicsonline.org/open-access/structural-variation-detection-from-next-generation-sequencing-2469-9853-S1-007.php?aid=69055

121. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014 Jan 1;42(D1):D986–92.

122. Mak SST, Gopalakrishnan S, Carøe C, Geng C, Liu S, Sinding MHS, et al. Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. GigaScience [Internet]. 2017 Aug 1 [cited 2021 Sep 23];6(8). Available from: https://doi.org/10.1093/gigascience/gix049

123. Zhou Y, Liu C, Zhou R, Lu A, Huang B, Liu L, et al. SEQdata-BEACON: a comprehensive database of sequencing performance and statistical tools for performance evaluation and yield simulation in BGISEQ-500. BioData Min. 2019 Nov 15;12(1):21.

124. Loewe L, Hill WG. The population genetics of mutations: good, bad and indifferent. Philos Trans R Soc B Biol Sci. 2010 Apr 27;365(1544):1153–67.

125. Volfovsky N, Oleksyk TK, Cruz KC, Truelove AL, Stephens RM, Smith MW. Genome and gene alterations by insertions and deletions in the evolution of human and chimpanzee chromosome 22. BMC Genomics. 2009 Jan 26;10(1):51.

126. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D1005–12.

127. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016 Jan 4;44(D1):D862–8.

128. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet. 2013 Oct 1;132(10):1077–130.

129. Lobo I. Same genetic mutation, different genetic disease phenotype [Internet]. Nature Education. 2008 [cited 2021 Sep 23]. Available from: http://www.nature.com/scitable/topicpage/same-genetic-mutation-different-genetic-disease-phenotype-938

130. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017 Jul 6;101(1):5–22.

131. Marigorta UM, Rodríguez JA, Gibson G, Navarro A. Replicability and Prediction: Lessons and Challenges from GWAS. Trends Genet. 2018 Jul 1;34(7):504–17.

132. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet. 2017 Oct;18(10):599–612.

133. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017 Jun 15;169(7):1177–86.

134. Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. Philos Trans R Soc B Biol Sci. 2010 Jan 12;365(1537):185–205.

135. McVean GA, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR, Chakravarti A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov;491(7422):56–65.

136. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020 Mar 20;367(6484):eaay5012.

137. Nugent A, Conatser KR, Turner LL, Nugent JT, Sarino EMB, Ricks-Santi LJ. Reporting of race in genome and exome sequencing studies of cancer: a scoping review of the literature. Genet Med. 2019 Dec;21(12):2676–80.

138. Spratt DE, Chan T, Waldron L, Speers C, Feng FY, Ogunwobi OO, et al. Racial/Ethnic Disparities in Genomic Sequencing. JAMA Oncol. 2016 Aug 1;2(8):1070–4.

139. Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016 Oct;538(7624):238–42.

140. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016 Oct;538(7624):201–6.

141. Chen CY, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL. Improved ancestry inference using weights from external reference panels. Bioinformatics. 2013 Jun 1;29(11):1399–406.

142. Smith MW, O'Brien SJ. Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat Rev Genet. 2005 Aug;6(8):623–32.

143. Stephens JC, Briscoe D, O'Brien SJ. Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am J Hum Genet. 1994 Oct;55(4):809–24.

144. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep 1;19(9):1655–64.

145. Infinium Global Screening Array-24 v1.0 BeadChip. :7.

146. 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 Bioinforma. 2013;43(1):11.10.1-11.10.33.

147. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep 1;38(16):e164.

148. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012 Apr 1;6(2):80–92.

149. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. 2016;37(3):235–41.

150. Zheng-Bradley X, Streeter I, Fairley S, Richardson D, Clarke L, Flicek P, et al. Alignment of 1000 Genomes Project reads to reference assembly GRCh38. GigaScience [Internet]. 2017 Jul 1 [cited 2021 Sep 23];6(7). Available from: https://doi.org/10.1093/gigascience/gix038

151. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet. 2012 Mar 15;3:35.

152. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904–9.

153. McKinney W. Data Structures for Statistical Computing in Python. In Austin, Texas; 2010 [cited 2021 Sep 23]. p. 56–61. Available from: https://conference.scipy.org/proceedings/scipy2010/mckinney.html

154. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020 Mar;17(3):261–72.

155. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep 1;81(3):559–75.

156. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science. 2001 Feb 16;291(5507):1304–51.

157. Kempfer R, Pombo A. Methods for mapping 3D chromosome architecture. Nat Rev Genet. 2020 Apr;21(4):207–26.

158. Alkes Group [Internet]. [cited 2021 Sep 23]. Available from: https://alkesgroup.broadinstitute.org/EIGENSOFT/

159. Depienne C, Mandel JL. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? Am J Hum Genet. 2021 May 6;108(5):764–85.

160. Gymrek M. A genomic view of short tandem repeats. Curr Opin Genet Dev. 2017 Jun 1;44:9–16.

161. Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, et al. The impact of short tandem repeat variation on gene expression. Nat Genet. 2019 Nov;51(11):1652–9.

162. Course MM, Gudsnuk K, Smukowski SN, Winston K, Desai N, Ross JP, et al. Evolution of a Human-Specific Tandem Repeat Associated with ALS. Am J Hum Genet. 2020 Sep 3;107(3):445–60.

163. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019 Dec 2;47(21):10994–1006.

164. Renton AE, Majounie E, Waite A, Simón-Sánchez J, Rollinson S, Gibbs JR, et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron. 2011 Oct 20;72(2):257–68.

165. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron. 2011 Oct 20;72(2):245–56.

166. Seixas AI, Loureiro JR, Costa C, Ordóñez-Ugalde A, Marcelino H, Oliveira CL, et al. A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of DAB1, Mapping to SCA37, Causes Spinocerebellar Ataxia. Am J Hum Genet. 2017 Jul 6;101(1):87–103.

167. Mantere T, Kersten S, Hoischen A. Long-Read Sequencing Emerging in Medical Genetics. Front Genet [Internet]. 2019 [cited 2022 May 8];10. Available from: https://www.frontiersin.org/article/10.3389/fgene.2019.00426

168. Trost B, Engchuan W, Nguyen CM, Thiruvahindrapuram B, Dolzhenko E, Backstrom I, et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. Nature. 2020 Oct;586(7827):80–6.

169. Dashnow H, Pedersen BS, Hiatt L, Brown J, Beecroft SJ, Ravenscroft G, et al. STRling: a k-mer counting approach that detects short tandem repeat expansions at known and novel loci [Internet]. bioRxiv; 2021 [cited 2022 May 7]. p. 2021.11.18.469113. Available from: https://www.biorxiv.org/content/10.1101/2021.11.18.469113v1

170. Cao MD, Balasubramanian S, Bodén M. Sequencing technologies and tools for short tandem repeat variation detection. Brief Bioinform. 2015 Mar 1;16(2):193–204.

171. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 2012 Jun;22(6):1154–62.

172. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017 Jun;14(6):590–2.

173. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008 May 1;453(7191):56–64.

174. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science. 2021 Apr 2;372(6537):eabf7117.

175. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999 Jan 15;27(2):573–80.

176. Schaper E, Kajava AV, Hauser A, Anisimova M. Repeat or not repeat?—Statistical validation of tandem repeat prediction in genomic sequences. Nucleic Acids Res. 2012 Nov 1;40(20):10005–17.

177. Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance | Briefings in Bioinformatics | Oxford Academic [Internet]. [cited 2022 May 7]. Available from: https://academic.oup.com/bib/article/14/1/67/306476?login=true

178. Hamada M, Ono Y, Asai K, Frith MC. Training alignment parameters for arbitrary sequencers with LAST-TRAIN. Bioinformatics. 2017 Mar 15;33(6):926–8.

179. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio [Internet]. 2013 May 26 [cited 2022 May 8]; Available from: http://arxiv.org/abs/1303.3997

180. BEDTools: a flexible suite of utilities for comparing genomic features | Bioinformatics | Oxford Academic [Internet]. [cited 2022 May 7]. Available from: https://academic.oup.com/bioinformatics/article/26/6/841/244688

181. Cortese A, Tozza S, Yau WY, Rossi S, Beecroft SJ, Jaunmuktane Z, et al. Cerebellar ataxia, neuropathy, vestibular areflexia syndrome due to RFC1 repeat expansion. Brain. 2020 Feb 1;143(2):480–90.

