**Title**
Unravelling the structural complexity of the human genome using linked-read sequencing and optical mapping technologies



**Author**
Wong, Hang Yin Karen

**Publication Date**
2020

Peer reviewed|Thesis/dissertation

# Unravelling the structural complexity of the human genome using linked-read sequencing and optical mapping technologies

by
Hang Yin Karen Wong

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Neil Risch
E3D0E6171DC74F3...

Neil Risch
Chair

DocuSigned by:
Pui Kwok

Pui Kwok

DocuSigned by:
Jeffrey D Wall
9CC4C1B0D513432...

Jeffrey D Wall

Committee Members

## ACKNOWLEDGEMENTS

This dissertation would not have been possible without the many individuals who have inspired me along the way. First, I would like to thank my PhD advisor Dr. Pui-Yan Kwok for guiding me and providing support throughout my scientific career and beyond. Amid his busy schedule, he is always available and ready to discuss my work with great enthusiasm. He has built an incredible network of talented people, always giving me the opportunities to connect and learn from those with different backgrounds. The projects in the lab are constantly at the forefront of scientific development and I cannot thank him enough for giving me the resources and freedom to grow as an independent scientist. I owe a debt of gratitude most of all to my dissertation and qualifying exam committee members: Drs. Neil Risch, Jeff Wall, Nadav Ahituv, Joseph Shieh, and Elad Ziv. They generously shared their time to give me guidance on my projects and welcomed me to discuss new ideas.

I am extremely grateful to the many members of the Kwok lab, past and present, from whom I have learned valuable knowledge. I must thank Michal Levy-Sakin and Yulia Mostovoy for their generosity and patience in mentoring me. The long discussions, from conceptualizing ideas to debugging, had sparked my fundamental interests and curiosity in bioinformatics and human genomics. I have also benefited immensely from Annie Poon and Catherine Chu, for teaching me hands-on skills in various types of sequencing and optical mapping technologies. I also thank Stephen Chow and Walfred Ma for all of their extensive time and effort spent helping me with my dissertation work. Additionally, this work would not have been possible without the CVRI IT team, many of

whom are willing to go above and beyond to answer my questions. My PhD work would have taken many times longer without their expert IT advice.

In these past several years, I also have had the wonderful privilege to work with the finest scientists including Drs. Clive Pullinger, John Kane, Jennifer Puck, and Monica Penon on various projects, and every single one has enriched my scientific acumen. I will remember all the outrageous stories John Kane had shared with me throughout the years.

I also want to thank all of my collaborators in Taiwan for brainstorming, helping with data generation, and taking me out every time I visited. I would like to thank them all for making me feel like home even when I was 7,000 miles away, especially Elin Wang, Erh-Chan Yeh, Nancy Wei, and Hsiao-Jung Kao.

None of these would have been possible without the help and support from my friends and BMS peers, who would always laugh at my cynical jokes, push me through the good times and bad. To Kat Yu and Didi Zhu, I will always remember the fun memories we created in Taiwan and Japan. To Keith Cheung, while I did not adhere to your advice, you literally changed my career trajectory when I had to make the most important decision in 2016. Thank you for offering to help when I stumbled in the Torres lab.

Last but not least, support coming from my family has always been the main driver of my success. I want to thank my parents for making tremendous sacrifices to give me new opportunities in the states. They have always been supportive but never pushed me to do anything without my own will. Finally, I wish to thank my partner, Danny Depe, for his unwavering support and patience over the years. His constant ray

of sunshine can blast through the darkest clouds even when I am overwhelmed by frustrations. His perpetual optimism and genuine happiness are simply contagious, and these small doses of positive energy have kept me sane during the most stressful time of the journey. I dedicate this dissertation to him, as he gave me the courage to overcome challenges and achieve new heights.

## CONTRIBUTIONS

This body of work contains the following published or submitted materials:

1. **Wong, K.H.Y.**, Levy-Sakin, M., Ma, W., Gonzaludo, N., Mak, A.C.Y., Vaka, D., Poon, A., Chu, C., Lao, R., Balamir, M., Grenville, Z., Wong, N., Kane, J.P., Kwok, P.-Y. Three patients with homozygous familial hypercholesterolemia: Genomic sequencing and kindred analysis. *Mol. Genet. Genomic Med*. (2019). DOI: 10.1002/mgg3.1007.
2. **Wong, K.H.Y.**, Ma, W., Wei, C.-Y., Yeh, E.-C., Lin, W.-J., Wang, E.H.F., Su, J.-P., Hsieh, F.-J., Kao, H.-J., Chen, H.-H., Chow, S.K., Young, E., Chu, C., Poon, A., Yang, C.-F., Lin, D.-S., Hu, Y.-F., Wu, J.-Y., Lee, N.-C., Hwu, W.-L., Boffelli, D., Martin, D., Xiao, M., Kwok, P.-Y. Towards a reference genome that captures global genetic diversity. *In revision*.
3. Shieh, J.T.*, Penon-Portmann, M.*, **Wong, K.H.Y.***, Levy-Sakin, M., Verghese, V., Slavotinek, A., Gallagher., R., Mendelsohn, B., Tenney J., Beleford, D., Perry, H., Chow, S.K., Sharo, A.G., Qi, Z., Yu, J., Klein, O., Martin, D., Kwok, P.-Y., Boffelli, D. Full genome assemblies provide a single comprehensive test for rare genetic disorders. *In revision*.
4. **Wong, K.H.Y.**, Levy-Sakin, M., Kwok, P.-Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat. Commun*. (2018). DOI:10.1038/s41467-018-05513-w.

   * All authors contributed equally

*“I see my path, but I don't know where it leads. Not knowing where I'm going is what inspires me to travel it.”*

-Rosalia de Castro

## ABSTRACT

### Unravelling the structural complexity of the human genome using linked-read sequencing and optical mapping technologies

**Hang Yin Karen Wong**

Elucidating the full spectrum of genetic variations across the human population is a fundamental pursuit in scientific research as it underlies the complex interplay between genotype and phenotype. However, most resequencing studies performed to date rely on short read technologies and the lack of long-range sequence information precludes comprehensive structural variations (SVs) analysis. While numerous SV algorithms can pinpoint deletion breakpoints with high sensitivity, it is much more challenging to detect other types of SVs as they cannot be directly inferred through an alignment-based approach. In this dissertation, we leveraged state-of-the-art technologies that allow for long-range genome sequencing and mapping to comprehensively evaluate SVs in the context of population genetics and genomic medicine.

To improve the current representation of the human reference genome, we utilized 10x Genomics (10xG) whole genome linked-read sequencing to generate *de novo* assemblies of 328 genomes from around the world. We breakpoint-resolved 18Mb of genomic sequences missing from the reference genome—aka Non-reference Unique Insertions (NUIs)—and linearly integrated them into the GRCh38 primary chromosomal assemblies so that these NUIs can be annotated based on the local genomic context.

We demonstrated that many of these NUIs can be found in the human transcriptome and hence are likely to have functional significance. Our proof-of-concept reference representation will allow researchers to identify biologically relevant polymorphisms beyond what is currently detected, thus enhancing the interpretability of all existing and future short-read whole genome sequencing datasets.

Furthermore, we applied either linked-read sequencing or in conjunction with Bionano optical mapping in two precision medicine projects to identify disease-causal variants that previously evaded detection. In the first project, whole genome linked-read sequencing was performed on two families with homozygous familial hypercholesterolemia to determine the underlying genetic cause of the disease. In the second project, we applied both technologies on 50 undiagnosed children with suspected genetic diseases. Our automated informatics pipeline identified 18 clinical diagnoses, with which 22% of these cases were attributed to cryptic SVs. Our results substantiate the use of long-range sequencing and mapping in patients with rare genetic diseases. Their applications in genomic medicine can bring tremendous precision to clinical diagnosis, thereby fulfilling the promise of individualized medicine.

## TABLE OF CONTENTS

## LIST OF TABLES

### Chapter 3:

## LIST OF FIGURES

**Chapter 2:**



**Chapter 3:**



**Chapter 4:**

# CHAPTER 1: INTRODUCTION TO GENOMICS

When Gregor Mendel established the fundamental rules of heredity in the 1800s through conducting a series of experiments on pea plants, the field of modern genetics was born. While genetics became a well-established scientific discipline involving the targeted study of a specific gene, it was discovered later that many phenotypic traits are contributed based on the complex interplay between different sets of genes. In contrast to genetics, genomics is a newly developed discipline that analyzes the entire collection of genes residing in the DNA of an organism. Genomics is more feasible in the recent decades with the advent of high-performance computing and advanced sequencing technology. Genomic research now permeates the scientific literature and many of the discoveries have directly shaped the clinical practices in the pursuit of individualized medicine. In this chapter, we will discuss some of the seminal work contributing to the success of genomics and the application of high throughput sequencing in the healthcare industry as genomic medicine takes hold.

## 1.1 The Human Genome Project

The blueprint of life was revealed in 2002 as scientists across the globe collaboratively put together, for the first time, a draft sequence of the human genome[1]. The flagship product of this endeavor is the human reference genome composed of roughly 3 billion nucleotides. While the end product is far from perfect, it enables new scientific discoveries at an unparalleled speed and scale. However, this remarkably profound undertaking came with a hefty price tag of roughly 2.7 billion USD.

The draft sequence of the human genome was constructed through a brute-force process known as the “hierarchical shotgun sequencing”. In this approach, amplified genomic DNA from anonymous human donors is first sheared into fragments of approximately 200kb. These large fragments are cloned into bacterial artificial chromosomes (BAC) and propagated in bacteria, to build a clone library as a stable resource. Next, a molecular map of each BAC clone was constructed by screening it for sequence-tagged sites (STSs) and digesting it with restriction enzyme to identify signature fragment sizes. The result is a molecular “fingerprint” unique to each clone. Scientists can compare these fingerprints to order the clones based on overlapping ends. Using fluorescence *in situ* hybridization, an array of clones was defined per chromosome, and a minimal tiling path was picked to select a number of BAC clones for sequencing. Chosen BAC clones were randomly sheared and the resulting DNA fragments, each around 1 to 2kb in length, were sub-cloned into vectors to generate a shotgun library for Sanger sequencing, which was the most accurate sequencing method that yielded the longest read length at that time. Once the DNA sequence was determined, researchers used the overlapping bases to assemble individual pieces of DNA into larger chunks, or contigs (contiguous sequences). The finished product contains 341 gaps and span 99% of the human genome with a predicted error rate of 0.001%[2]. Many of the remaining gaps were enriched in regions with low sequence complexity. While tremendously costly and laborious, this undertaking laid out one of the most important scientific frameworks for all future human genetic studies to come.

The human reference genome is inarguably the most complete representation of the human genome sequenced to date. However, the reference genome is a set of

composite haplotypes generated based on several anonymous DNA donors, of which a male donor of African American ancestry, namely RP11, made up about 70% of the reference sequence[3]. The lack of genetic diversity, as well as the missing genomic content in the reference genome, became alarming in recent years as biases towards Caucasians in genetic research is increasingly pronounced[4]. The latest genome reference build GRCh38 represents some of these missing sequences in the form of alternative contigs, and more are being added with every minor release. However, most downstream tools do not handle any of the resulting alignments in alternative, unplaced, or unlocalized contigs and are thereby removed.

In light of the shortcomings of the human reference genome, different research groups have proposed alternative methods to represent the human reference genome. One that has come under the spotlight recently is the concept to build genome graphs for improved read mapping and variant calling[5]. However, there are a plethora of challenges genome graphs have to overcome before they can fully be realized. Some of these challenges include 1) new toolkits have to be developed to handle graph-based alignment for all downstream next-generation sequencing (NGS) applications; 2) a new universal coordinate system has to be reinvented to encompass all the possible paths in a graph-based reference; 3) it has to be scalable and efficient to handle a large amount of sequencing data, and 4) the scientific community has to adapt to this new reference structure. On the other hand, there have been other efforts to build population-specific references. However, these representations only benefit genetically homogenous populations and the different coordinate systems are barriers to translate and generalize genetic findings. In light of these challenges, we proposed to augment the current

human reference genome by incorporating all the missing sequences in the form of the longest human linear haplotype, as discussed in Chapter 2.

## 1.2 High throughput sequencing

### 1.2.1 Next-generation sequencing

Within a span of about a decade, the cost of whole genome sequencing plummeted from 2.7 billion down to $1,000 in 2014. At the turn of the century, new DNA sequencing methods were being developed with the goal of reaching throughput high enough for whole genome application. These new sequencing technologies were collectively known as the “next-generation” or “second-generation” sequencing. Four major players emerged around that time—pyrosequencing (Roche 454 sequencing), ion semiconductor sequencing (Ion Torrent), sequencing by ligation (SOLiD), and sequencing by synthesis (Illumina sequencing).

All next-generation sequencing (NGS) technologies leverage shotgun sequencing technology, first shearing the genome randomly into small fragments of several hundred bases. A sequencing library is then prepared to generate enough DNA material for sequencing. During sequencing, tens of millions of sequencing reactions are carried out in parallel, thus achieving the throughput necessary to sequence an entire genome. Raw reads are then aligned to the reference genome to infer genetic differences. At the sequencing stage, the Ion torrent system is unique in that it determines the base identity by measuring change in pH due to the release of a hydrogen ion, while the other three systems require the use of fluorescently labelled nucleotide that can then be detected by an optical instrument upon integration. Since

their introduction in the 2000s, Illumina sequencing has been dominating the short-read market with ever declining sequencing cost, which in turn opens up a broad spectrum of novel sequencing applications in the scientific and medical communities.

### 1.2.2 Third-generation sequencing

In the sub $1,000 genome era, the demand for high throughput sequencing continues to rise steadily but potential competition is afoot for the short-read sequencing market. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are independently developing long read sequencing methods that can overcome the major shortcoming of NGS technologies—the limited read length. These breakthroughs allow researchers to interrogate regions of the genomes that are previously inaccessible by short reads such as low complexity and GC-rich sequences. Furthermore, long sequence span is particularly useful for *de novo* assembly—a process that reconstruct the genome sequence without using a reference structure—and haplotype phasing.

PacBio is powered by a technology called single molecule real time sequencing (SMRT). SMRT sequencing works by immobilizing polymerases at the bottom of zero-mode waveguides (ZMWs) in a SMRT cell and millions of DNA fragments are read one base at a time. Upon integrating a phosphor-linked nucleotide, light is emitted and measured in real time. This technology can produce sequencing reads near 12kb (need citation). Early-stage PacBio sequencing using the continuous long read (CLR) mode is error-prone but is amenable to higher sequencing coverage as most errors are randomly distributed. Soon after, the company developed the circular consensus sequencing (CCS) method that sequences the same DNA fragment for multiple rounds.

Albeit slower and more costly, sequencing with this mode can produce high fidelity (HiFi) reads with an average read concordance of 99.8%[6]. The improved accuracy negates the need for short-read polishing during genome assembly. However, even with the latest PacBio Sequel II system, the throughput capacity is still comparably lower compared to the Illumina NovaSeq.

While PacBio is the only company that provides highly accurate long read sequencing, the long-read technology from ONT has also been steadily improving. One key feature of nanopore sequencing is that molecules of any length can be sequenced theoretically with no compromise in accuracy. The longest sequenced read using the nanopore system exceeds 2Mb in length[7]. Nevertheless, ONT long reads still suffer from high error rate, ranging between 5%-15%[8,9]. Continuous development in base calling algorithms and pore chemistry warrant future attention.

### 1.2.3 Linked-read sequencing

An alternative method, 10x Genomics (10xG) “Linked-Read” sequencing, is a powerful advance over existing short-read technologies in that this approach is scalable and it preserves long-range sequence information. A microfluidics platform is used to partition and barcode high molecular weight genomic DNA (~100kb) so that reads deriving from the same fragment will share an identical barcode. The barcoded library is then sequenced on an Illumina sequencer, which is both high throughput and inexpensive relative to other technologies. By leveraging the unique identifier, highly accurate *de novo* assemblies can be achieved even though the number of gaps is considerably higher compared to those generated from long reads.

### 1.3 Genetic variations in the human genome

Prior to the completion of the HGP, several other international collaborative efforts, such as the International HapMap Project[10] and the 1000 Genome Project[11], took off to elucidate patterns of human genetic variations. Initial work mainly focused on cataloging common single nucleotide polymorphisms (SNPs) in different human populations. Such effort led to the development of array-based genotyping technologies and allowed specific SNP alleles to be re-genotyped. As a result, it became feasible to perform genome-wide associated studies on large cohorts of cases and controls through re-genotyping hundreds of thousands of SNPs simultaneously. While SNP is the most common and well-characterized form of genetic variations in the human genome, scientists soon realized that more complex forms of variations, namely insertion, deletion, inversion, and translocation, are also prevalent in the human genome. These are collectively known as structural variations (SVs), denoting structural changes that are larger than 50bp. It has been reported that there are as many as 27,622 SVs per human genome[12]. SVs can clearly contribute to disease pathogenesis when they disrupt genic or functional elements in a genome. However, there are also studies attributing different Mendelian diseases to SVs residing in introns with no known functions[13,14].

Unlike SNPs, SVs are not readily detectable by short read sequencing, which is still the most cost efficient and widely adopted sequencing technology to date. SV detection using alignment-based analysis is based on assessing read depth, read pair orientation, and split reads[15,16]. SV detection algorithms generally work well only on deletions because a short read can capture both breakpoints. In order to resolve other type of SVs, a full reconstruction of the haplotype, or *de novo* assembly, is generally

required. Repetitive DNA poses yet another constraint. For example, algorithms inherently cannot infer the size of a tandem repeat if it exceeds the read length. Due to these limitations, SVs have been under-appreciated until long read or linked-read sequencing started to gain popularity. As sequencing reads get longer, more SVs are discoverable and more clinically relevant variants can be identified. In the following chapters, we will discuss SVs identification using linked-reads and its application on clinical genomics for disease diagnosis.

### 1.4 Application of high throughput sequencing in the era of precision medicine

With the advent of improved technologies, the application of high throughput DNA sequencing in genomic medicine has become increasingly widespread. As the price tag of these technologies continues to plunge, sequencing for disease diagnosis is gaining momentum in different clinical specialties, rare disease genetics, oncology, reproductive health, and infectious disease in particular. While part of the growth is fueled by the low sequencing cost, the data-driven nature in massively parallel sequencing enables a systematic exploration of the exome or genome of a patient.

In genomic medicine, rare disease diagnosis has traditionally been imprecise and difficult. Direct Sanger sequencing on one or several genes has been the gold standard in clinical genomic laboratories since the 1980s. However, Sanger sequencing can only achieve diagnostic goals when a patient presents strong phenotypic or radiographic features indicative of a classical Mendelian disorder, which is often rare. Without a clear manifestation, misdiagnosis is disconcertingly common because physicians tend to practice with cognitive biases based on previous clinical experiences. Even worse, the

search for a diagnosis can easily turn into an odyssey with patients receiving multiple misdiagnoses and mistreatments, often unnecessary or even harmful to the patients, along the way. In the early 2000s, more advanced genomic technologies emerged, and the field began to shift from a hypothesis-driven diagnostic paradigm to a data-driven one. High-density DNA microarrays such as array comparative genomic hybridization (array CGH) and single nucleotide polymorphism (SNP) genotyping underpin many aspects of genomic medicine. They are frequently used in the context of evaluating intellectual disability, multiple congenital anomalies, and many neuropsychiatric disorders even up to this day. These technologies can detect large copy number variation (CNV) of ~400kb[17]. Although microarray-based technology has increased diagnostic yields, it is not a stand-alone diagnostic tool as the resolution is quite low and it fails to detect subtle changes in the genome.

Over the ensuing decade, massively parallel DNA sequencing, or next-generation sequencing (NGS) technologies have drastically improved. In 2014, the cost of whole genome sequencing dipped below $1,000[18]. With improved accuracy and throughput, NGS technology has revolutionized genomic medicine, and for the first time, it allows researchers to interrogate virtually all variants in a genome. The incremental gain in WES/WGS popularity in genomic medicine stems from the fact that a human genome can be surveyed without knowing the disease *a priori*. This hypothesis-free approach is especially useful in rare disease diagnostics and by performing trio-based sequencing of unaffected parents and an affected offspring, physicians can identify *de novo* or mosaic mutations that are previously not amenable to linkage analysis. By harnessing the power of NGS, the diagnostic rate of patients with rare genetic disorders

has increased to 25-30%[19,20], of which 30-72% of the diagnoses can result in changes in clinical management or even outcomes[21]. Despite the many successful stories in clinical genomics, the use of NGS technology in genomic medicine is not without limitations, and a substantial fraction of the sequenced patients remains undiagnosed. While there are many plausible explanations for these unsolved cases, two major factors limit the diagnostic outcome of patients sequenced with NGS technologies. First, the use of short-read sequencing is inadequate to detect large genomic variants that may be responsible for the disease. Second, not knowing the genetic basis of the disease etiology makes rare Mendelian disorders undiagnosable. Both of these points will be elaborated further in Chapter 3.