182. Orr HT, Chung M yi, Banfi S, Kwiatkowski TJ, Servadio A, Beaudet AL, et al. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat Genet. 1993 Jul;4(3):221–6.

183. Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-Cendes I, et al. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat Genet. 1996 Nov;14(3):269–76.

184. Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet. 1994 Nov;8(3):221–8.

185. Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C, et al. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the α1A-voltage-dependent calcium channel. Nat Genet. 1997 Jan;15(1):62–9.

186. Testa CM, Jankovic J. Huntington disease: A quarter century of progress since the gene discovery. J Neurol Sci. 2019 Jan 15;396:52–68.

187. Brook JD, McCurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, et al. Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell. 1992 Feb 21;68(4):799–808.

188. Liquori CL, Ricker K, Moseley ML, Jacobsen JF, Kress W, Naylor SL, et al. Myotonic Dystrophy Type 2 Caused by a CCTG Expansion in Intron 1 of ZNF9. Science. 2001 Aug 3;293(5531):864–7.

189. Campuzano V, Montermini L, Moltò MD, Pianese L, Cossée M, Cavalcanti F, et al. Friedreich's Ataxia: Autosomal Recessive Disease Caused by an Intronic GAA Triplet Repeat Expansion. Science. 1996 Mar 8;271(5254):1423–7.

190. Gossye H, Engelborghs S, Van Broeckhoven C, van der Zee J. C9orf72 Frontotemporal Dementia and/or Amyotrophic Lateral Sclerosis. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Gripp KW, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2022 May 8]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK268647/

191. Talbott EO, Malek AM, Lacomis D. The epidemiology of amyotrophic lateral sclerosis. Handb Clin Neurol. 2016;138:225–38.

192. Talbott EO, Malek AM, Lacomis D. Chapter 13 - The epidemiology of amyotrophic lateral sclerosis. In: Aminoff MJ, Boller F, Swaab DF, editors. Handbook of Clinical Neurology [Internet]. Elsevier; 2016 [cited 2022 Apr 17]. p. 225–38. (Neuroepidemiology; vol. 138). Available from: https://www.sciencedirect.com/science/article/pii/B9780128029732000136

193. Arthur KC, Calvo A, Price TR, Geiger JT, Chiò A, Traynor BJ. Projected increase in amyotrophic lateral sclerosis from 2015 to 2040. Nat Commun. 2016 Aug 11;7:12408.

194. Kiernan MC, Vucic S, Cheah BC, Turner MR, Eisen A, Hardiman O, et al. Amyotrophic lateral sclerosis. The Lancet. 2011 Mar 12;377(9769):942–55.

195. Beleza-Meireles A, Al-Chalabi A. Genetic studies of amyotrophic lateral sclerosis: Controversies and perspectives. Amyotroph Lateral Scler. 2009 Jan 1;10(1):1–14.

196. Maruyama H, Morino H, Ito H, Izumi Y, Kato H, Watanabe Y, et al. Mutations of optineurin in amyotrophic lateral sclerosis. Nature. 2010 May;465(7295):223–6.

197. Sreedharan J, Blair IP, Tripathi VB, Hu X, Vance C, Rogelj B, et al. TDP-43 Mutations in Familial and Sporadic Amyotrophic Lateral Sclerosis. Science. 2008 Mar 21;319(5870):1668–72.

198. Kwiatkowski TJ, Bosco DA, LeClerc AL, Tamrazian E, Vanderburg CR, Russ C, et al. Mutations in the FUS/TLS Gene on Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis. Science. 2009 Feb 27;323(5918):1205–8.

199. Vance C, Rogelj B, Hortobágyi T, De Vos KJ, Nishimura AL, Sreedharan J, et al. Mutations in FUS, an RNA Processing Protein, Cause Familial Amyotrophic Lateral Sclerosis Type 6. Science. 2009 Feb 27;323(5918):1208–11.

200. van Es MA, Veldink JH, Saris CGJ, Blauw HM, van Vught PWJ, Birve A, et al. Genome-wide association study identifies 19p13.3 (UNC13A) and 9p21.2 as susceptibility loci for sporadic amyotrophic lateral sclerosis. Nat Genet. 2009 Oct;41(10):1083–7.