# CHAPTER 2: TOWARDS A REFERENCE GENOME THAT CAPTURES GLOBAL GENETIC DIVERSITY

## 2.1 Abstract

Diversity and inclusion are fundamental goals in human genetics and biomedical research. However, the current human reference genome is predominantly derived from a single individual, and thus does not adequately reflect human genetic diversity. Here, we analyzed the largest collection of high-quality human assemblies of genetically divergent human populations to identify missing sequences in the human reference genome with breakpoint resolution and understand the full scope of Non-reference Unique Insertions (NUIs). Our collection consists of 338 high-quality genome assemblies, including 329 genomes generated with 10x Genomics Linked-Reads and 9 genomes sequenced on the PacBio platform. We identified 127,727 recurrent NUIs (≥10bp) spanning 18,048,877bp, many of which disrupt exons and known regulatory elements. To improve genome annotations and sequencing interpretations, we linearly integrated these NUIs into the chromosomal assemblies and constructed a Human Diversity Reference (HDR). Leveraging the HDR, an average of 402,573 previously unmapped reads can be recovered for a given genome sequenced to ~40X coverage, facilitating detection of an average 27,658 novel single nucleotide polymorphisms per genome. Transcriptomic diversity among these NUIs can also be directly assessed. We successfully mapped tens of thousands of previously discarded RNA-Seq reads to the HDR, and identified NUI transcription evidence in 4,781 gene loci, underlining the importance of NUIs in functional genomics. Our extensive datasets covering a variety of

human genomes and the intuitive structure of the HDR are important advances toward a comprehensive reference representation of global human genetic diversity.

## 2.2 Introduction

The completion of the human reference genome in 2003 marked a major milestone in biomedical research. The human reference genome represents a first attempt to produce a telomere to telomere assembly of the human genome and provides a universal coordinate system to which all human sequences from subsequent work are aligned and compared. It also forms the basis for all studies that annotate genes, define regulatory elements, compare genome diversity, and search for disease-causing mutations. As the cost of short-read sequencing drops, millions of people have already been sequenced and their DNA sequences vetted against the human reference genome to identify genetic factors associated with health and disease.

However, some factors limit the utility of the human reference genome for genome analysis. First, it is a composite haplotype derived from a small number of donors recruited at one location in the US, with 70% of its sequences coming from a single DNA donor[3], and so it does not capture the diversity of the world's population. Second, it does not contain numerous unique sequences found in multiple individuals but not in the human reference genome[12,22-27]. Some of these Non-reference Unique Insertions (NUIs) are in genes and affect gene annotation[22,23]. Importantly, unlike deletions, insertions are cryptic events that are not readily identified by structural variation (SV) callers unless a *de novo* assembly is generated. Third, some highly variable regions of the human genome consist of long, near-identical sequences that

are highly variable from person to person. Consequently, short-read alignment to the human reference genome often does not reflect the true configuration of these regions in the sequenced individual[28,29].

As many individuals have been sequenced around the world, it is common knowledge that a significant portion of whole genome sequencing (WGS) reads from each individual is discarded because they cannot be aligned to the human reference genome, even though the discarded reads are not contaminants. Thus, all WGS data analyses to-date are based on an “incomplete” human reference genome, potentially missing important signals residing in NUIs. The latest genome build GRCh38 includes some of these missing sequences as alternative contigs, and more are being added with every minor release. Efforts to utilize the alternative contigs in WGS analyses are underway, including the approach of building genome graphs for improved read mapping and variant calling[30].

To build a more useful human reference genome for WGS analysis, we have constructed a Human Diversity Reference (HDR) that incorporates NUIs derived from all high-quality genome assemblies known to-date into a linear genome. Here we assembled 305 diverse human genomes in-house and analyzed an additional 33 assemblies from the public domain. In total, the HDR includes 127,727 recurrent NUIs (≥10bp) that are placed uniquely in the GRCh38 chromosomal assemblies and span 18 Mb. The HDR is able to recover tens of thousands of previously unmapped WGS and RNA-Seq reads, thus allowing interrogation of sequences that would otherwise evade detection. We show that discarded RNA-Seq reads can align to 4,781 gene-containing loci, and that as many as 22,551 NUIs directly disrupt known regulatory elements. Not

only can the HDR improve genome annotations, it can directly benefit a wide range of research communities because the use of the HDR is intuitive. This work presents an important first step towards a human reference genome that represents the diversity of human populations.

### 2.3 Materials and Methods

*Sample collection and data generation*

Samples used in this study were collected from one of the eight sources: 1KGP samples sequenced in house using cell lines purchased from Coriell (52), Full Genome Analysis Project (FGAP) asymptomatic participants (99), Taiwan Precision Medicine Initiative (154), Illumina's Polaris Project (22), 10x Genomics (10xG) website (2), Audano et al (7), Seo et al (1), and Shi et al (1) (**Table 2.1**; see original manuscript). The 52 Coriell cell lines were selected to cover the breadth of genetic diversity across the four major continents – one male and one female were chosen from each of the 26 1KGP populations. FGAP participants were referred to UCSF and they gave written informed consent prior to their enrollment in the study, which was approved by the Human Research Institutional Review Board as part of the UCSF Human Research Protection Program. The 154 Taiwanese subjects of Han Chinese ancestry in this study were recruited from the Taiwan Biobank (128), National Taiwan University Hospital (22), Taipei General Hospital (2), and Mackay Memorial Hospital (2). This study was approved by the Institutional Review Board of the respective recruitment hospitals and Academia Sinica, and ethical approval was granted by the Internal Review Board of the

Taiwan Biobank. All other sequencing datasets included in this study were obtained from public sources.

Sample relatedness were assessed among the cohort sequenced with 10xG. Sample-merged VCF file generated from 10xG Longranger was converted to plink format and subject to kinship calculation using KING v2. Samples were removed if the pairwise estimated kinship coefficient exceeds 0.0442, which indicates at least third-degree relationships.

*10xG assemblies*

Purchased 1KGP cell lines, primary cells collected for FGAP and TPMI were processed according to the 10xG Chromium Genome User Guide and as published previously[31]. Samples were sequenced to an average of 60x coverage for this study using 2x150 paired-end reads (**Supplementary Figure S2.1**; see end of chapter). Barcode-aware alignment was performed using Longranger v2 with the "wgs" pipeline. *De novo* assemblies were generated using Supernova v2. For the Polaris Project samples, barcode-aware alignment BAM files were downloaded and converted to FASTQ format using bamtofastq utility designed for 10xG Linked-Reads. Processed FASTQ files were subsequently used as input for Supernova. Polaris samples with mean inferred molecule length below 40kb were excluded. Two additional *de novo* assemblies were directly obtained from the 10xG website. The pseudohap FASTA output was used for all downstream analysis.

*PacBio assemblies*

PacBio assemblies were downloaded from Audano et al.[22], Seo et al.[32], and Shi et al[33]. The motivation to include PacBio assemblies was to supplement 10xG assembles in regions of low complexity.

*BioNano optical maps*

We generated Bionano optical maps with one or a combination of enzymes if DNA materials were available (**Supplementary Table S1**; see original manuscript). Labelling and staining were performed according to manufacturer's protocol. SV calls were generated using Bionano Solve v3.4.

*Reference terminology*

Hereafter, we refer to the chromosomal assemblies 1-22, X, and Y as the core assemblies throughout the paper. We intentionally avoided the term "primary assemblies" because the reference genome consortium uses this term to define the assembled chromosomes as well as unlocalized and unplaced sequences. We focus our analysis on the core chromosomes because the vast majority of the downstream processing and annotation tools ignore patches and alternative contigs.

*Definitions of Non-reference unique insertions*

NUIs are defined as follows:

1. Insert size ≥10bp

2. Reference breakpoints not overlapping segmental duplications (downloaded as part of the pre-built reference package provided by 10xG)
3. Reference breakpoints not overlapping SV filter file (downloaded as part of the pre-built reference package provided by 10xG)
4. Identified in at least two samples
5. Pass filtering criteria as defined in the section “Sequence annotations for filtering” under pipeline architecture

*Pipeline architecture*

The pipeline consists of four major steps: per-sample insertion calling, clustering and choosing representative non-redundant insertions, applying additional filters, and constructing the HDR (**Supplementary Figure S2.1**; see end of chapter).

*Per-sample insertion calling*

Nucmer[34] was used to align pseudohap FASTA files generated by Supernova, or assemblies directly downloaded in the case for PB, to genome build GRCh38 reference provided by 10xG. Insertions were detected using a modified version of Assemblytics[35]. Insert size is defined as the difference between the query gap size and the reference gap size (**Supplementary Figure S2.2**; see end of chapter). Calls in decoy, alternate, unplaced, and unlocalized contigs were excluded. We additionally applied two filters—SV blacklist, and segmental duplications as defined by 10xG—to remove potentially spurious calls. Insertions ≥10bp was used as the size cutoff for this study and the

primary reason is that most SV callers are able to detect insertions <10bp with high specificity and sensitivity.

*Clustering and choosing representative non-redundant insertions*

All insertions whose reference breakpoints overlapped by at least 1bp were grouped into a component (**Supplementary Figure S2.3a**; see end of chapter). Reference breakpoints within a component do not necessarily have to be identical because even highly similar insertions can have different reference breakpoints, depending on how repetitive the adjacent sequences are. Singletons (components with sequences from only one sample) are discarded at this point. While singleton insertions contribute significantly to the overall insertion count, it remains unclear whether these sequences are private in their respective genomes or results of misassemblies. Importantly, we aim to produce a highly specific set of insertion calls and thereby we require insertions to be called in at least two individuals. This will remove most of the potentially spurious insertion calls because it is extremely unlikely for two independent human lineages to have the same insertion in a given genomic locus.

Within a component, all sequences were extracted, repeat-masked[36] (-species human -qq -xsmall -noint), and passed on to the multiple alignment algorithm kalign[37]. During sequence extraction, paddings were added to insertion sequences to improve the accuracy of multiple alignment. First, the minimum reference start site and the maximum reference end site within a component were calculated and 50bp were added to both ends (**Supplementary Figure S2.3b**; see end of chapter). We note that reference paddings are only added if at least 1 insertion in a component is longer than

50bp. This is primarily because shorter sequences tend to share identical NUI breakpoints and multiple alignment can stack these sequences more accurately and efficiently. During repeat masking, only low complexity DNA and simple repeats were masked for scoring purposes.

To pick the most representative sequence in each component, we first calculated a global score for every insertion in a component based on multiple alignment results. The global scoring scheme is as follows: Match for an unmasked base +2; match for a masked base +0.2; mismatch -1, undetermined base (N) -0.5; gap extension +0. Only the inserted sequence segments are used for scoring, hence longer sequences are advantageous. In order to maximize the number of unique bases, masked bases are given smaller weights. Each undetermined base (N) gets a penalty equivalent to half of a mismatch so that a fully resolved sequence will be more likely to have a higher score. All sequences within a component are sorted numerically in descending order based on the global score. The insertion with the highest score will be selected as the representative sequence.

Within the component, the pipeline then performs a recursive clustering algorithm to assign the rest of the insertions to different clusters based on sequence similarity using the following scoring scheme: Match +1, mismatch -4, undetermined +0, gap opening -4. Only the inserted segment of the sequence being compared is used for calculation. First, the sequence with the highest global score, aka the representative sequence, will form the top sequence in cluster 1. Pairwise comparisons were then performed across all unassigned insertions in the order of the sorted global scoring results. Any insertions with at least 80% sequence identity with the top sequence of

cluster 1 will be assigned to the same cluster. A new cluster is generated every time when a sequence does not reach the 80% sequence identity threshold. This process iterates until all sequences are assigned to a cluster. More than one representative sequence may be picked per component if the top sequence among the clusters do not overlap based on the reference coordinates. This can happen when two different insertions are grouped into a single component by one insertion with a large reference gap. Of note, if an NUI is found in multiple haplotypes but the insertions are exactly identical, the first sequence in the alphabetically sorted sample ID list will be picked.

*Sequence annotations for filtering*

Representative insertions were annotated with RepeatMasker (-species human) and Tandem Repeat Finder[38] (2 7 7 80 10 50 2000 -f -m -h -d). We removed insertions if they fulfilled at least one criterion:

1. TRF reports > 80% masked bases and less than 100 non-masked bases
2. RepeatMasker reports > 80% combined masked bases in satellites, simple repeat, and low complexity categories; and less than 100 non-masked bases
3. If more than 10 undetermined bases are present, we required at least 50 fully resolved bases in the NUI
4. If more than 100 undetermined bases are present, we required at least 100 fully resolved bases in the NUI

In addition, we removed all insertions if there were more than 5 insertions in any given 200bp windows. These are often associated with regions of low complexity. Additionally, if more than 2 insertions were found in any given 50bp windows, only the longest

insertions were kept. For insertions smaller than 50bp, a breakpoint consistency filter was applied. High-copy number repeat elements such as SINE and LINE were included in our NUI dataset because most of them are short, and they generally had diverged to a degree that allowed pair-end reads to map distinctively[39]. We note that NUIs smaller than 30bp might be too short for RepeatMasker and/or Tandem Repeat Finder to repeat mask effectively.

*Constructing the human diversity reference*

To integrate missing sequences into the HDR, we first combined 127,727 NUIs with the 233 filled genome gap sequences (described in "Reference gap closure" section). Redundant gap-filled sequences were first removed since there are instances when one large assembly scaffold overlaps multiple reference gaps (180 gap-filled sequences remaining). The GRCh38 reference coordinates of the gap-filled sequences were then overlapped with those of the NUIs, and overlapping NUI sequences were removed (127,702 NUIs remaining). Filtered NUIs and gap-filled sequences were then linearly integrated into the GRCh38 core chromosomes. For overlapping and gapped NUIs, the entire sequence from minimum to maximum reference breakpoints were replaced (**Supplementary Figure S2.4b** & **c**).

*Sequence annotation*

NUI repeat type was classified based on the single most represented repeat element in a sequence and it had to make up at least 50% of the overall sequence content as reported by RepeatMasker. Genes were annotated using NCBI Ref Seq All (version

date 2018-8-10; downloaded from UCSC genome browser table). If an NUI overlapped more than one gene, the gene label was prioritized based on the following order: coding exon, 3'UTR, 5'UTR, and intron. Regulatory annotations were performed based on the Ensembl Regulatory Build[40] (ftp://ftp.ensembl.org/pub/release-98/regulation/homo_sapiens/homo_sapiens.GRCh38.Regulatory_Build.regulatory_features.20190329.gff.gz).

*SV genotyping*

The four SVs illustrated in "Detailed genomic analysis of exonic insertions" were genotyped on 70 randomly selected SGDP individuals using Paragraph, a graph-based SV genotyper for short-read sequence data[41]. These samples were aligned to the HDR as described in "whole genome sequence mapping analysis" section. We manually generated the sample manifest file and the required input vcf file describing the four NUIs as deletions. We found that Paragraph generally works well when NUIs have unique breakpoints. However, it did not work as expected when the NUI does not appear to have clearly defined breakpoints (**Supplementary Figure S2.5**).

*Modeling NUI diversity and population structure.*

NUI diversity modelling was performed as described previously[42]. In brief, we performed down-sampling to obtain the average NUI count in each super-population and we proposed a mathematical model to simulate the NUI discovery frequency. NUI projection was performed by fitting the last 10 points of our down-sampled dataset to predict future NUI growth.

*Whole genome sequence mapping analysis*

We downloaded 70 WGS datasets, 10 from each studied population, from the Simons Genome Diversity Project for mapping comparison. Reads were aligned to the GRCh38 core reference (Supplementary notes; see original manuscript) using bwa mem and unmapped read pairs were extracted for quality filters and contamination screening. Specifically, reads were discarded if 1) less than 70% of their bases had a quality score of 20 or above, or 2) blast similarity search against the non-redundant nucleotide database indicated a bacteria or virus as the top hit. The remaining unmapped reads were compared against reads from the HDR alignment results. We additionally calculated the number of reads that were mapped to GRCh38 core reference but failed to map to the HDR.

*Identification of new polymorphic sites using GATK*

We used GATK v3.8-1 haplotypeCaller to identify variants on the 70 SGDP CRAM files which were previously aligned to the HDR using bwa mem. We specified variant calling only in the newly inserted regions in the HDR and we only included variants with a base quality of 20 or above. Of note, we did not evaluate the zygosity accuracy because GATK HaplotypeCaller could not distinguish variants that are hemizygous in loci harboring deletions.

*Transcriptomic diversity through RNA-sequencing*

We randomly downloaded 10 RNA-sequencing datasets from each of the 31 histological types available through GTEx (fallopian tube only had 7 sets of data). Raw FASTQ files

were aligned to the GRCh38 core reference using the STAR[43] aligner and unmapped reads were extracted. Extracted reads were then realigned to the HDR and proper read pairs with MAPQ 255 were kept. Downstream analysis was restricted to the new regions in the HDR and proper read pairs aligning to the genic regions were counted. Since there might be unannotated exons and UTRs residing in these NUIs, we counted all RNA-Seq read pairs as long as they were within the genic boundaries. We filtered out genes with less than 10 combined read pairs from all processed samples.

*Tissue-specific gene expression analysis*

Proper RNA-Seq read pairs aligning to genic regions were used for tissue-specific gene expression analysis and count values were $\log_2$ transformed. Tissue-specificity was determined based on comparing gene expression levels in a given tissue versus the baseline. We used the Welch two sample t-test to perform tissue versus baseline comparisons and we considered a gene to be tissue-specific if the BH adjusted P value is <0.05 and t >3. Gene ontology was performed using the "topGO" package available through R. All GO terms (Fisher unadjusted P value < 0.05) were reported for all tissue types.

*Insertion placement and size validation*

All insertions larger than 1kb without N-gaps were subject to orthogonal validation using Bionano optical mapping technology. Due to the resolution limit of this technology, two labels within a few thousand bases may be merged and the reported reference coordinates will be shifted by one label. To ensure optimal overlapping, 5kb were added

to the two ends of Bionano insertion calls. Assembly-based insertion coordinates were then overlapped with Bionano insertion calls. Insertions were considered concordant if the assembly-based insert size is within either 700bp or 20% of the Bionano insert size, whichever is smaller. Since Bionano maps were not generated for every sample, insertions were deemed validated if their locations and sizes were concordant with calls from different samples. We also removed calls if the BN between-label distance is greater than 500kb. These were excluded because the inferred SV size is far less accurate in sparsely labelled regions.

*Reference gap closure*

A list of reference gaps was downloaded from the UCSC genome browser track. We filtered this list for core chromosomes and removed 64 gaps with types short arm (chr13, 14, 15, 21, 22), telomere, or heterochromatin. Remaining reference gaps (539) were targeted for gap closure using the *de novo* assemblies. Assemblytics SV calls coordinates overlapping these genome gaps were used for gap closure. To close additional gaps, we extracted 10kb upstream and downstream of each reference gap and realign them to the assemblies using minimap2[44] with -c to generate the CIGAR string. Primary alignment with mapping quality 30 or above were used to identify the filled sequences by locating unique anchors flanking the gaps. A reference gap is considered minimized instead of fully closed if small N-gaps still remained in our assembly sequences.

*NUI comparison*

We compared our NUI dataset against two publicly available resources: 1000GP and gnomAD (insertion calls only). We obtained 1000GP data from two sources: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz (SVs) and ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/*.vcf.gz (indels). We ran remap to convert the coordinates from GRCh37 to GRCh38. Similarly, we obtained gnomAD data from two sources https://storage.googleapis.com/gnomad-public/release/3.0/vcf/genomes/gnomad.genomes.r3.0.sites.vcf.bgz (indels) and https://storage.googleapis.com/gnomad-public/papers/2019-sv/gnomad_v2.1_sv.sites.bed.gz (SVs). For small NUI comparisons, we required the insertion breakpoints to overlap by at least 1bp and the size ratio has to be between 0.9-1.1. For large NUIs, we implemented the same overlapping requirement but the size ratio is loosened to 0.5-2. For the 1000GP SV dataset, we only used insertion calls with the SVLEN tag in the INFO column.

## 2.4 Results

### Human genome sequencing, *de novo* assembly, and optical mapping

10x Genomics (10xG) whole genome Linked-Reads were generated in-house on 305 samples (**Supplementary Table S2.1**; see original manuscript). We also included 24 publicly available Linked-Read datasets, of which 22 were obtained from the Polaris project and 2 from the 10xG website. All genomes were sequenced to an average of ~60x coverage, with a mean inferred molecular length of 95.6kb (**Supplementary**

**Figure S2.1, Supplementary Table S2.1, Supplementary Table S2.2**; see end of chapter; see original manuscript). Pseudo-diploid *de novo* assemblies were generated with Supernova, yielding average scaffold and phased block N50s of 39.3Mb and 4.6Mb, respectively (**Supplementary Table S2.2**; see original manuscript). In addition, we supplemented our datasets with previously published PacBio (PB) haploid assemblies on 7 human and 2 hydatidiform mole genomes[22,32,33]. All data presented here were based on a combined set of 338 genomes. In addition, to validate the accuracy of these NUIs, we generated BioNano optical maps for 309 samples using a combination of BspQ1, BssS1, and DLE1 enzymes.

**Identification and characterization of recurrent non-reference unique insertion**

We define as non-reference unique insertions (NUIs) all recurrent insertions ≥10bp, not overlapping segmental duplications, and containing less than 80% low complexity sequences (see Methods). We developed a custom pipeline to select the longest representative NUI in a given genomic locus (**Supplementary Figure S2.2**; see end of chapter). Briefly, we grouped insertions with overlapping reference breakpoints from all individuals and aligned them against each other using multiple alignment. An alignment score based on length and quality of unique sequence was generated for each insertion of the locus. The insertion with the highest alignment score was chosen as the representative NUI.

In total, we identified across different human populations 127,727 recurrent NUIs spanning 18,048,877bp. For every genome sequenced with Linked-Reads, we found on average 14,121 NUIs (2,642,883bp). NUIs are abundant throughout the genome, but,

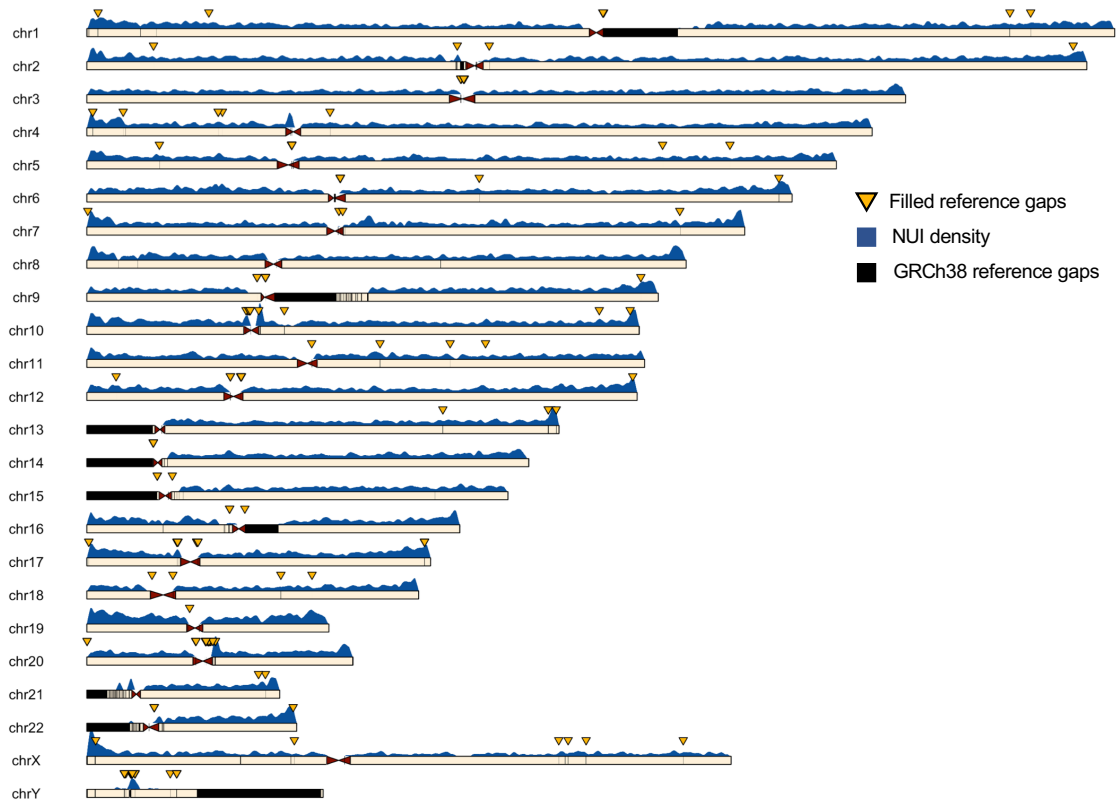



**Figure 2.1: Overview of NUI distributions.** Ideogram illustrating the distribution of NUIs in the 24 chromosomes. Black chromosomal segment refers to GRCh38 reference gaps and yellow down pointing arrows indicate gaps that can be fully closed or minimized with our pipeline. The blue density plot above the chromosomes depicts the number of NUIs identified using a sliding window of 1Mb with a 10-kb step size.

consistent with previous observations[22,23], they are biased towards the sub-telomeric and peri-centromeric regions of the chromosomes (**Figure 2.1**). High-copy repeats elements Alu and LINE contributed to two discernible peaks at 300bp and 6kb in the NUI size distributions (**Supplementary Figure S2.6**; see end of chapter). While 86.9% (110,996) of the NUIs were small (<50bp), they only made up 9.7% (1.7Mb) of the overall cumulative length (**Supplementary Figure S2.6d**). The number of NUIs ≥350bp (a size generally beyond the reach of short read sequencing) was 7,801, with a cumulative length of 14.3Mb (79.5% of total). NUIs (≥1kb) account for 12.2Mb (67.9%) of the cumulative length even though there were only 3,361 of them (2.6% of total).

Since representative NUIs are the top scoring sequences selected based on multiple alignment results, we assessed the distributions of these selected NUIs stratified by sequencing platforms and sample populations. PB sequences contributed 2.7% to the small NUI total bases (<50bp), which roughly equal to the proportion of PB sample in the collection (**Supplementary Table S2.3**; see original manuscript). In contrast, since our pipeline favors sequences that can be fully resolved, PB sequences contributed 20% of the large NUI category (≥50bp). Of note, 10xG assemblies often shared these insertions but, due to their inherent limitation in read length, they failed to resolve the entire NUIs in low complexity regions. We observed that Africans had the highest representative NUI contribution proportional to their sample size (**Supplementary Table S2.4**; see original manuscript). Africans contributed almost twice as many NUIs per capita as individuals from other groups.

To quantify NUI diversity and project further discovery in the five super-populations, we performed saturation analysis on the large NUIs. The East Asian population was the first to approach an asymptote while the African population had the steepest trend (**Supplementary Figure S2.7a**; see end of chapter). This result is within expectations because our East Asian samples were predominantly Southern Han Chinese (Taiwanese) samples (154 out of 180). Based on our analyses, sequencing an additional African sample with 10xG Linked-Reads would add an estimate of 94 NUIs, but an East Asian would add only 23 (**Supplementary Figure S2.7b**; see end of chapter).

To evaluate NUI specificity, we compared the large NUIs ≥1kb with no undetermined bases chosen for inclusion in the HDR using Bionano optical maps. In all,

86.2% of these variants were concordant with insertions identified by Bionano optical mapping on the basis of size and location (**Supplementary Table S2.5**; see original manuscript). Since Bionano insertions may overlap multiple SVs in close proximity and the insert size is the overall sum of the total size change, the reported insertion size may not be directly comparable. When we loosened the criteria to include NUIs whose sizes did not agree with Bionano optical maps, the concordance rate increased to 93.9% (**Supplementary Table S2.5**; see original manuscript).

Within our NUI dataset, 84.4% and 63.5% of the small and large NUIs, respectively, were never discovered in any of the PacBio samples (**Supplementary Figure S2.8a**). A small proportion of the large NUIs (1.3%) were found only in PacBio samples, suggesting that some of these NUIs were only resolvable by long reads. To further evaluate the novelty of our dataset, we compared our NUIs with two publicly available resources: 1000GP and gnomAD (**Supplementary Figure S2.8b; Supplementary Table S2.6**, see original manuscript). We found that 69.3% of the large NUIs were novel, as compared to 53.1% of the small NUIs.

**Non-reference unique insertions affecting genic and regulatory elements**

To assess the functional consequences of NUIs, we searched for those affecting genic and regulatory elements. Using the NCBI RefSeq genes and gene predictions database, we identified 59,332 NUIs overlapping genic loci. Among these, 521 recurrent NUIs are in coding exons, 1,738 in 3'UTR, 1,200 in 5'UTR, and 55,878 in introns (**Supplementary Table S2.7**; see original manuscript). Evolutionary constraints acting on these NUIs manifest as a very strong 3bp periodic trend in small exonic NUIs

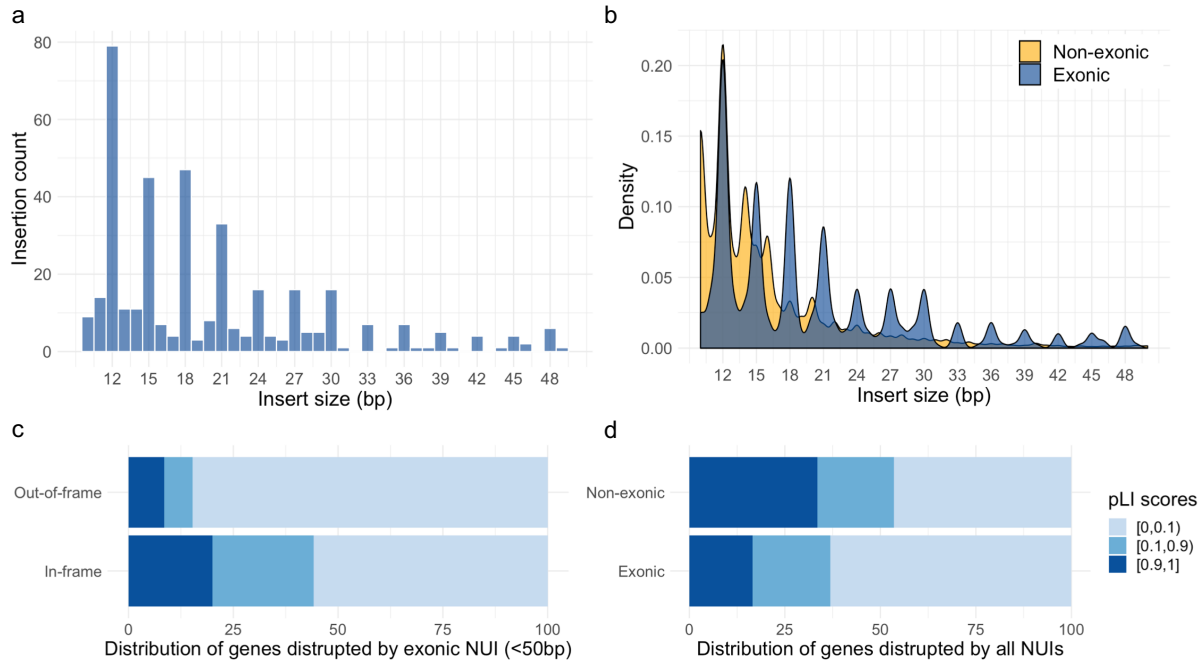



**Figure 2.2: Evolutionary constraints on exonic NUIs.** (a) Histogram of small NUIs (<50bp) disrupting exons. The 3bp periodic trend shows that the majority of these events are in-frame, suggesting that that a strong evolutionary constraint is in place to preserve the structural and functional integrities of the proteins. (b) Density plot comparing the 3bp periodic trend in exonic (blue) versus non-exonic (yellow) NUIs. (c) Horizontal barplot showing the pLI scores distributions of genes whose exons are disrupted by in-frame or out-of-frame small NUIs. (d) As in c but exonic and non-exonic NUIs of all size ranges are compared.

(<50bp) (**Figure 2.2a**). Other classes of NUIs did not exhibit this periodic trend (**Figure 2.2b**), suggesting preservation of exonic insertions that do not alter the coding frame and so are presumably less deleterious. Since NUIs were identified in reasonably healthy individuals, we hypothesized that the out-of-frame exonic NUIs were enriched in genes that were more tolerant of changes. As expected, only 5% of genes affected by out-of-frame small exonic NUIs are extremely intolerant to loss-of-function changes (with pLI scores ≥0.9) whereas 20% of genes affected by in-frame small exonic NUIs are extremely intolerant of loss-of-function changes (**Figure 2.2c**; Chi-squared p = 0.0002461, df = 2, Chi-squared = 16.619). We then performed the same analysis by

partitioning all NUIs into exonic and non-exonic. This revealed that a smaller proportion of exonic NUIs were in genes that were intolerant to loss-of-function changes while non-exonic NUIs were more likely to be found in both ends of the pLI spectrum (**Figure 2.2d**; Chi-squared p = 9.605e-13, df = 2, Chi-squared = 55.343).

We then assessed NUIs affecting regulatory features defined by the Ensembl regulatory build. Overall, we found 22,557 NUIs intersecting at least one regulatory feature (**Supplementary Table S2.8**; see original manuscript), again underlining the potential impact of these NUIs, which cannot be elucidated using the current reference genome.

**Detailed genomic analysis of exonic insertions**

Integrating NUIs into the reference genome can benefit genome annotation in multiple ways. To illustrate the key advantages of using the longest haplotype as the reference structure, we took a closer examination of four insertions affecting coding sequences. The first is a 2,090bp insertion located 4bp upstream of the start codon of *METTL21C*, a skeletal muscle-specific lysine methyltransferase that plays an important role in regulating protein degradation[45] (**Figure 2.3a**). This common insertion adds a new start codon, along with 19 amino acids that are in-frame with the existing exon1. A BLAST similarity search using the non-redundant nucleotide database revealed an identical amino acid sequence in the same gene in the chimpanzee RefSeq protein database (ID: XP_009425476.2), indicating that this insertion is conserved and thus likely to be ancestral and functional. Based on the GTEx RNA-Seq evidence in skeletal muscle, we were able to verify the presence of this sequence as part of the mRNA transcript

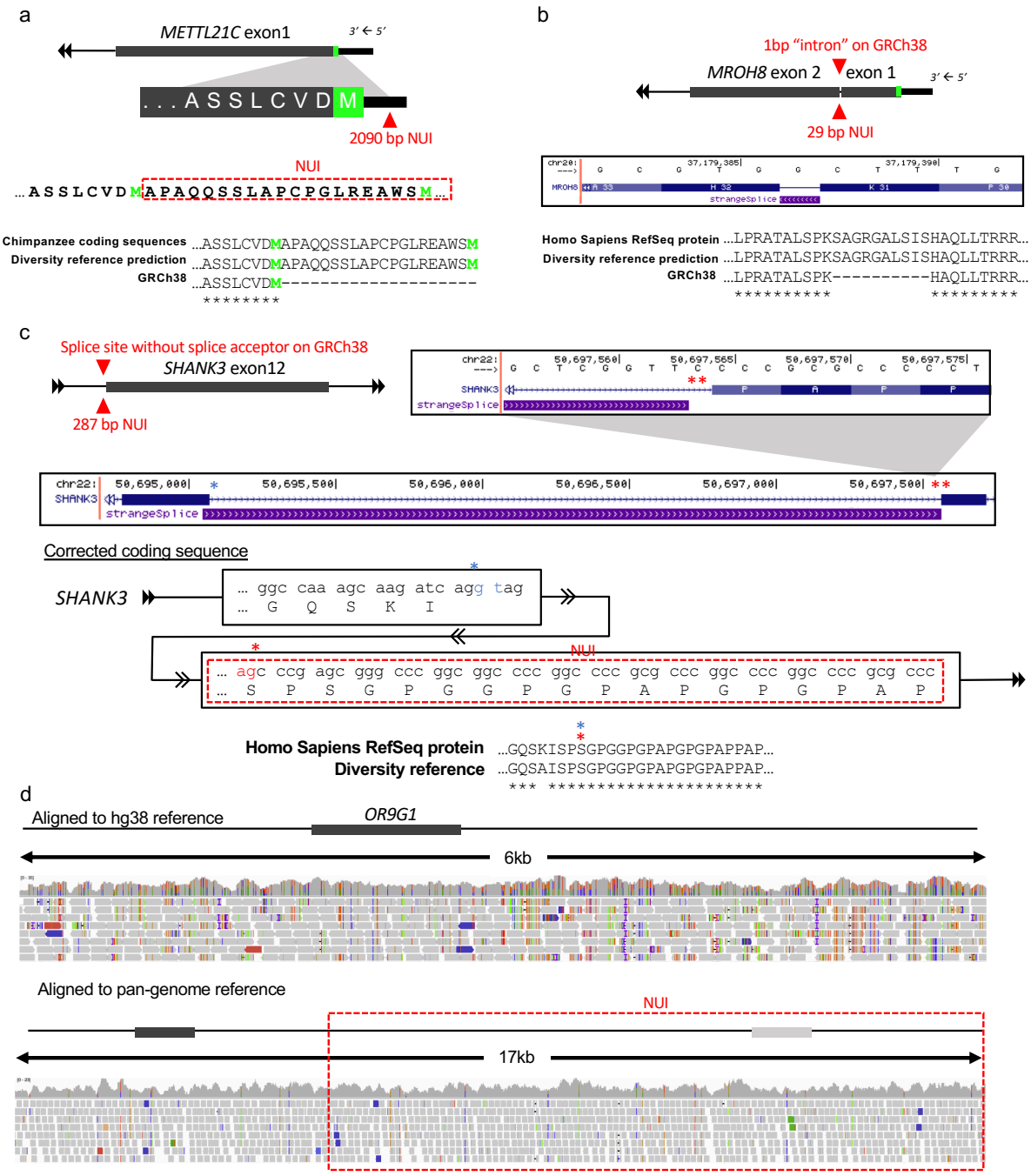



**Figure 2.3: Examples of NUIs affecting exonic annotations.** (a) A 2090bp insertion was identified 4bp upstream of the annotated start codon based on GRCh38 annotation. Upon integrating the NUI, the open reading frame is extended by 20 amino acids with a new methionine that could now be served as the start codon. Identical coding sequences were identified in the chimpanzee RefSeq protein database. Neon green denotes the annotated and the new start codon methionine. (b) A 29bp NUI was identified precisely at that 1bp "intron" between exon 1 and exon 2 of *MROH8*. It restores a single reading frame by merging the two exons. The new coding sequence

has a match in the human RefSeq protein database. (c) A 287bp NUI was located right at the 3' end (splice acceptor) of the intron between exon 11 and exon 12 of *SHANK3*. The UCSC genome browser screenshot showed a non-canonical splice site that did not match with the corresponding splice donor. Upon integrating the sequence boxed in red, the bona fide splice acceptor was restored. The full-length coding sequence had an almost identical match to a record in the human RefSeq protein database. Blue asterisk, splice donor; single red asterisk, restored splice acceptor; double red asterisks, original splice acceptor. (d) An example demonstrating the excessive number of variants caused by missing sequences. This missing sequence also contains an ortholog of *OR9G1* that is not present in the GRCh38 core reference. Gene structures are not drawn to scale.

(**Supplementary Figure S2.9**; see end of chapter). We also identified two alternate splicing events in the inserted sequence upstream of the original start codon. By restoring two different splice acceptors, we discovered the corresponding splice donors and novel exons upstream of the known *METTL21C* gene boundary. Additionally, we genotyped this insertion in 70 SGDP samples and found that 56/70 (80%) samples carried the homozygous NUIs, 10/70 (14.3%) were heterozygous, and 4/70 (5.7%) were homozygous ref allele. All Africans have the homozygous NUI allele at this locus (**Supplementary Figure S2.10; Supplementary Table S2.9**, see original manuscript).

The second example is a 29bp insertion in *MROH8*, a gene known to be associated with Alzheimer's disease[46] (**Figure 2.3b**). The current reference assembly has a 1bp "intron" between exon 1 and exon 2. By placing the NUI in its proper genomic context, the two exons were merged into a single reading frame with 10 amino acids in place of the "exon". Results from BLAST reveal that this coding sequence is identical to the RefSeq *MROH8* sequence record (ID: NP_689716.4). GTEx RNA-Seq reads confirmed the presence of this sequence in exon 1 of *MROH8* transcript (**Supplementary Figure S2.11**; see end of chapter). All genotyped individuals carried

the homozygous NUI allele (**Supplementary Table S2.9**), indicating that the reference version is likely a private mutation or a result of an assembly error.

In the third example (**Figure 2.3c**), missing sequences can prevent the correct splice site interpretation. In the reference genome, there is a non-canonical splice acceptor at the 3' end of the intronic sequence between exons 11 and 12 of *SHANK3*, a prominent candidate gene for autism, schizophrenia, Phelan-McDermid syndrome, and other disorders[47-51]. We identified a 287 bp NUI precisely at the non-canonical splice acceptor. This NUI extends the existing reading frame and restores the canonical splice acceptor that was incorrectly annotated on the reference genome (isoform #2 in **Figure 2.3c**). The correct sequence elongates exon 11 by 15 amino acids. On the 5' end of exon 11, the bona fide splice acceptor—AG—is restored. This coding sequence is supported by RefSeq entry (ID: NP_277052.1). Using the HDR, RNA-Seq reads from GTEx can now be mapped precisely across this junction (**Supplementary Figure S2.12**; see end of chapter). In this example, genotyping was complicated by the lack of clear SV breakpoints. The SV detection tool Paragraph reported 63/70 homozygous NUI and 7 ambiguous cases (**Supplementary Table S2.9**). Upon manual evaluation of the sequencing data, however, samples that were deemed homozygous NUI by Paragraph also had significant drop in coverage, suggesting that that these were likely to be heterozygous NUI (**Supplementary Figure S2.5**).

The final example (**Figure 2.3d**) demonstrates that missing sequences can lead to spurious variant calls by forcing reads to map to a suboptimal location. A 6kb OR9G1-containing locus was duplicated in the studied genome. The two loci are sufficiently divergent that short reads can map uniquely to the respective copies, and

the resulting alignment was substantially improved using the HDR. When aligning to the GRCh38 reference, numerous variants are falsely identified because reads derived from both copies were forced to align to the only locus on the reference genome. All genotyped samples carried the homozygous NUI allele, indicating that the reference sequence is either a mistake or a private mutation (**Supplementary Table S2.9**).