201. Benyamin B, He J, Zhao Q, Gratten J, Garton F, Leo PJ, et al. Cross-ethnic meta-analysis identifies association of the GPX3-TNIP1 locus with amyotrophic lateral sclerosis. Nat Commun. 2017 Sep 20;8(1):611.

202. van Rheenen W, van der Spek RAA, Bakker MK, van Vugt JJFA, Hop PJ, Zwamborn RAJ, et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. Nat Genet. 2021 Dec;53(12):1636–48.

203. La Spada AR, Paulson HL, Fischbeck KH. Trinucleotide repeat expansion in neurological disease. Ann Neurol. 1994;36(6):814–22.

204. Majounie E, Renton AE, Mok K, Dopper EG, Waite A, Rollinson S, et al. Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. Lancet Neurol. 2012 Apr 1;11(4):323–30.

205. Fritsche LG, Gruber SB, Wu Z, Schmidt EM, Zawistowski M, Moser SE, et al. Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. Am J Hum Genet. 2018 Jun 7;102(6):1048–61.

206. Saini S, Mitra I, Mousavi N, Fotsing SF, Gymrek M. A reference haplotype panel for genome-wide imputation of short tandem repeats. Nat Commun. 2018 Oct 23;9(1):4397.

207. Border R, Smolen A, Corley RP, Stallings MC, Brown SA, Conger RD, et al. Imputation of behavioral candidate gene repeat variants in 486,551 publicly-available UK Biobank individuals. Eur J Hum Genet. 2019 Jun;27(6):963–9.

208. Marina H, Suarez-Vega A, Pelayo R, Gutiérrez-Gil B, Reverter A, Esteban-Blanco C, et al. Accuracy of Imputation of Microsatellite Markers from a 50K SNP Chip in Spanish Assaf Sheep. Animals. 2021 Jan;11(1):86.

209. Koboldt DC, Ding L, Mardis ER, Wilson RK. Challenges of sequencing human genomes. Brief Bioinform. 2010 Sep;11(5):484–98.

210. Wang M, Beck CR, English AC, Meng Q, Buhay C, Han Y, et al. PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. BMC Genomics. 2015 Mar 19;16(1):214.

211. Pineda SS, Lee H, Fitzwalter BE, Mohammadi S, Pregent LJ, Gardashli ME, et al. Single-cell profiling of the human primary motor cortex in ALS and FTLD [Internet]. Neuroscience; 2021 Jul [cited 2022 May 9]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.07.07.451374

212. Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun. 2019 Nov 28;10(1):5436.

213. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016 Nov;48(11):1443–8.

214. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016 Oct;48(10):1284–7.

215. Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum Genet. 2007 Nov 1;81(5):1084–97.

216. 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 Jan 1;32(Database issue):D493-496.

217. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.

218. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. 2012 Sep 1;22(9):1760–74.

219. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. Science. 2015 May 8;348(6235):660–5.

220. Zhu XR, Netzer R, Böhlke K, Liu Q, Pongs O. Structural and functional characterization of Kv6.2 a new gamma-subunit of voltage-gated potassium channel. Receptors Channels. 1999;6(5):337–50.

221. Bean BP. The action potential in mammalian central neurons. Nat Rev Neurosci. 2007 Jun;8(6):451–65.

222. LoRusso E, Hickman JJ, Guo X. Ion channel dysfunction and altered motoneuron excitability in ALS. Neurol Disord Epilepsy J. 2019;3(2):124.

223. Young PE, Kum Jew S, Buckland ME, Pamphlett R, Suter CM. Epigenetic differences between monozygotic twins discordant for amyotrophic lateral sclerosis (ALS) provide clues to disease pathogenesis. PLoS ONE. 2017 Aug 10;12(8):e0182638.

224. Hornstra IK, Nelson DL, Warren ST, Yang TP. High resolution methylation analysis of the FMR1 gene trinucleotide repeat region in fragile X syndrome. Hum Mol Genet. 1993 Oct;2(10):1659–65.

225. Russ J, Liu EY, Wu K, Neal D, Suh E, Irwin DJ, et al. Hypermethylation of repeat expanded C9orf72 is a clinical and molecular disease modifier. Acta Neuropathol (Berl). 2015 Jan 1;129(1):39–52.