**Mapping improvement and the identification of novel polymorphic site in the Human Diversity Reference**

To examine the extent to which the HDR can improve the alignment rate of sequence reads to the human genome, we obtained 70 Simons Genome Diversity Project (SGDP) WGS datasets and aligned sequencing reads to the human genome reference and to the HDR. On average, 402,573 previously unmapped reads from each individual could be aligned to the HDR (**Supplementary Table S2.10**; see original manuscript) for any given genome sequenced to ~40x coverage depth, indicating that a significant proportion of sequences that are invisible in GRCh38 can now be used directly for identification of genetic polymorphisms. We also assessed the number of perturbed reads as a result of aligning to the HDR. The number of lost sequencing reads averaged at 63,619 per sample (**Supplementary Table S2.10**; see original manuscript). In other words, for every gained read, only 0.16 read is lost.

We applied GATK HaplotypeCaller to ascertain SNPs and indels within the NUIs. We found that each SGDP sample harbored an average of 27,658 SNPs and 11,219 indels in the NUIs (**Supplementary Table S2.11**; see original manuscript). In both variant classes, Africans have 5,238 and 929 more SNPs and indels than the non-

African averages (SNPs ANOVA P value = 8.53e-14, df = 6, F value = 22.05; followed by Tukey (adjusted P values): America-Africa (1.57e-11), CentralAsiaSiberia-Africa (3.94e-9), EastAsia-Africa(5.95e-11), Oceania-Africa (2.08e-11), SouthAsia-Africa (8.20e-8), WestEurasia-Africa (6.67e-9), SouthAsia-America (3.98e-2); Indels ANOVA P values = 0.0066, df = 6, F value 3.324; followed by Tukey (adjusted P values): America-Africa (0.0418), Oceania-Africa (0.00264); **Supplementary Figure S2.13**; see end of chapter), which is consistent with the known genetic diversity in Africa.

**Transcriptomic analysis using previously unmapped reads**

We leveraged the GTEx RNA-Seq dataset to assess the functional significance of the NUIs across different tissue types. We first aligned all RNA-seq reads to the core GRCh38 reference genome; we then extracted the unmapped reads for re-alignment to the HDR. We identified unique RNA-seq read alignment in 4,781 genes across all tissues, meaning that a significant proportion of all human genes have unannotated exons or expressed functional elements that can now be recovered with the HDR. Of these, 3,689 (77%) genes have OMIM annotations, providing further evidence that these genes have known phenotypic relationships that are potentially relevant to clinical diagnosis.

**Tissue-specific gene expression analysis**

We sought to explore whether the RNA-Seq reads that align to the HDR but not the canonical reference can inform biological relationships between genes and the specialized functions of different tissue types. Here we define tissue-specific gene

expression as those showing significantly enhanced expression levels compared to the baseline across tissues (see Methods). Of all the tissues tested, the testis has the highest number of tissue-specific genes with 115 protein-coding genes and 26 non-coding RNA (**Figure 2.4a**). This result is consistent with previous literature findings[52]. Majority of the testis-specific genes play significant roles in the regulation of cell division and the formation of male gametes (**Supplementary Table S2.12**; see original manuscript). In contrast, the stomach had the fewest number of tissue-specific expressed genes (3 protein-coding genes). Two of the three genes are MUC1 and MUC6. Both genes are key members of the glycoprotein Mucin family genes that make up the major component of the gastrointestinal mucus gel layer[53].

To further explore the biological significance of tissue-specific isoforms encoded by NUIs, we performed a detailed gene ontology (GO) analysis using tissue-specific genes. Not surprisingly, GO profiles showed highly relevant biological functions with respect to the tissue origins. As examples, GO terms such as chemical synaptic transmission, neurotransmitter transport, and nervous system development were among the most significant ones in the brain. In the spleen, immune-related processes were significantly enriched while epidermis development genes were among the top in the skin (**Figure 2.4b**). In the muscle, cellular components such as sarcomeres and other myofibril structures yielded the strongest signals. All significant and suggestive GO terms are summarized in **Supplementary Table S2.12** (see original manuscript). Taken together, these results highlight the transcriptomic diversity found within the NUIs, the biological relevance of these missing sequences, and the importance of incorporating them into the HDR for comprehensive functional analysis.

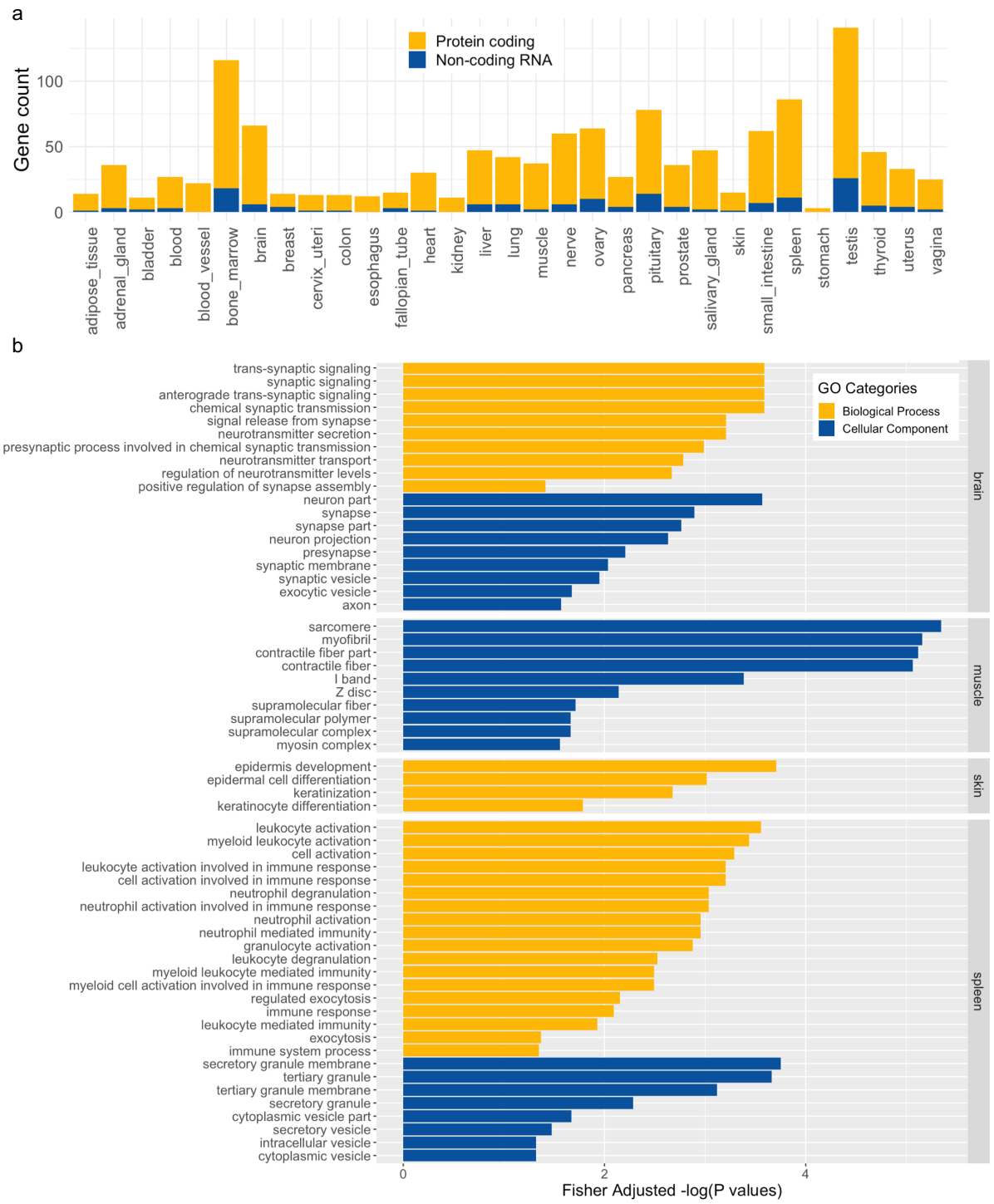



**Figure 2.4: GTEx RNA-Seq analysis using previously unmapped reads.** (a) Barplot showing the total number of tissue-specific expressed genes across 31 different histological sites. Yellow—protein coding genes; blue—non-coding RNA. (b) Top gene ontology terms in four different histological sites.

### Gap closing on the reference genome

We targeted 539 gaps in the core reference assemblies for gap closure. We fully filled 220 gaps and minimized an additional 13 (43%) where we were able to identify high confident anchors flanking both ends of the gaps (**Figure 2.1; Supplementary Table S2.13**; see original manuscript).

## 2.5 Discussion

The human reference genome is an invaluable resource in biomedical research, given its quality and the wealth of information it contains. However, in the Human Genome Project completed in 2003, technical limitations restricted sequencing to the euchromatic regions of the genome. Similarly, the prohibitive cost and effort of sequencing at the time made it impractical to generate more than one reference-quality human genome at that time. The current human reference genome is a composite genome with every locus deriving from a single haplotype, which does not provide a complete picture of the genetic diversity of human populations. Recent advances in long read sequencing allow *de novo* assemblies of the human genome, but the cost is still high for routine studies. Massively parallel short read sequencing remains a cost-effective approach to analyze the human genome, but its utility largely depends on a well annotated human reference genome.

To bridge this gap, we sequenced and analyzed 338 *de novo* assembled human genomes and identified a significant number of Non-reference Unique Insertions (NUIs) that are missing from the human reference genome. Building upon earlier efforts to discover missing sequences, our work not only identifies these NUIs but also places

them uniquely in the human reference genome. By creating a linear Human Diversity Reference (HDR) that incorporates all 127,727 NUIs discovered to-date, we have created a new reference that captures the global genetic diversity and expands the interpretable space of the human genome. As examples, we demonstrated that previously unmapped RNA-Seq reads from 4,781 genes can be aligned to the HDR and that 22,557 NUIs directly disrupt known regulatory features of the genome. These results have important implications. First, inclusion of NUIs in the human reference genome enables more accurate gene and functional element annotations. This is because NUIs are annotated based on the local sequence context, their effects on neighboring genes or regulatory elements cannot be discerned without knowing their location in the genome. Non-reference sequences constructed with short reads in earlier efforts cannot be confidently anchored on the reference because the scaffolds are generally very fragmented[54,55], so their functional impact are undetermined. Second, the HDR provides a more comprehensive reference without the need for new software tools or shifting the current paradigm of alignment-based analysis; this is possible because NUIs are insertions that simply extend the length of the human genome assembly but do not alter its linear structure. Third, integrating these NUIs into the HDR provides new opportunities to detect variants in previously intractable sequences and reduces the number of incorrectly called sequence variants due to misalignments. In order to capture more genetic diversity in future studies, our population analysis strongly suggests that sequencing additional African genomes would yield the greatest increase in the number of NUIs. Finally, our strong emphasis on incorporating unique sequences is pivotal as adding in highly repetitive sequences can increase alignment ambiguity,

especially since short-read sequencing has been the standard for genomic analyses for years.

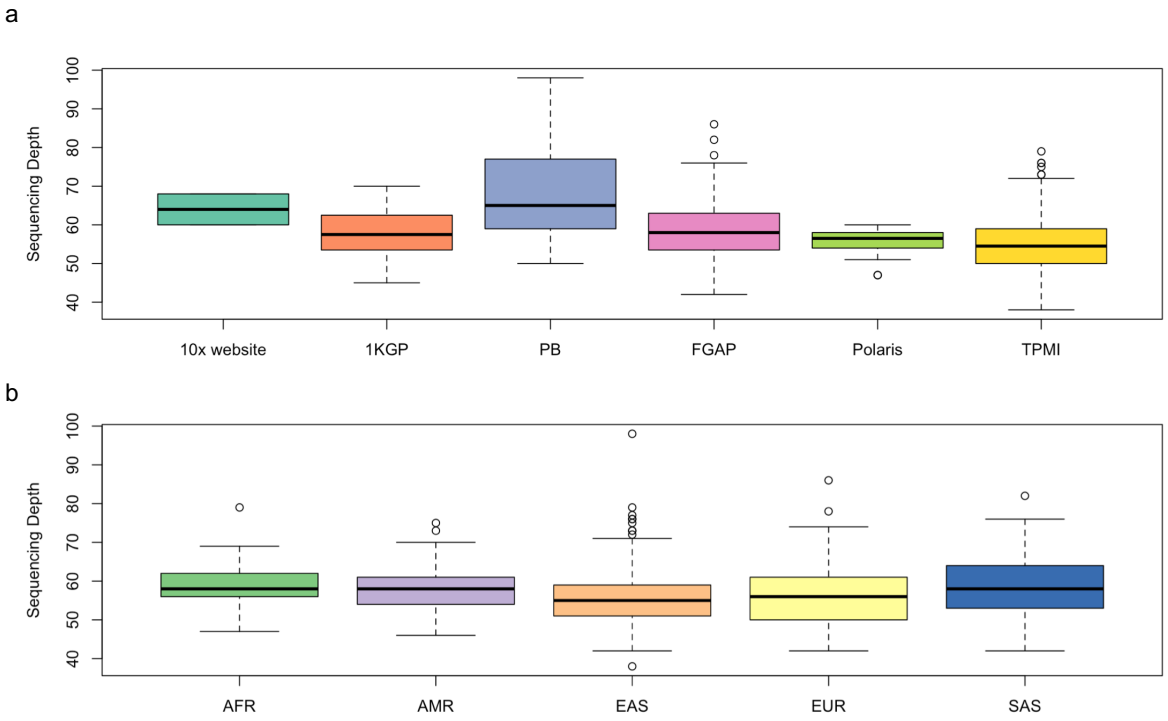
We explored the utility of creating a population-specific reference genome as part of the study using the 154 Han Chinese samples that we sequenced and assembled. We found that the Han Chinese population reference genome indeed captures the population's genetic diversity of the population, but the HDR that encompasses additional populations performs just as well when WGS from Chinese individuals are aligned against it.

While our approach is fast and cost-effective in identifying a broad spectrum of genetic variations across the world's populations, the *de novo* assembler cannot fully reconstruct contigs in regions of low complexity or extreme GC content due to the inherent limitations of Illumina sequencing. As a result, the NUIs incorporated into the HDR contain about 6% of undetermined bases, or N-gaps. Furthermore, the HDR does not tackle the complex (large and highly repetitive) regions of the genome because short-read sequence alignment and SV analysis break down in these regions. In fact, the highly variable nature of these regions means that there are numerous haplotypes in the population, making it impossible to infer the correct configuration of sequences in these regions of an individual based on alignment[29]. Simply put, it remains unclear how the full representation of these highly complex sequences can benefit genomic analysis when short-read data—the current standard in sequencing—cannot align well to a single locus on the reference. If analysis of complex regions is warranted, sequencing and mapping on multiple platforms to construct a *de novo* assembly has to be performed[12].

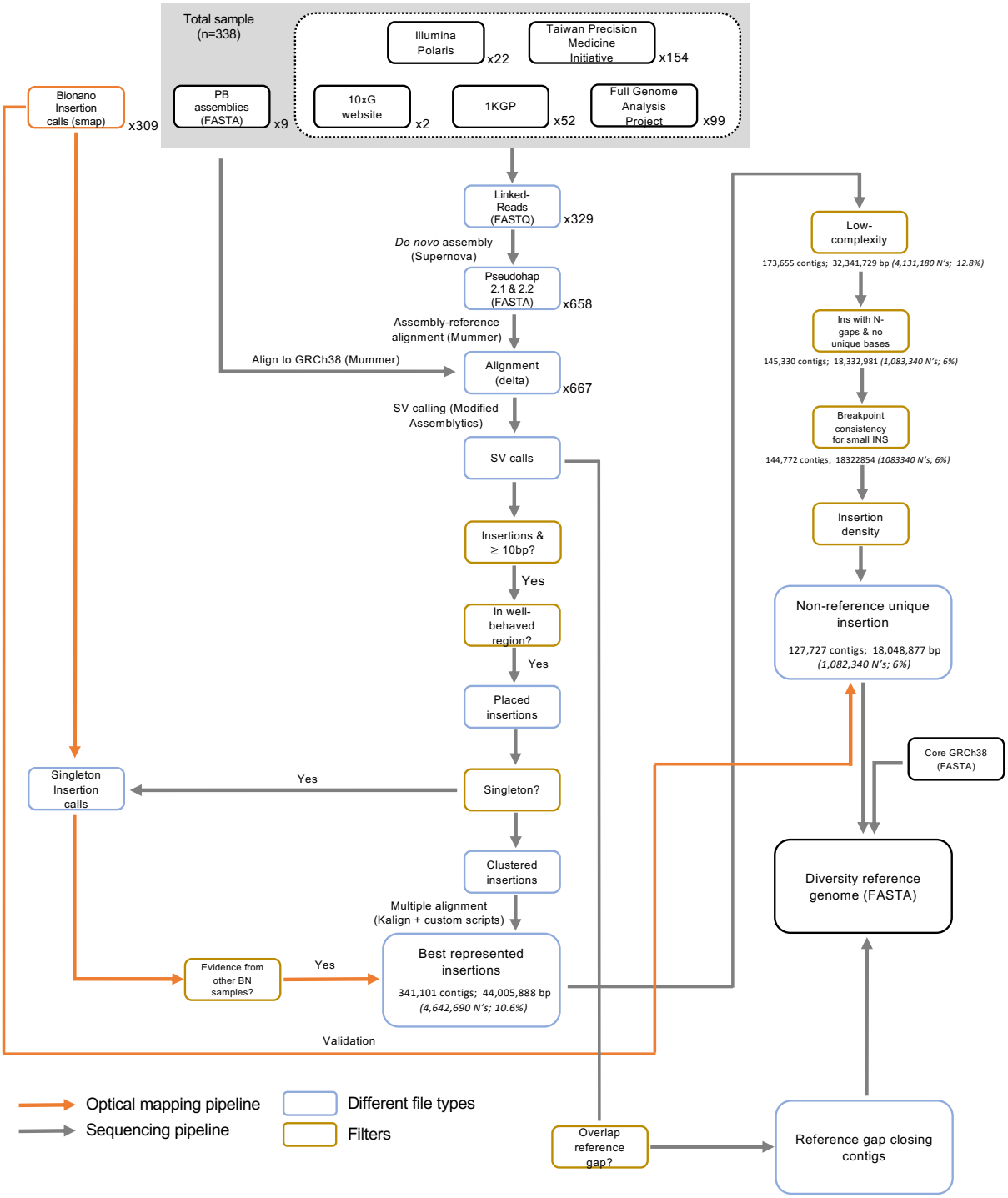
As opposed to a graph-based reference structure, the HDR is not intended to represent all genetic variations found in the human genome in a single compact format (please see Sherman et al[56] for current challenges associated with graph-based references). Instead, the HDR serves as a bridge between the present GRCh38 and a future when graph-based reference structure is fully realized. In fact, it is possible to adopt the longest linear haplotype as a permanent reference coordinate system, and other variations can be added incrementally to this fixed set of coordinates in a graph structure. This way, newly discovered variants will not disrupt the existing coordinate system but can also be represented in one coherent entity.

With the unprecedented increase in size and scope of genome sequencing studies, there is an urgent need for an improved reference that can capture the global diversity found in different human populations. Our proof-of-concept reference representation will allow researchers to identify biologically relevant polymorphisms beyond what is currently detected. The HDR will also guide the design of capture probes for whole exome sequencing and gene panel sequencing, as better gene annotations will likely identify new exons for some of the key genes in disease settings. Most importantly, our HDR will not only benefit all ensuing WGS datasets, but it can also enhance the value in the millions of existing sequencing datasets that are readily available for re-analysis. Taken together, we have demonstrated that our long-read sequencing and mapping approach is an efficient and cost-effective way to generate hundreds of *de novo* assemblies from diverse human populations, thereby providing a path forward for future diagnostic and therapeutic opportunities.

## Supplementary Figures

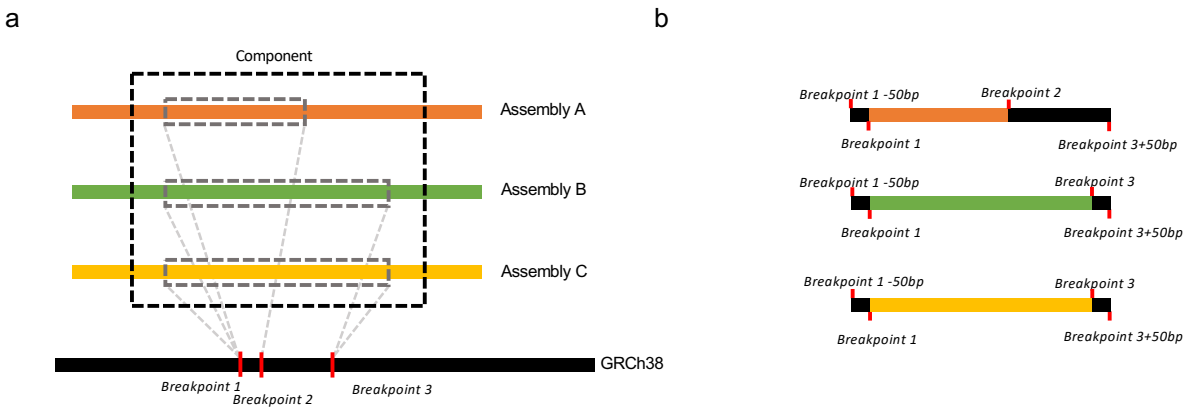



**Supplementary Figure S2.1: Sequencing depth across all samples.** Boxplots depicting sample sequencing depth stratified by (a) data sources and (b) populations. PB-PacBio; FGAP-Full Genome Analysis Project; TPMI-Taiwan Precision Medicine Initiative; AFR-Africans; AMR-Admixed Americans; EAS-East Asians; EUR-Europeans; SAS-South Asians.

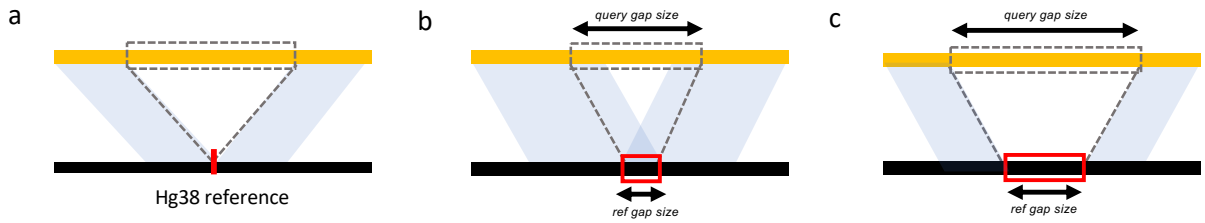



**Supplementary Figure S2.2: NUI calling flowchart.** 10xG Linked-Reads were assembled into pseudo-diploid *de novo* assemblies using Supernova, which were aligned to the GRCh38 reference genome. PacBio assemblies were also aligned at this stage. Insertions were identified using a modified version of Assemblytics. We filtered out calls based on size, genomic location, and sample frequency. The resulting insertions were merged across samples. Multiple alignment was performed to identify the single best insertion per genomic locus. If an insertion was initially identified in only

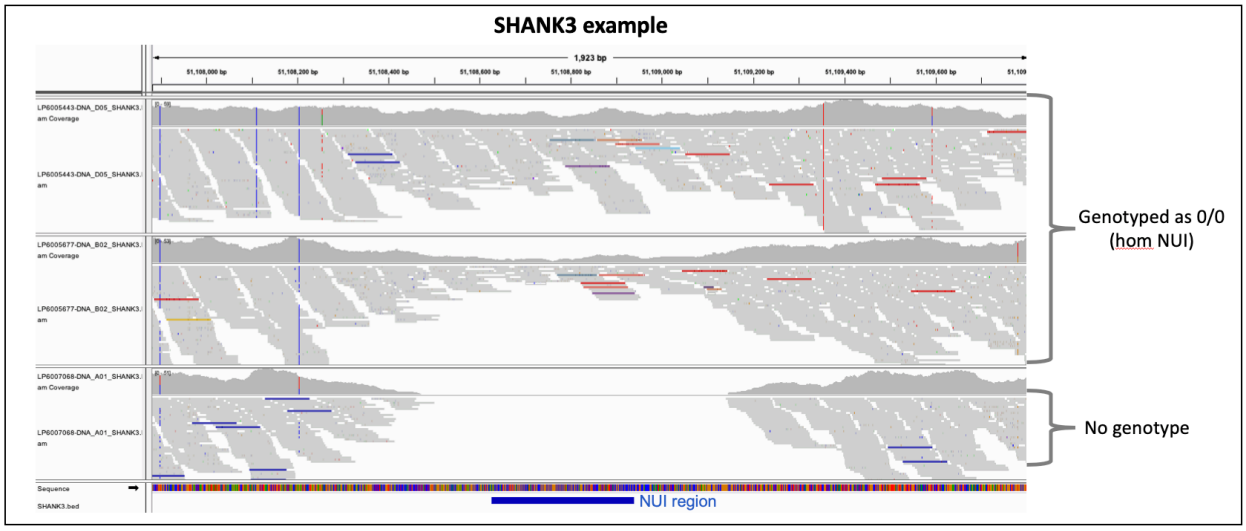
one sample but we found Bionano evidence in multiple samples, these were added back to our NUI call set. All insertions underwent another round of filtering and the ones that passed all the filters were included in the final call set. Additionally, we attempted to close reference gaps using the SV calls generated by Assemblytics. Gaps filled by our pipeline were also integrated into our Human Diversity Reference. Well-behaved regions were defined by the two SV filter lists provided by 10xG (see Methods).



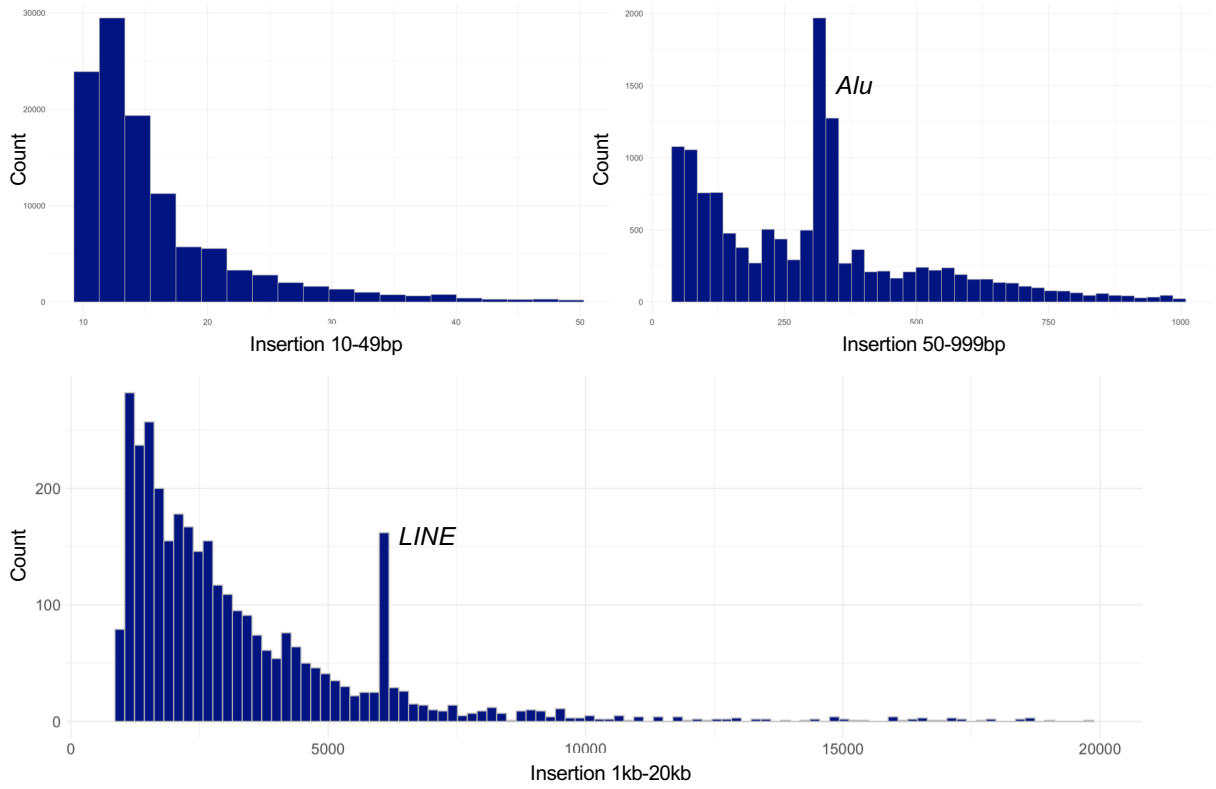

**Supplementary Figure S2.3: Schematic diagrams describing NUI grouping methods.** (a) Any insertions whose breakpoints overlap on the reference are grouped into a component. (b) Reference paddings were added to the individual insertion sequence to increase multiple alignment accuracy.



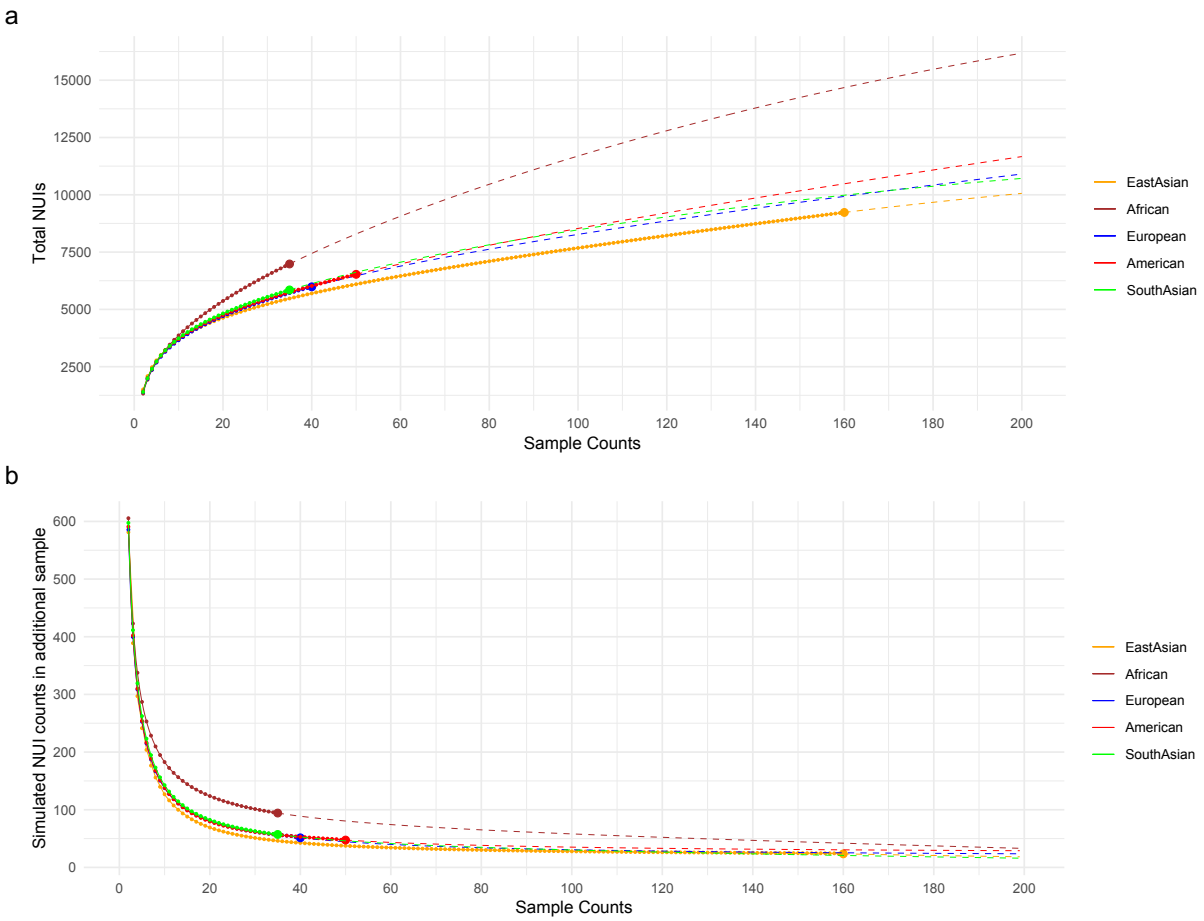

**Supplementary Figure S2.4: Schematic diagrams illustrating the three types of NUIs categorized based on alignment configurations.** Diagrams showing (a) a simple NUI where there is neither overlap nor gap on the reference between the two alignment blocks shaded in light blue; (b) an overlapping NUI where the two alignment blocks overlap on the reference; and (c) a gapped NUI where there is a gap on the reference flanked by the alignment blocks.



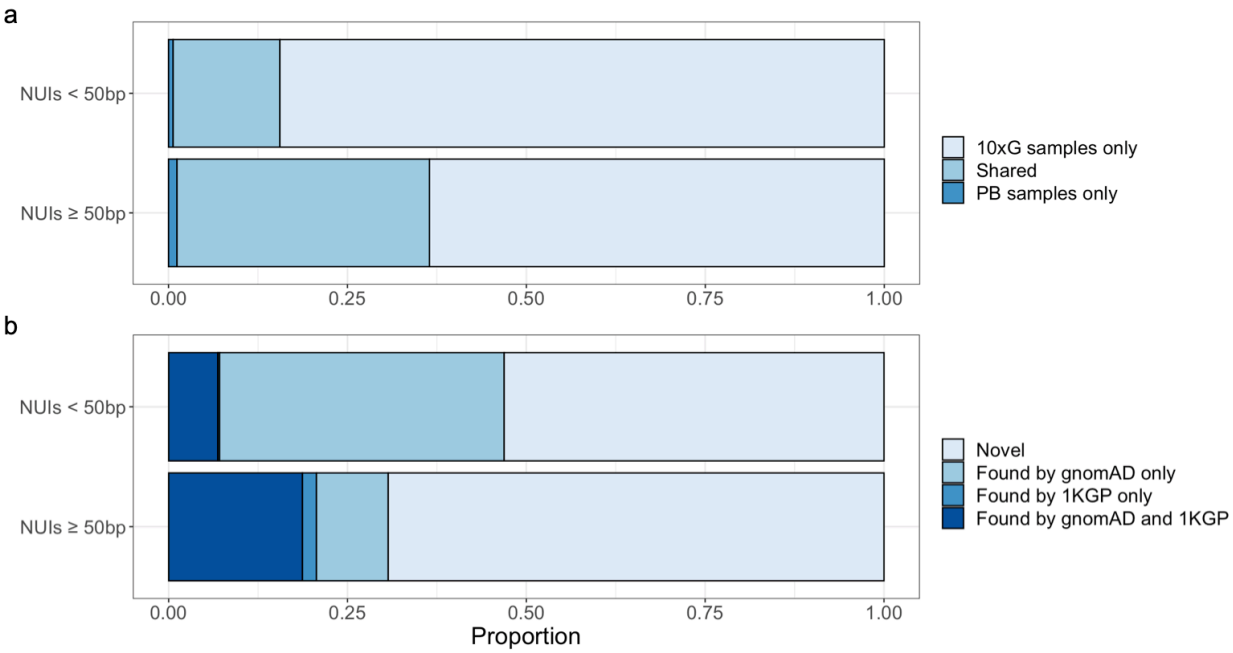

**Supplementary Figure S2.5: Ambiguous Paragraph genotype.** Paragraph reported homozygous NUI (0/0) in the top two samples. Upon manual evaluation, it appears that the second sample has a significant drop in sequence read coverage, and thus is likely to be heterozygous. The last sample could not be genotyped as reads were not aligned at the NUI breakpoints.




**Supplementary Figure S2.6: NUI size distribution.** Barplots illustrating the distributions of NUIs that are (a) 10-49bp in size; (b) 50-999bp in size; and (c) 1kb-20kb in size. Alu and LINE signature peaks were labelled. NUIs larger than 20kb were omitted.

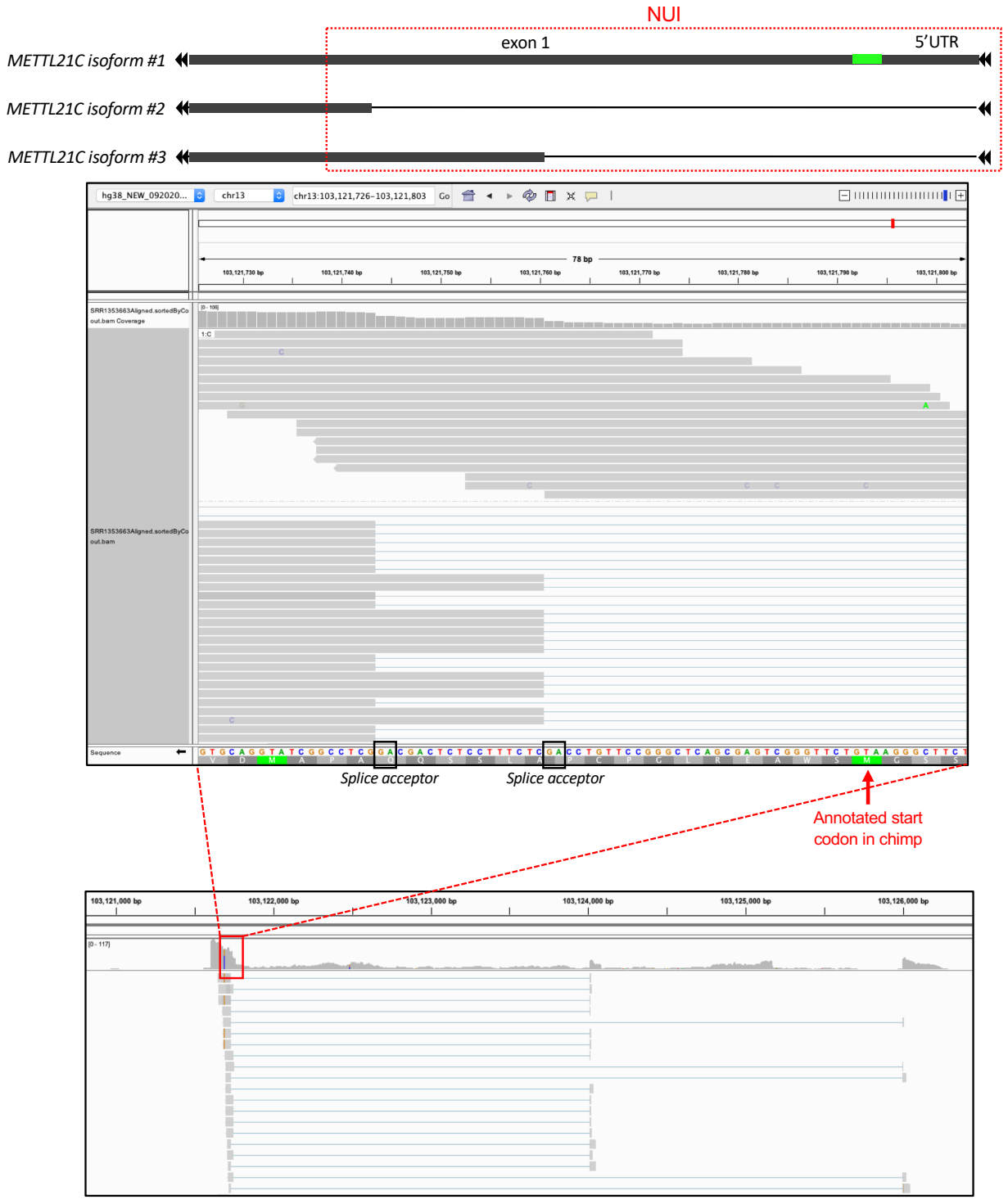



**Supplementary Figure S2.7: NUI saturation analysis.** (a) NUI projection depicting the expected total NUI counts versus number of analyzed samples. (b) Simulated new NUI count when an additional sample is sequenced. Only NUIs ≥50bp were used in these two analyses.



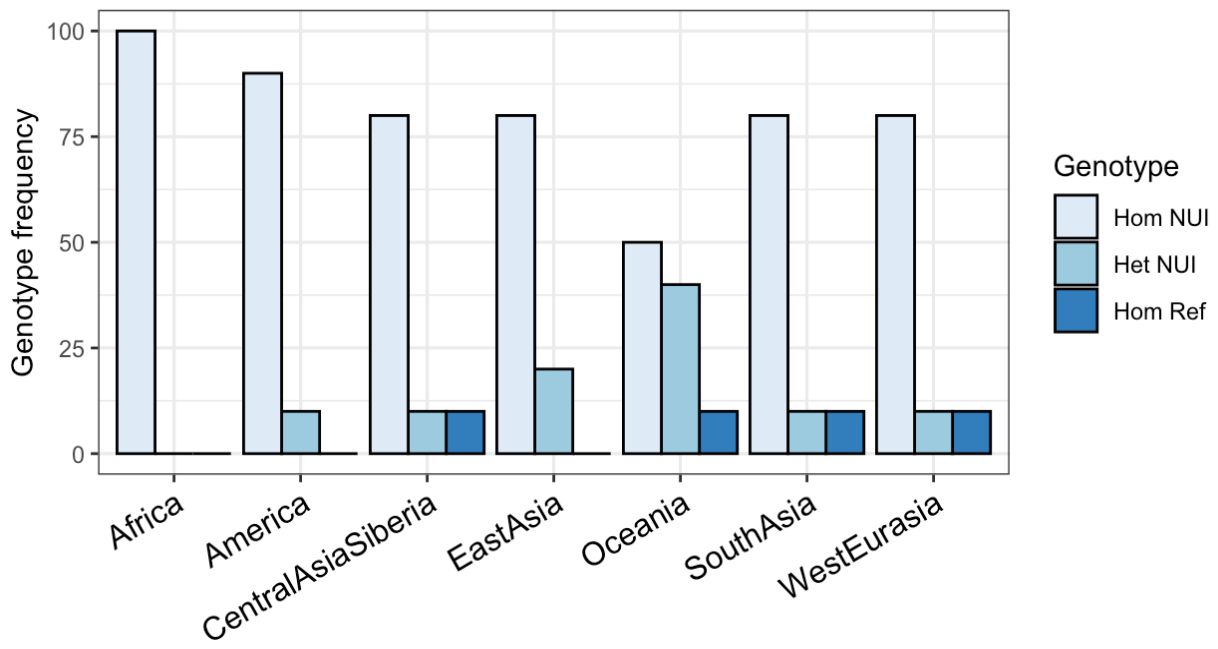

**Supplementary Figure S2.8: NUI dataset comparisons.** (a) Barplot depicting the proportions of NUIs found in PacBio samples analyzed through our pipeline. (b) Barplot depicting the proportions of NUIs found in either gnomAD, 1KGP, or both.