226. Giesselmann P, Brändl B, Raimondeau E, Bowen R, Rohrandt C, Tandon R, et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. Nat Biotechnol. 2019 Dec;37(12):1478–81.

227. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011 May;12(5):363–76.

228. Yu X, Guda K, Willis J, Veigl M, Wang Z, Markowitz S, et al. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? BioData Min. 2012 Jun 18;5(1):6.

229. Mejzini R, Flynn LL, Pitout IL, Fletcher S, Wilton SD, Akkari PA. ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? Front Neurosci [Internet]. 2019 [cited 2022 Apr 15];13. Available from: https://www.frontiersin.org/article/10.3389/fnins.2019.01310

230. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018 May;19(5):286–98.

231. Zhou ZD, Jankovic J, Ashizawa T, Tan EK. Neurodegenerative diseases associated with non-coding CGG tandem repeat expansions. Nat Rev Neurol. 2022 Mar;18(3):145–57.

232. Paulson H. Chapter 9 - Repeat expansion diseases. In: Geschwind DH, Paulson HL, Klein C, editors. Handbook of Clinical Neurology [Internet]. Elsevier; 2018 [cited 2022 May 8]. p. 105–23. (Neurogenetics, Part I; vol. 147). Available from: https://www.sciencedirect.com/science/article/pii/B9780444632333000099

233. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nat Genet. 2019 Aug;51(8):1215–21.

234. Stevanovski I, Chintalaphani SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, et al. Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. Sci Adv. 8(9):eabm5386.

235. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022 Apr;376(6588):44–53.

236. Eichler EE, Clark RA, She X. An assessment of the sequence gaps: Unfinished business in a finished human genome. Nat Rev Genet. 2004 May;5(5):345–54.

237. Ellegren H. Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet. 2000 Apr;24(4):400–2.

238. Huang QY, Xu FH, Shen H, Deng HY, Liu YJ, Liu YZ, et al. Mutation Patterns at Dinucleotide Microsatellite Loci in Humans. Am J Hum Genet. 2002 Mar;70(3):625–34.

239. Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, et al. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet. 2000 May;66(5):1580–8.

240. Single-Cell Sequencing Tackles Basic and Biomedical Questions [Internet]. [cited 2022 May 7]. Available from: https://www.science.org/doi/10.1126/science.336.6084.976

241. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May;6(5):377–82.

242. Zhang L, Cui X, Schmitt K, Hubert R, Navidi W, Arnheim N. Whole genome amplification from a single cell: implications for genetic analysis. Proc Natl Acad Sci U S A. 1992 Jul 1;89(13):5847–51.

243. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020 Feb 7;21(1):31.

244. Biezuner T, Raz O, Amir S, Milo L, Adar R, Fried Y, et al. Comparison of seven single cell whole genome amplification commercial kits using targeted sequencing. Sci Rep. 2021 Aug 25;11(1):17171.

245. Hou Y, Wu K, Shi X, Li F, Song L, Wu H, et al. Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. GigaScience [Internet]. 2015 [cited 2022 May 8];4(1). Available from: http://www.scopus.com/inward/record.url?scp=84979520403&partnerID=8YFLogxK

246. Huang L, Ma F, Chapman A, Lu S, Xie XS. Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. Annu Rev Genomics Hum Genet. 2015;16:79–102.

247. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. Nature. 2020 May;581(7809):444–51.

248. Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biol. 2019 Aug 9;20(1):159.

249. Chen R, Butte AJ. The reference human genome demonstrates high risk of type 1 diabetes and other disorders. Pac Symp Biocomput Pac Symp Biocomput. 2011;231–42.

250. A star is born: the updated Human Reference Genome : Methagora [Internet]. [cited 2022 May 7]. Available from: http://blogs.nature.com/methagora/2013/12/the_updated_human_reference_genome.html

251. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. Trends Genet. 2009 Nov 1;25(11):489–94.

252. Carlson CS, Matise TC, North KE, Haiman CA, Fesinmeyer MD, Buyske S, et al. Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. PLOS Biol. 2013 Sep 17;11(9):e1001661.

253. Graham SE, Clarke SL, Wu KHH, Kanoni S, Zajac GJM, Ramdas S, et al. The power of genetic diversity in genome-wide association studies of lipids. Nature. 2021 Dec;600(7890):675–9.