**GRCh38 insertion coordinates** : chr13:102694503-102694508
**Corresponding pan-genome reference coordinates**: chr13:103121739-103123834



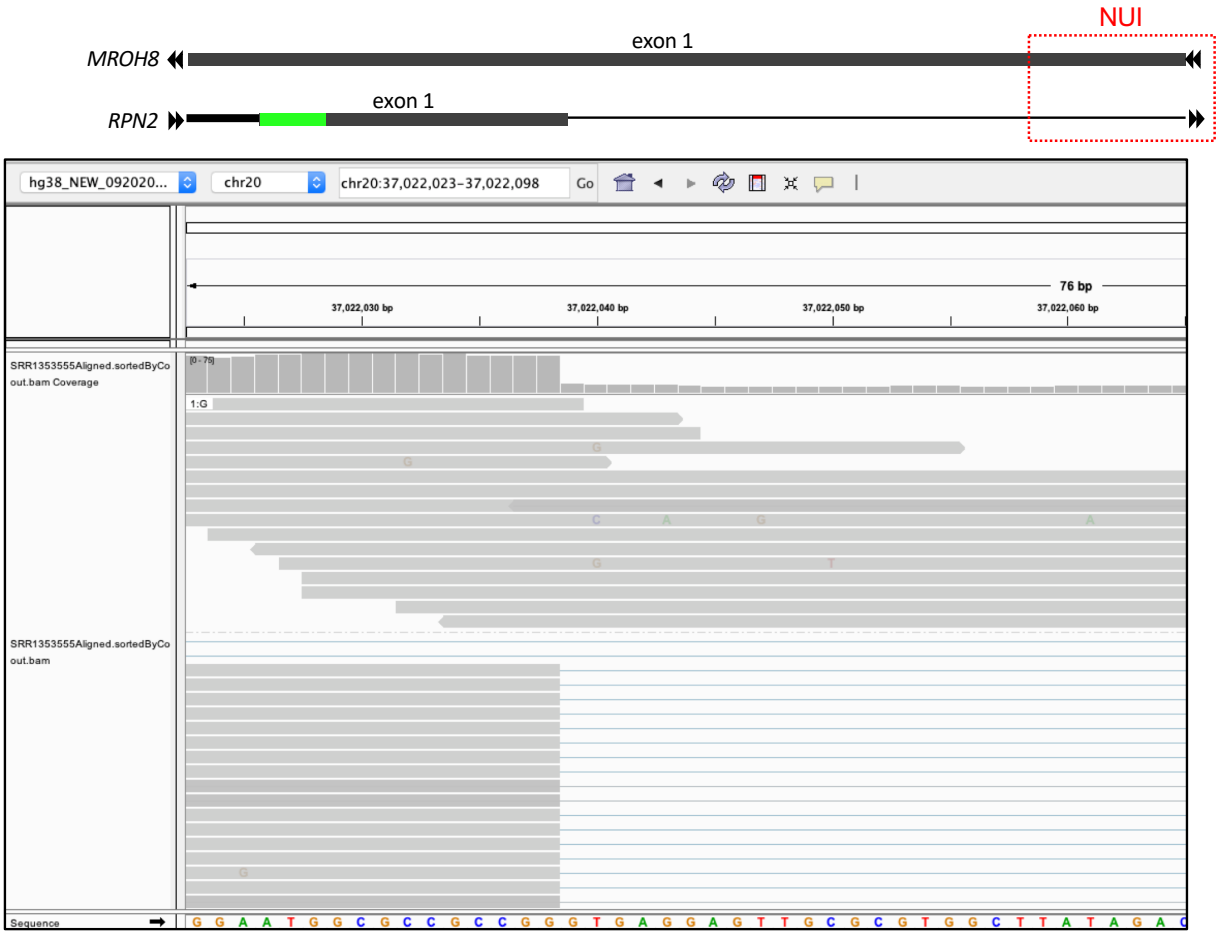

**Supplementary Figure S2.9: GTEx skeletal muscle RNA-Seq reads uncovered novel splice sites in *METTL21C* NUI.** Schematic showing three novel isoforms identified upon integrating the NUI. Isoform #1 extends the original exon 1 reading

frame by an additional 20 amino acids. An alternate start codon may be used to translate the coding gene. Isoform #2 and #3 have splice acceptors 5bp and 22bp into the integrated NUI sequence. Additional exons were also identified upstream of the original gene boundary defined by GRCh38.



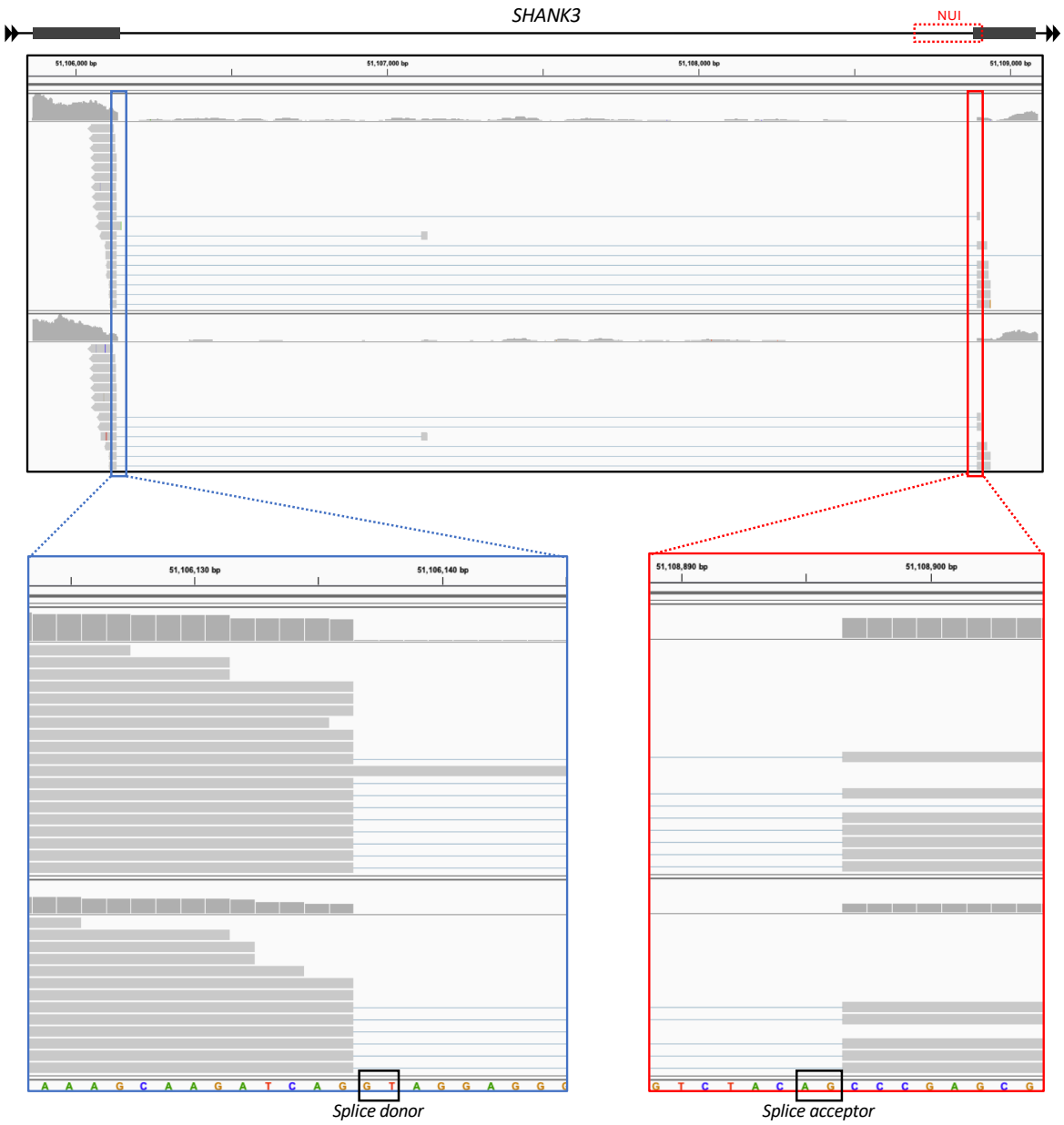

**Supplementary Figure S2.10: METTL21C NUI genotype distribution across 70 Simons Genome Diversity Project (SGDP) samples.** 10 samples were randomly selected from each representative population. Genotypes were determined using the SV genotyper Paragraph.

**GRCh38 insertion coordinates** : chr20:37179389-37179390
**Corresponding pan-genome reference coordinates**: chr20:37022058-37022088

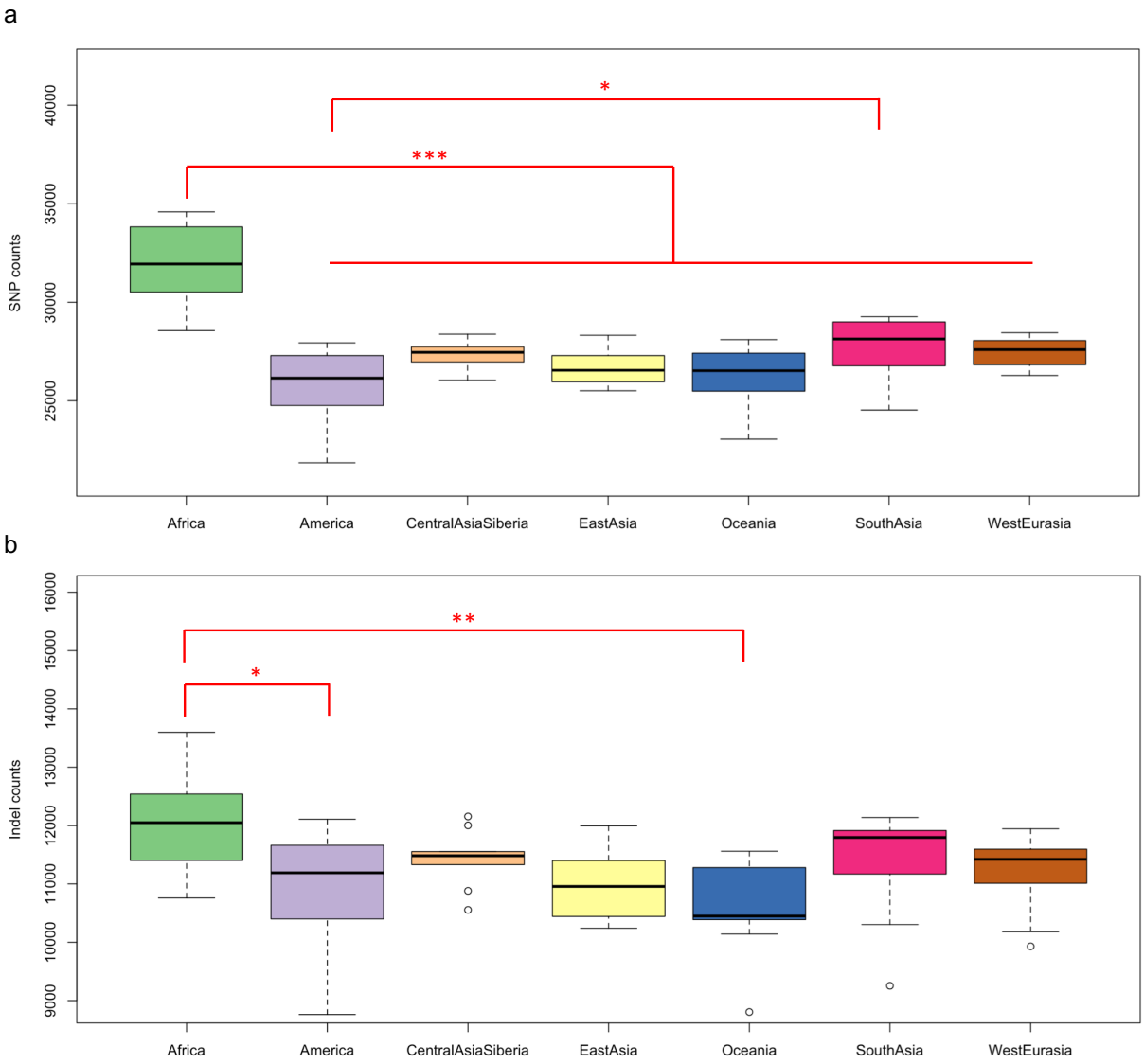



**Supplementary Figure S2.11: GTEx RNA-Seq reads aligning to the NUI in *MROH8*.** IGV screenshot showing RNA-Seq reads spanning the NUI junction corresponding to *MROH8* exon 1 (top half of the panel).

**GRCh38 insertion coordinates** : chr22:50697539-50697564
**Corresponding pan-genome reference coordinates**: chr22:51108627-51108939




**Supplementary Figure S2.12: GTEx RNA-Seq reads aligning to the NUI in *SHANK3*.** IGV screenshot showing RNA-Seq reads spanning precisely across the two exons. Reads were split right before and after the splice donor and the splice acceptor.




**Supplementary Figure S2.13: Novel polymorphic sites within NUIs.** Boxplots illustrating the (a) SNP and (b) indel counts stratified by populations defined by the Simons Genome Diversity Project study group. Statistical significance was determined by ANOVA followed by Tukey. * p <0.05; ** p<0.01, *** p<0.001.

# CHAPTER 3: FULL GENOME ANALYSIS FOR RARE GENETIC DISEASES

## 3.1 Abstract

An estimated 25 million individuals in the United States are affected by rare diseases, many of which have a monogenic basis. Current genetic testing methods provide a molecular diagnosis in only a small proportion of cases. Full Genome Analysis (FGA) is a method that combines whole-genome sequencing and long-range assembly and detects small coding and noncoding variants, breakpoints for structural variants, and long-range linkage. We built a genetic variant interpretation pipeline in-house and tested FGA's utility for molecular diagnosis of rare diseases in a clinical setting on a set of 50 undiagnosed rare disease patients. FGA expanded genetic test yield by identifying causal variants in select cases that were not resolved using short-read technologies. We found that it detected and mapped structural variants which had been missed, including non-coding duplication and translocations. As an analytical approach for the clinical setting, etiologic copy number variants were more easily identified, although deletions were detected at the same efficiency by short and long read sequencing. FGA provided additional phasing and allele-specific mutation information that was missing from short-read technologies, as illustrated by a diagnostic case where two deleterious variants are in *trans* and located 180 kb apart. Genome assemblies provide comprehensive information that can replace multiple tests for rare monogenic disorders and expands the range of variants that can be detected. The results of this study suggest that the high-quality genomic data produced will facilitate the diagnostic process as the set of disease-associated variants expands.

## 3.2 Introduction

Current approaches to diagnosis of presumptively genetic conditions include microarray analysis and short-read sequencing of exomes or genomes[57-59]. Although the diagnostic yield from these methods is generally good, they leave many cases unresolved[57,60]. The yield can be augmented by re-analysis against recently discovered disease-causing variants[61], or by using family-based analysis to identify *de novo* variants[59], but improvements are modest.

Two factors are principally responsible for diagnostic failures using short-read sequencing. First, short-read sequencing does not give a complete picture of the genome. For example, exome sequencing does not detect the majority of structural variants, cannot create chromosomal maps, rarely detects repeats, misses variants in exons that are not captured efficiently, and misses all non-exonic variants[23,28,29,32,62]. Even whole genome sequencing, which provides up to 9% additional diagnostic yield compared to exome sequencing[60], cannot detect all structural variants (especially duplications, inversions, and translocations), create chromosomal maps, or provide phasing information. In addition, detection of structural variants by short-read whole genome sequencing requires additional analytical processes that are not practical for a clinical setting[63]. Second, genetic diagnosis of rare disorders often entails “experiments of one,” where many sequence variants found in the proband must be vetted against current knowledge to decide if any of the variants meet the diagnostic criteria. Our incomplete biological knowledge limits our ability to identify the “causal variant” for any particular patient. Until we understand the functional consequences of all variants, or

more patients with the same phenotype are found for a variant/gene, many candidate variants remain variants of uncertain significance.

Newly developed long DNA sequencing and mapping technologies can help solve the problem of incomplete genome analysis. Whole genome sequencing methods that produce long contigs (sets of adjacent DNA segments that together represent a consensus region of DNA) promise several advantages over short-read sequencing alone[23,28,32,62,64]. First, long-read methods can resolve more easily large SVs(including large deletions, large insertions, translocations, and inversions), eliminating the need for additional genetic tests[12,28]. Second, long-read sequencing detects more readily insertions/deletions (indels) of intermediate size (500 bp to 50 kb)[62], ones that escape detection by clinical microarray testing because they are too small, and by short-read sequencing because they are too large. Third, the methodology of genome reconstruction provides opportunities to detect rearrangement variants that have evaded detection[62]. Fourth, single-basepair resolution of rearrangement breakpoints allows determination of the precise location of each insertion or duplication, making it possible to see if the structural variants disrupt genes or other sequences of functional significance. Finally, long contigs provide phasing information to determine the haplotype on which a variant occurs (e.g. maternal or paternal chromosomes) and can help resolve inheritance patterns of disease-causing alleles[64].

Here we test the diagnostic capabilities of an approach that we call “Full Genome Analysis” (FGA), which combines linked-read sequencing technology and optical mapping to produce contigs with a median length of ~100Mb. To analyze the data in an unbiased and comprehensive fashion, an automated genetic variant interpretation

pipeline was built to select and prioritize variants based on the individualized medical history. We show that FGA, when applied to patients with rare disorders in a clinical setting, leads to new diagnoses in cases where whole-exome analysis did not return a diagnosis. We further show that FGA is capable of detecting a variety of genomic variants, including translocations, intermediate-sized copy number variants, and phased biallelic variants, that are responsible for disease but usually missed by analysis of short-read sequencing data alone. Based on its ease of diagnosis, FGA opens the prospect of resolving more complex parts of the genome and identifying a more comprehensive set of genetic variants in rare disorder diagnosis[23,28,29].

### 3.3 Materials and Methods

*DNA extraction and preparation*

High molecular-weight DNA was extracted and isolated using the Bionano Prep Blood Isolation Kit following the manufacturer protocol (Bionano Genomics). Bionano optical mapping libraries were prepared using the Direct Label and Staining Technologies (DLS) following the manufacturer protocol (Bionano Genomics). 10x Genomics linked-read sequencing libraries were built as published[6] using the GemCode platform (10x Genomics).

*Optical mapping data generation and processing*

Optical mapping on the Bionano Irys and Saphyr platforms was used to produce *de novo* assemblies and identify structural variants and rearrangements. DNA was labeled using Nick, Label, Repair and Stain (NLRS) and/or Direct Label and Staining

Technologies (DLS). The first uses a nicking endonuclease that recognizes a specific 6-7 basepair sequence and creates a single-strand nick, filled with fluorescent nucleotides. The second uses a single direct-labeling enzymatic reaction to attach a fluorophore to a specific 6-basepair DNA sequence motif. Labeled DNA libraries were loaded onto the Bionano Genomics Irys™ Chip or Saphyr™ Chip, linearized and visualized using the Irys™ or Saphyr™ system, which detects the fluorescent labels along each molecule. Single molecule maps were assembled *de novo* into genome maps using Bionano Solve with the default settings[29]. Genome map assemblies were aligned to reference (hg38). Genome assembly, alignment, and variant calling were performed using IrysSolve. SVs were filtered based on a set of diverse controls labelled with the same enzyme using the 50% reciprocal rule.

*Linked-read data generation and processing*

Sequencing data was obtained from 10x Genomics linked-read libraries sequenced to ~60X coverage using an Illumina sequencer. Reads were aligned to GRCh38 using LongRanger and variants were identified using the callers integrated in the 10xG pipeline (GATK Haplotype caller for SNPs and indels).

*Automated clinical interpretation and prioritization of sequencing and mapping data*

An automated clinical interpretation and prioritization pipeline was built in-house using custom and publicly available software (**Supplementary Figure S3.1**; see end of chapter). Electronic health records of the probands were exported into JSON format for parsing with clinical natural language processing algorithm ClinPhen[65]. HPO

hierarchical terms separated by 1 degree were also included as part of the clinical phenome. Every HPO term (*h)* is assigned a weight, which is defined as the inverse of the total number of disease genes associated with it.

$$weight\ of\ h = \frac{1}{total\ number\ of\ genes\ associated\ with\ h}$$

Next, we overlapped the clinical phenome of the proband with a list of known phenotypic features associated with mutations in a given gene (*G*). The overlapping terms were used to calculate a gene sum score to identify and rank clinically relevant genes.

$$Gene\ sum\ score\ of\ gene\ G = \sum_{1}^{n} weight(h_n)$$

SNP and indel exonic variants identified from the proband were overlapped with the ranked gene list. The same strategy was applied to all SVs overlapping genic exons. Scores were normalized for comparing and calculating confidence scores. All SVs were vetted against a set of regions known to be associated with deletion and duplication syndromes. Additionally, all translocations and inversions were reported by default. Prioritized variants were reviewed manually to determine which one was diagnostic.

*Approvals and Phenotypic Assessment*

The study was approved by the Institutional review board of Children's Hospital Oakland and University of California, San Francisco (UCSF), Committee for Human Subjects Research. Recruitment was from UCSF Benioff Children's Hospital Medical Genetics and Genomics clinics. We focused on cases of two types, chosen to demonstrate the capability of FGA: cases in which whole-exome sequencing had not returned a causal

variant; sporadic cases from the pediatric population that are suspected to have a genetic basis, but fall into no clear syndrome and have no clear candidate target for conventional genetic diagnosis. Individuals with undiagnosed conditions and unaffected parents were offered testing and underwent a consent process prior to blood draw. Phenotypic evaluation was performed by clinical review by at least two genetics professionals, and human phenotype ontology terms were curated for each case.

### 3.4 Results

#### Automated genetic variant interpretation pipeline

An automated pathogenicity assessment pipeline was built in-house to filter and prioritize variants based on the reported clinical features of recruited patients. In addition to analyzing single nucleotide variants (SNVs) and small indels, this pipeline was designed specifically to handle structural variations that often evade detection by short-read technologies. Our SV modules integrate the two complementary (linked-read sequencing and optical mapping) datasets and are capable of reporting clinically relevant translocations, inversions, deletions, duplications, insertions and other types of complex SVs of virtually all sizes. This pipeline ranks variants for every case and considers each potential inheritance pattern (details in methods and supplementary methods).

#### Full Genome Analysis

Using the automated genetic variant interpretation pipeline, we performed full genome analysis (FGA) on 50 undiagnosed cases to determine the diagnostic yield and ask if it

could solve cases that are currently difficult to solve with short-read sequencing technologies. The test was carried out in a clinical environment, allowing clinicians to propose unsolved cases for further genomic sequencing; cases were included only if prior testing was negative and characteristics of the case did not suggest any further specific test. Of the 50 cases, 20 had negative prior trio whole exome sequencing and 42 prior negative microarray. Overall, our automated interpretation pipeline helped diagnose 18 cases (14 SNPs/indels and 4 SVs). All diagnostic cases but three were ranked as the top variants in their respective inheritance/SV groups (**Table 3.1**). Additionally, more than half of the remaining undiagnosed cases have at least 1 prime SV candidate for immediate follow-up.

| Case | Variant type | Diagnostic variant rankings |
|---|---|---|
| 0103 | Indel | 1 out of 4 in de novo |
| 0203 | SNP | 1 out of 10 in de novo |
| 0503 | SNP | 1 out of 4 in de novo |
| 0703 | SV - translocation | Not ranked[#] |
| 1003 | SNP | Not ranked[#] |
| 1503 | Indel | 2 out of 8 in recessive |
| 1703 | SV - duplication | 1 out of 1 in duplication |
| 1903 | SNP | 1 out of 5 in recessive |
| 2103 | SNP | 1 out of 4 in recessive |
| 2403 | SNP/Indel | 1 out of 8 in compound het |
| 2503 | SNP | 1 out of 23 in de novo |
| 2704 | SNP/SNP | 1 out of 3 in compound het |
| 3103 | SNP | 1 out of 12 in de novo |
| 3603 | SNP | 1 out of 1 in recessive |
| 4103 | Indel | 3 out of 15 in de novo |
| 4203 | SV - deletion | 1 out of 1 in deletion |
| 4903 | SNP | 12 out of 92 in inherited het |
| 5103 | SV - deletion | 1 out of 1 in deletion |

**Table 3.1: Automated variant interpretation pipeline performance.** If a patient has more than one genetic diagnosis, only the higher-ranking variant is reported. The pound sign (#) indicates that the variant was reported but not ranked by the pipeline due to lag in annotation.

FGA solved three classes of cases where short-read sequencing or microarray analyses had previously failed to detect the causal variants: 1) cryptic heterozygous

structural variants (e.g. *NHEJ1-IHH*, *WAC*), particularly variants of intermediate size; 2) translocations; and 3) missed heterozygous variants, for example in *trans* for recessive disorders (e.g. *TSPEAR*). Here we describe examples of diagnostic findings.

**Non-Coding Structural Variation**

In a 9-month-old female with craniosynostosis and syndactyly, we found a rare 32 kb heterozygous *de novo* intronic duplication within the *NHEJ1* gene. FGA identified the breakpoints of a duplication at chr2: 219,102,933 - 219,134,970 (genome version GRCh38) in the 2q35 band (**Figure 3.1**). Similar duplications have been described in cases of craniosynostosis and syndactyly, where variant localization was accomplished by familial mapping studies[66,67] (Chromosome 2q35 Duplication Syndrome, OMIM #185900); the breakpoints identified here narrow the critical region of the *NHEJ1* intron that is important for the condition[66,67]. FGA also found that the intronic duplication occurred on the paternal allele of our case. This heterozygous *de novo* duplication was detected by both optical mapping (Bionano) and linked-read sequencing (10x Genomics) technologies. FGA also indicated that the duplication occurred adjacent to the original segment in tandem (**Figure 3.1**). This mid-sized structural variant (~32kb) cannot be detected by standard microarray analysis because it is small and intronic. It also escaped detection by our short-read whole genome sequencing copy number variant calling but was easily identified by FGA (**Table 3.2**).

The duplication affects an enhancer for the Indian Hedgehog (*IHH*) gene, located upstream in the third intron of the neighboring *NHEJ1*[66]. ENCODE data support the enhancer function of this intronic region (**Supplementary Figure S3.2**; see end of

chapter). The structural variant breakpoints defined by FGA narrow the critical intronic region for this condition.

| | Short-read WGS CNV | Linked-reads | De novo assembly |
|---|---|---|---|
| **Variant calls:** | | | |
| Total number | 1945 | 264 | 225 |
| High quality | 1415 | 2 | n/a |
| Mean size ± SE (bp) | 54,409 ± 21,115 | 78,861 ± 3,124 | 97,654 ± 22,204 |
| **Diagnostic variant:** | | | |
| Identified | no | yes | yes |
| Correct SV type | - | yes | yes |
| Correct zygosity | - | yes | no |

**Table 3.2. Comparison of duplication calls between short-read WGS CNV and genome assembly technologies.** Table includes calls for the 32 kb intronic *NHEJ1* duplication case.

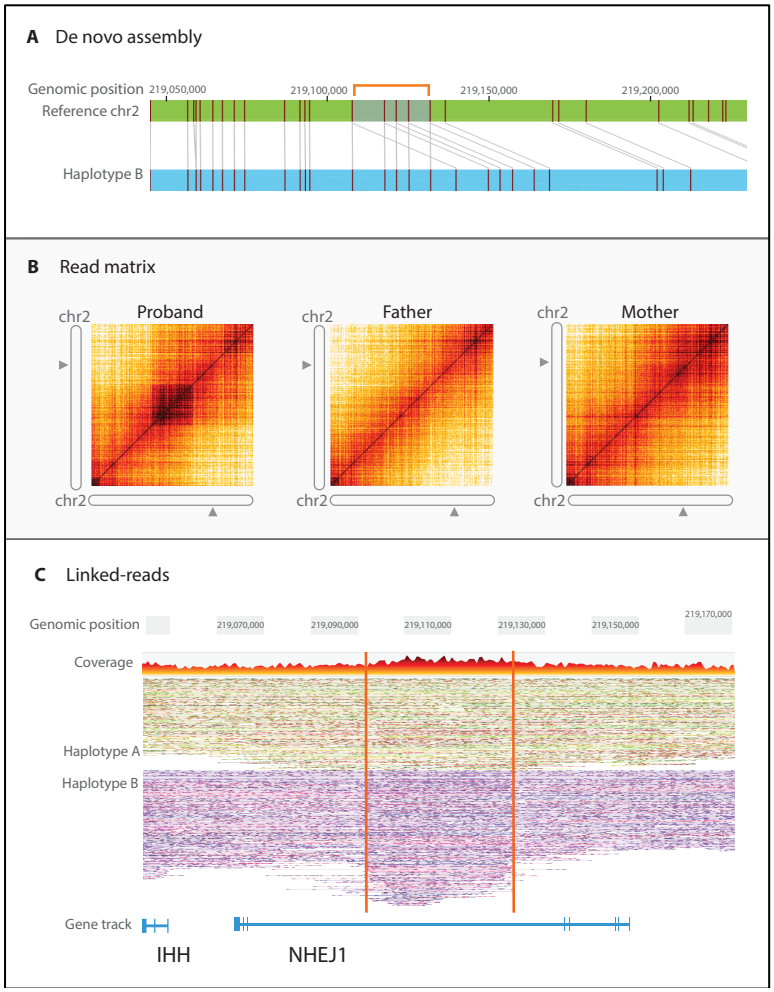



**Figure 3.1: Heterozygous, intronic tandem duplication (32kb) in NHEJ1.** This region (chr2: 219,102,933 - 219,134,970, 2q35, GRCh38) corresponds to an *IHH*

upstream enhancer. (A) depicts a de novo assembly (light blue) and its alignment to reference (green). The labeled motifs in the reference genome (vertical maroon lines) are duplicated in Haplotype B and their orientation demonstrates the duplication occurred adjacent to the original sequence or in tandem. (B) Shows a matrix view or heatmap of linked reads. The dark orange square in the far-left panel (proband's panel), illustrates a higher density of barcode overlap in the read matrix compared to parents, indicating the variant occurred for the first time in the proband or de novo. (C) contains phased haplotypes generated by linked-read data. Haplotype B, in purple, contains a region with higher number of linked-reads due to sequence duplication.

**Genomic Rearrangements**

Current clinical sequencing pipelines have difficulty detecting translocations without specific additional informatic analyses. In contrast, we were readily able to detect translocations using FGA and our standard informatic pipeline. For example, we found a germline translocation between chromosomes 1 and 9 in a 2-year-old male with a history of neuroblastoma and developmental delay who had negative microarray and exome sequencing (**Figure 3.2** – both linked-read genome sequencing and optical mapping support the translocations). Trio analysis indicated that the translocation occurred *de novo*; it was subsequently verified by cytogenetic chromosome analysis (**Supplementary Figure S3.3**; karyotype: 46,XY,t(1;9)(p32.3;p21); see end of chapter). FGA identified the precise breakpoint locations on chromosome 1 and chromosome 9 (chr1: 49,553,194 and chr9: 29,096,674, respectively, genome version GRCh38). The breakpoints were found to be non-exonic, occurring in an intronic region of *AGBL4* and an upstream/untranslated region near *LINGO2*. FGA also revealed that the translocations occurred on the paternal allele, with a small breakpoint deletion suggesting non-homologous end joining, with an additional maternally inherited intronic deletion present on the other allele. *AGBL4*, encoding a cytosolic carboxypeptidase, has a potential role in neuroblastoma and developmental delay. Copy number alteration has

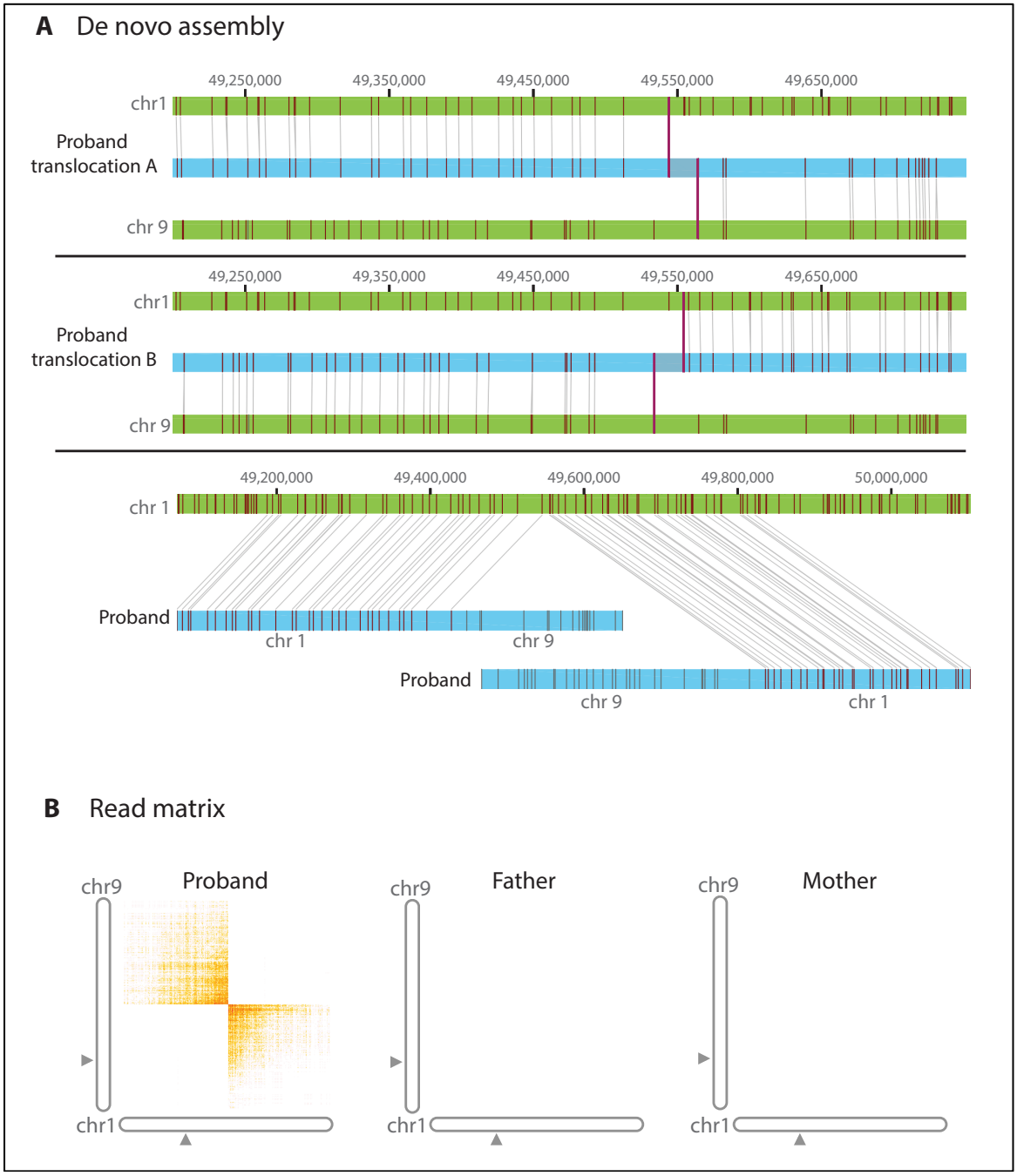



**Figure 3.2. Structural rearrangement detection with de novo assembly and linked reads; t(1:9)(p33,p21).** (A) *De novo* assemblies of chromosome 9 and 1. Gene positions in grey at the top, reference genome in green (reference GRCh38), proband assembly map blue, vertical maroon line shows matching label patterns. The first and second panel show two *de novo* assembly maps that align to reference chromosomes 1 and 9 and the translocation breakpoint where the alignment switches. The third panel depicts two assembly maps in chromosome 1 with segments that align and do not align to reference due to the translocation. (B) Matrix view of linked reads showing unexpected barcode overlap (in orange) between chromosome 1 and 9, corresponding to the point of fusion between the two. This overlap is absent in the parents.

also been reported in *LINGO2* in neuroblastoma cell lines[68,69]. Although several other genes in this case had *de novo* variants (*MYH11, GABRA2, TFE3*), none of these were clearly etiologic. Neuroblastoma predisposition genes *ALK* and *PHOX2B* were also negative[70]. The *de novo* translocation suggested a new etiology for this condition, which could be explored in future studies[71,72]. Interestingly, the short read WGS copy number calling shows hundreds of potential inter-chromosomal break points that needed further analysis for diagnostic use. In contrast, FGA had at least 8-fold fewer candidates (**Supplementary Table S3.1**; see end of chapter). Furthermore, *de novo* assembly from FGA promptly identified the event as a translocation as opposed to short-read WGS CNV which called many breakpoint junctions**.**

Translocations have important implications for future reproductive risks. A diagnostic strategy that encompasses what chromosome analysis and microarray do in a single diagnostic test could also serve to detect balanced events which are important for family planning in carriers. More complex rearrangements were also promptly detected among significantly fewer candidates and localized with FGA (unbalanced insertional translocation) providing novel variants for future characterization (**Supplementary Figure S3.4, Supplementary Table S3.2**; see end of chapter).

**Deletions**

FGA was also capable of pinpointing genomic breakpoints of deletion type structural variants. For example, clinically significant copy number variants identified by FGA include a 36 kb deletion disrupting *TANGO2* (OMIM #616878) in cousins with a history of episodic rhabdomyolysis, metabolic acidosis, ketosis and liver failure (**Figure 3.3**),

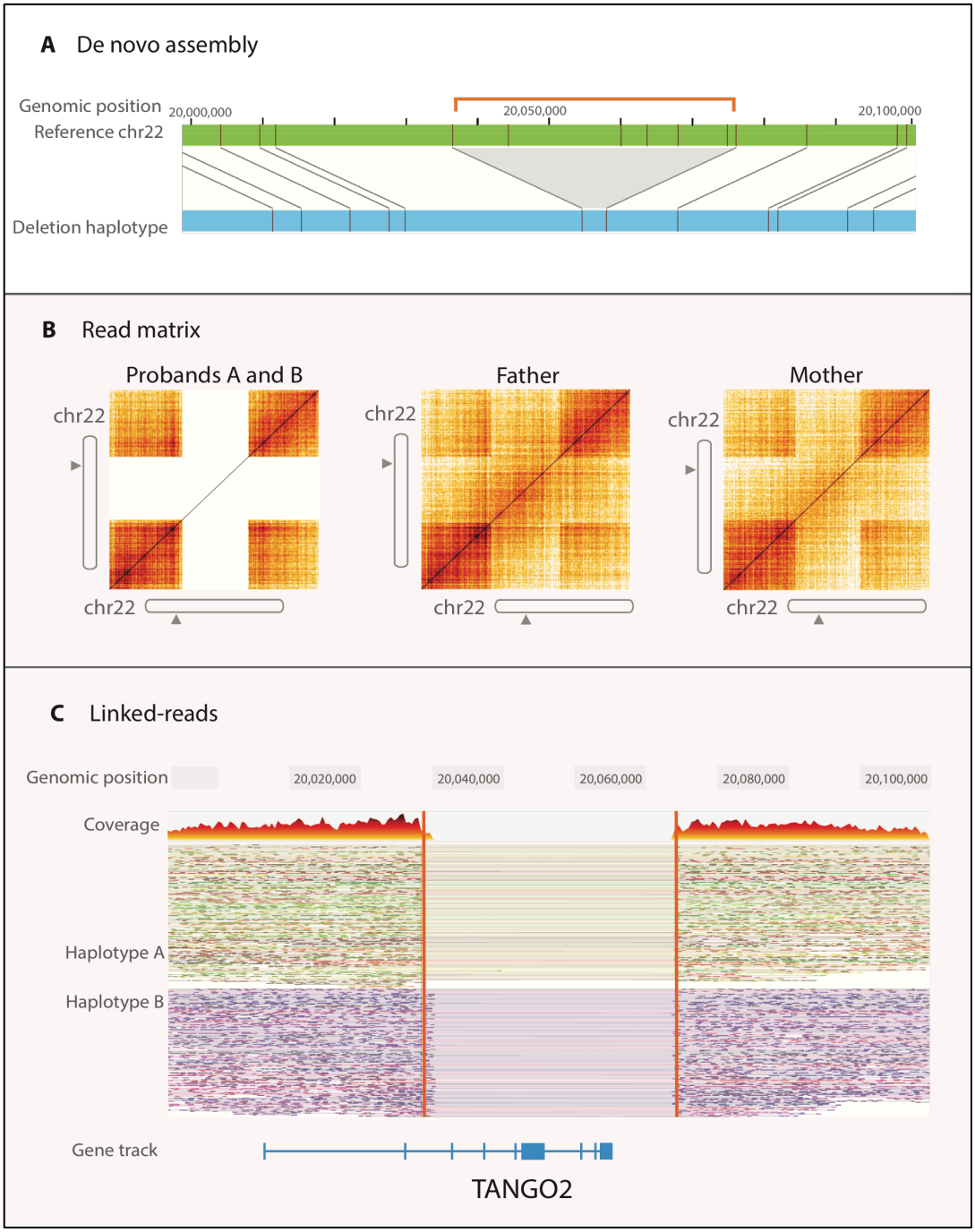



**Figure 3.3. Homozygous 36kb deletion disrupting TANGO2 (chr22: 20,039,637 – 20,075,714 and chr22: 20,041,469 – 20,075,432, genome version GRCh38).** (A) De novo assembly (light blue) demonstrates missing sequence labels with respect to reference (green). The orange bracket and gray triangle shows the deleted region. (B) Matrix view with absent signal from intervening region demonstrates probands have a homozygous deletion. (C) Deletion is also seen by drop in coverage in both haplotypes in linked-read data.

and a 1480 bp *de novo* deletion in *WAC* (Desanto-Shinawi syndrome, OMIM #616708) found in a male patient with seizures, hypotonia, developmental delay and nonfamilial features like low-set ears and brachycephaly (**Supplementary Figure S3.5**; see end of chapter). FGA simultaneously identified and verified such variants and breakpoints without additional copy number prediction tools or external validation required by current short-read sequencing pipelines. Indeed, short-read sequencing copy number calling was able to detect these deletions albeit with sometimes incorrect zygosity (**Supplementary Tables S3.3 and S3.4**; see end of chapter). FGA had an advantage compared to short-read sequencing in identifying deletions in challenging regions of the genome near segmental duplications.

**Small Variant Detection and Biallelic Phased Variants**

FGA also yielded coding variants, similar to short-read WGS, however phasing was now uniquely possible given the longer DNA segments. Discerning that variants reside on separate chromosomes is critical for diagnoses of compound heterozygous recessive variants; FGA is capable of making this determination in a single proband test. We identified two *TSPEAR* variants in a young female with oligo/hypodontia, missing 15 adult teeth, but no previous family history. The two *TSPEAR* NM_144991 variants were found 180 kb away from each other. FGA phasing clearly showed that the variants occurred in *trans*, suggesting that both parents are heterozygous carriers. The maternal variant is a 10 base insertion which leads to frameshift, c.51_52_insGCCGGGGGCC, p.Gly17ProArgTer, while the paternal variant is nonsense, c.1281G>A, p.Trp427*; together the variants imply a bi-allelic etiology (**Figure 3.4**). The large phasing block

(chr21: 29,801,272 - 44,927,448, genome version GRCh38) created by FGA was able to discern the two haplotypes and determine that the variants are in *trans,* even with the affected individual's sequence only; this observation explains the pattern of inheritance from unaffected parents. Correspondingly, we confirmed the maternal/paternal origin of each of the variants by analysis of the parents, verifying carrier status of the parents as well. *TSPEAR* has recently been associated with tooth agenesis, and loss-of-function variants in *TSPEAR* are associated with ectodermal dysplasia 14, hair/tooth type, with or without hypohidrosis[68,73] (OMIM #618180). Indeed, WGS may have identified these variants, but only phasing using FGA is sufficient to make a diagnosis on proband alone. In detecting compound heterozygous variant, phasing information is valuable since one of the variants might be *de novo* and data from both parents are not always available.

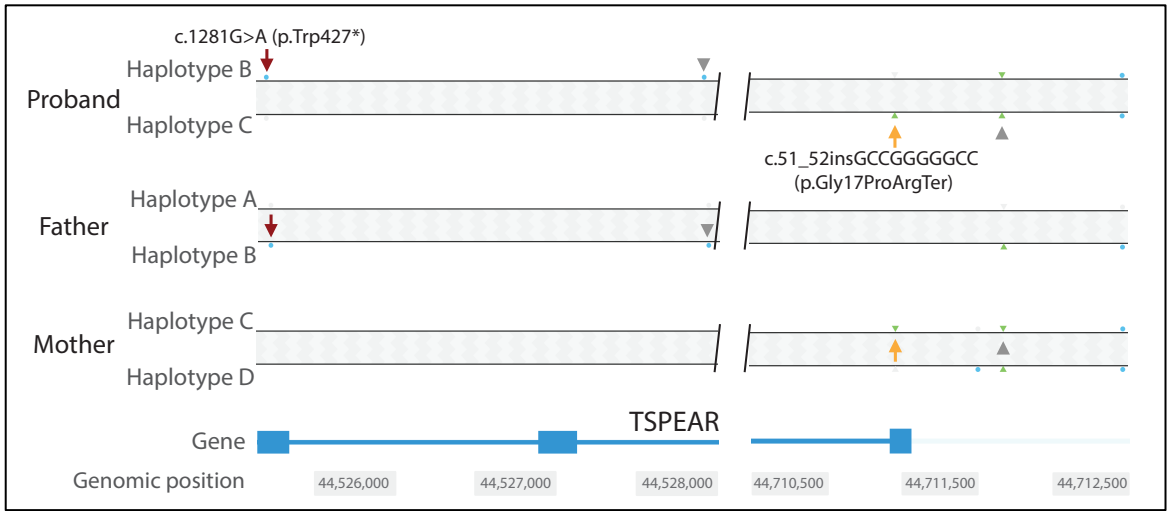



**Figure 3.4: Variant haplotype distinction.** Sequencing results reveal compound heterozygous *TSPEAR* variants (NM_144991). Phasing was possible for etiologic variants 184,756 bp apart inside a unique phasing block of 15.1 Mbp. Maroon and yellow arrows point to each variant. Grey arrowheads point to SNPs that confirm *trans* orientation.

By detecting additional variation beyond current clinical testing modalities, FGA yields novel genomic information for discovery in undiagnosed patients. The total diagnostic yield was 36% of the 50 cases tested, with FGA detecting both new structural

variants and single-nucleotide variants that were missed by previous short read annotation. FGA diagnosed 25% of exome negative cases (5 of 20 cases). We also identified candidate variants in more than half of the cases for future follow-up. Thus, our clinical pilot shows that FGA detects and localizes duplications that are missed by WGS and can easily identify translocations and phase variants across long-distances. With variant detection from longer DNA technologies, we can improve detection of diagnostic variants and provide higher resolution genome maps for future studies.

### 3.5 Discussion

In genomic medicine, rare disease diagnostics has traditionally been imprecise and biased largely due to the rarity of the disorders. Here, we described the FGA approach with automated analysis using linked-read sequencing and optical mapping to evaluate the full spectrum of genetic variants implicated in rare genetic diseases. The automated interpretation pipeline integrates newer technologies into the diagnostic realm by enabling a streamlined variant detection protocol and minimizing biases introduced during the process. This data-driven approach results in a drastic decrease in human intervention time and ensures that every patient was evaluated thoroughly. We find that genome assemblies can be used in clinical testing strategies detecting all types of genetic variants concurrently. For individuals with undiagnosed conditions, these technologies encompass what is currently provided by the combination of chromosome analysis (karyotyping), microarray study, and clinical short-read WGS. By identifying novel SVs and phasing, it provides diagnostic information beyond what current clinical tests do. The use of a combination of methods can also provide internal validation of

SVs, bypassing the need of additional time and blood drawn for additional testing. By constructing *de novo* genome assemblies and facilitating the identification of variants that do not map to the genome reference, these technologies can also provide additional information for future analysis. These strengths make the technologies highly suitable for early implementation in diagnostic evaluations, particularly if a specific genetic condition or type of variant is not immediately suspected.

FGA has distinct practical advantages over traditional genetic testing strategies because it can detect phased variants for recessive conditions, as well as the full spectrum of structural variants including balanced translocations. FGA makes it possible to effectively test probands even when parents/family members are not available for testing. This situation occurs often in intensive care units or in other settings where rapid diagnosis is vital to clinical care. Phasing allows variants to be identified along maternally or paternally inherited chromosomes, or for two potential disease variants to be evaluated for *cis* or *trans* configuration; this information can be critical in an evaluation of clinical significance. FGA also returns a high-quality genome reconstruction in a single test; it detects all types of structural variants without the need to perform additional tests (chromosome analysis, microarray and sequencing) and without the limitations in variant size that restrict the utility of those tests. Finally, FGA resolves complex or novel regions of the genome much more readily than with short-read technologies, enabling resolution of regions not resolved by short-read technologies[12,28].

The application of FGA in genomic medicine is not without limitations. First, our automated variant interpretation pipeline is based on existing annotation databases.

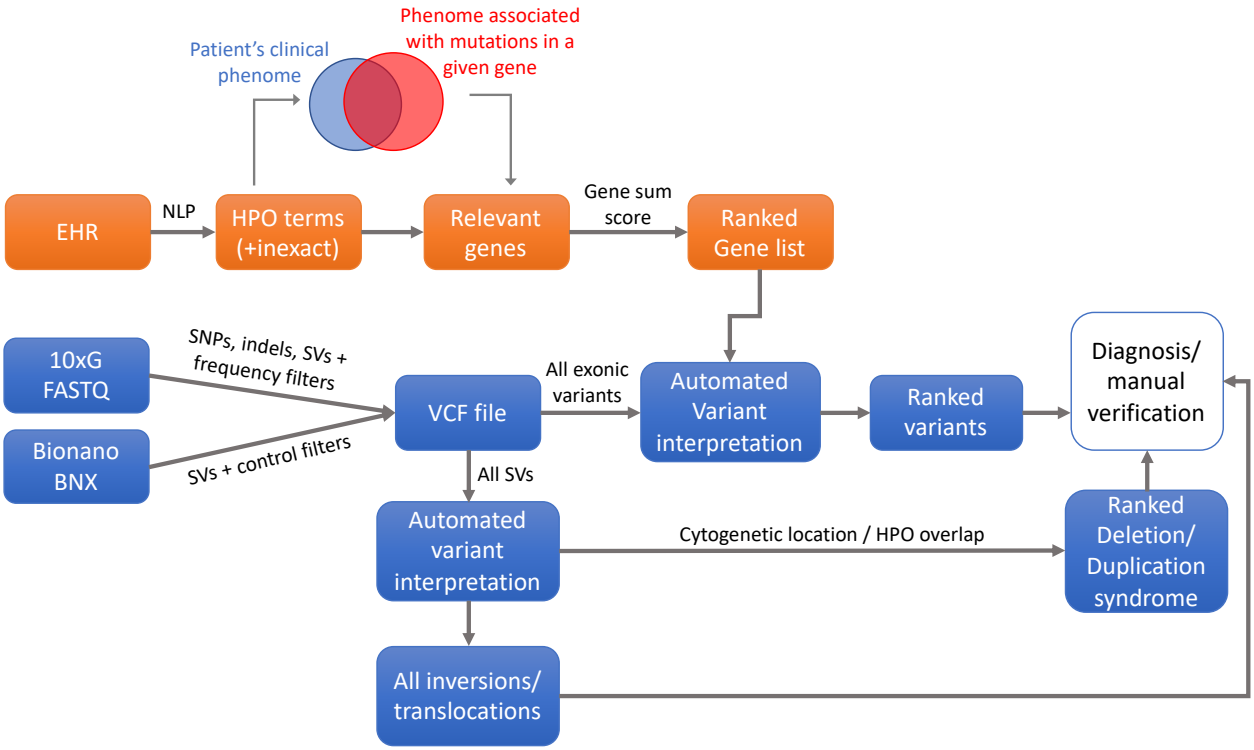
Genetic variations cannot be ranked or annotated if they are not found in these resources. Second, even with the use of long molecules averaged at 200-300kb in our optical mapping experiments, they are not long enough to resolve the large, near-identical segmental duplications in some of the most complex regions of the human genome. A small number of these complex regions remain inaccessible despite using long-range sequencing and mapping technologies[24]. Third, the current human reference genome is a set of composite haplotypes generated from 8 anonymous DNA donors. As such, there are functionally important sequences found in many people around the world but are missing from the reference genome[27,74]. Since the reference genome serves as the benchmark for all analyses, missing sequences are never assessed, thus making variants in these regions undiagnosable.

Although the number of diagnostic cases attributable to SVs is not striking in the current analysis, we did identify at least one highly probable SV candidates on more than half of the undiagnosed patients. These cases do not meet the ACMG diagnostic criteria due to several reasons. SVs overlapping similar regions do not always produce the same phenotype. This is particularly limiting since most SVs are not recurrent and thus do not share identical breakpoints. Furthermore, unless a critical region can be established or the syndrome is associated with very distinctive phenotypes, it is unclear whether an SV can be diagnostic even if it is *de novo* and involves millions of base pairs. Most importantly, SV and CNV databases are strikingly sparse and inconsistent in contrast to SNV databases. This substantially dampens our confidence in determining which SVs are diagnostic.

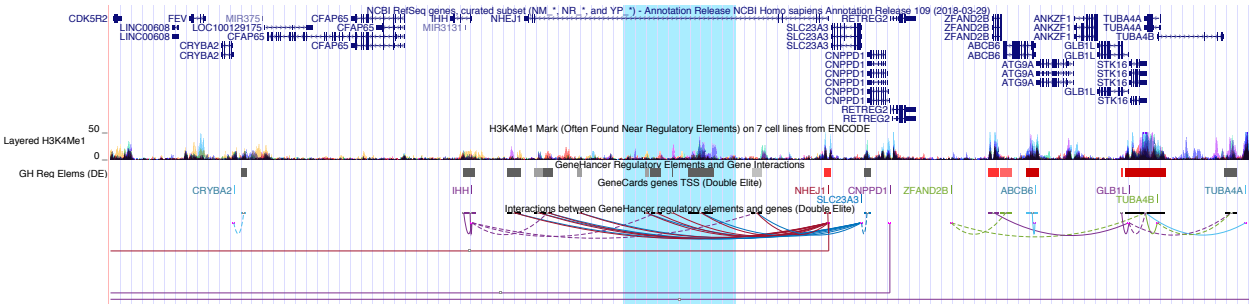
We can expect that whole-genome sequencing will become the method of choice for genetic diagnosis, given the greater number of variants that it is capable of detecting relative to exome sequencing or microarray analysis. In choosing a technology for the acquisition of whole-genome data, one should consider not only costs and the complexity of the analytical workflow, but also the completeness of the data and the continuing value of the data for future reanalysis. The inherent amount of missing data in genomes generated by short-read sequencing reduces their utility for clinical diagnosis. Data reanalysis is becoming a successful strategy to identify variants that cause disease in a patient's genome; as our understanding of deleterious variants grow, it is possible to revisit previously acquired data and assign functional significance to previously discarded variants. FGA's ability to acquire a more extensive set of variants increases the likelihood that future reanalysis will be productive. More importantly, by identifying previously unknown variants, FGA makes it possible to explore their functional significance.

The increase in diagnostic yield produced by FGA in this study, attributable to advances such as phasing and structural variant detection, has made it possible to solve cases that were negative by short-read sequencing. Full realization of FGA's potential to provide comprehensive detection of clinical variants will require a combination of automated capture of phenotypic terms with expanded expertise in variant interpretation. Comprehensive assessment of the genome in every undiagnosed patient would rapidly produce both genome maps of annotated functional variants and new diagnostic information. The result would be a better understanding of population variation, and improved diagnostics for direct clinical care.

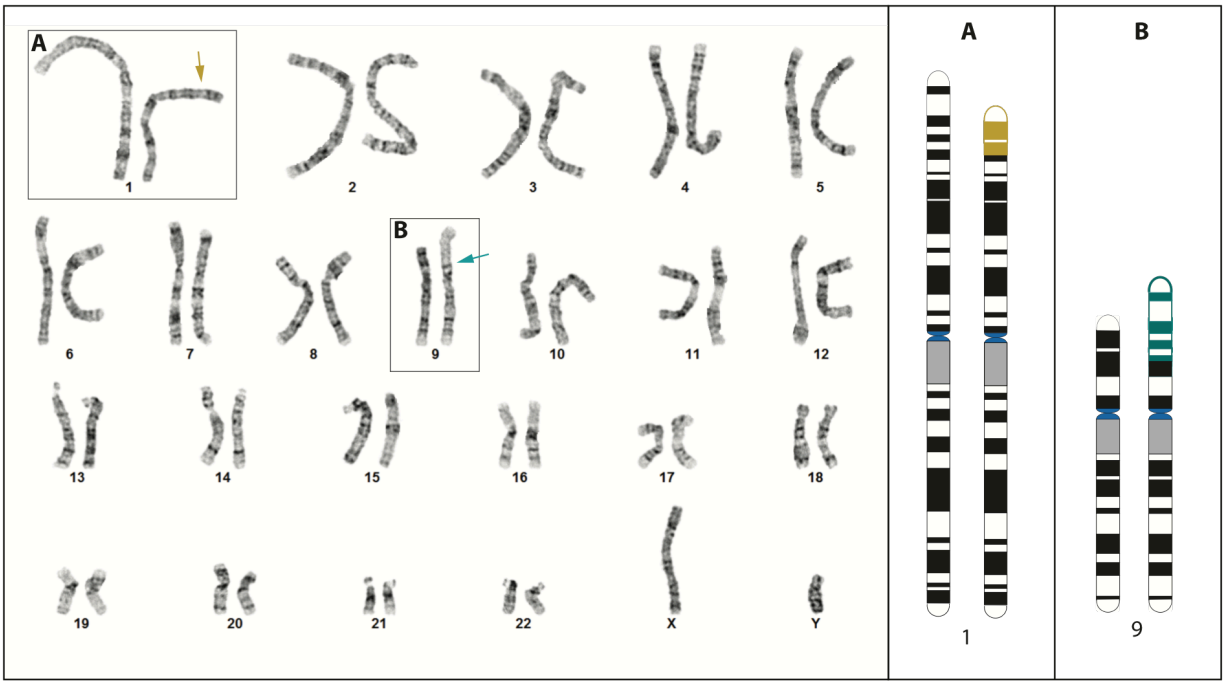
## Supplementary Figures



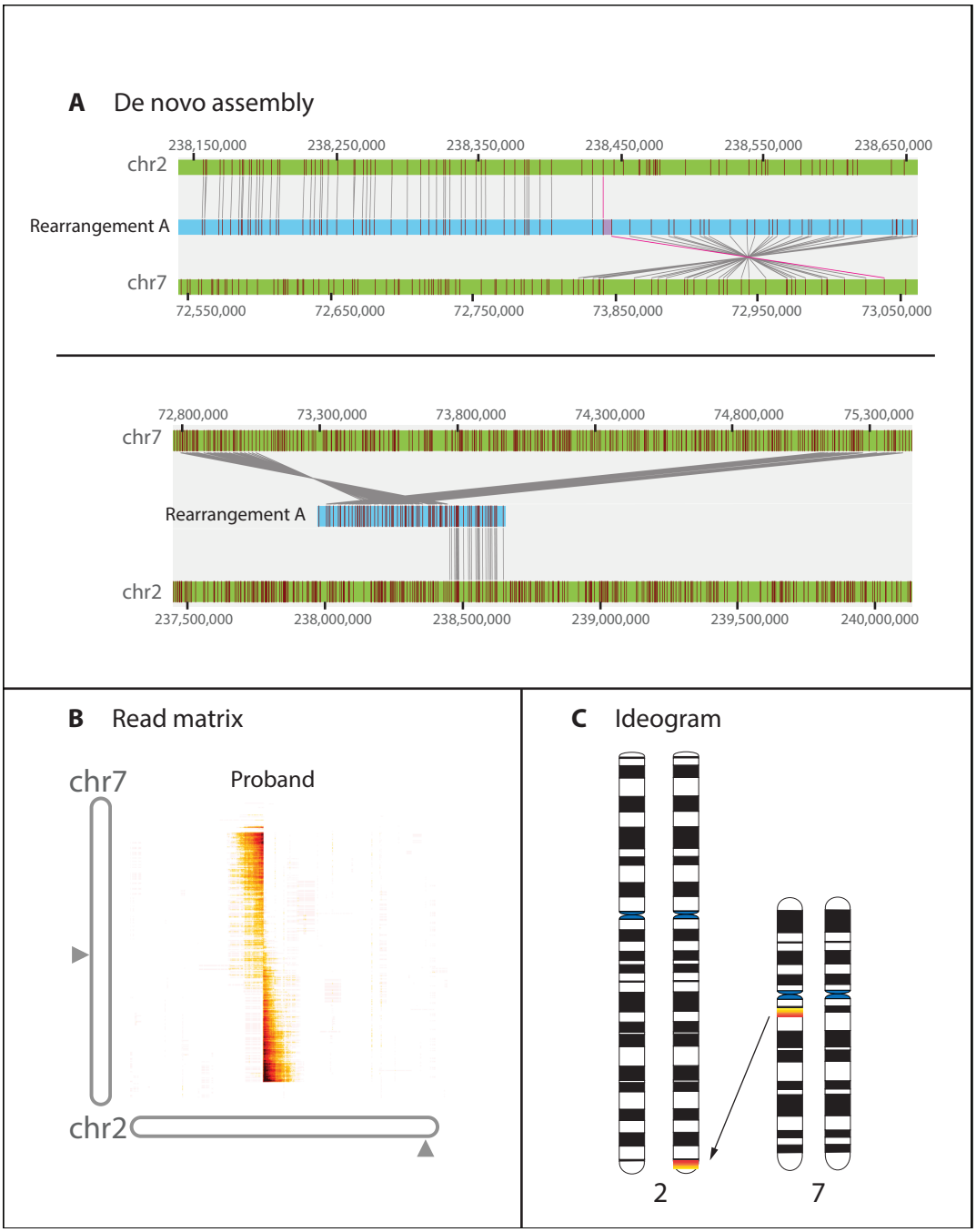

**Supplementary Figure S3.1: Automated computational pipeline for clinical variant interpretation and prioritization.** In orange, the electronic health records from the probands were exported as JSON format. Unstructured clinical notes were translated into HPO terms using clinical natural language processing tool. Related HPO hierarchical terms separated by 1 degree were also included as part of the clinical phenome, which was then overlapped with a list of known phenotypical features associated with mutations in a given gene. The overlapping HPO terms were used to calculate a gene sum score to rank potentially relevant genes. Such score was derived using the information content of an HPO term, which is defined as the inverse of the probability of observing the phenotype in a given database. In blue, raw data from 10x Genomics linked-reads and Bionano optical maps was processed using default settings. Variants were called and filtered based on allele frequency and the pre-defined gene list. All rare SVs were vetted against regions known to be associated with deletion or duplication syndromes. All translocations and inversions were reported.



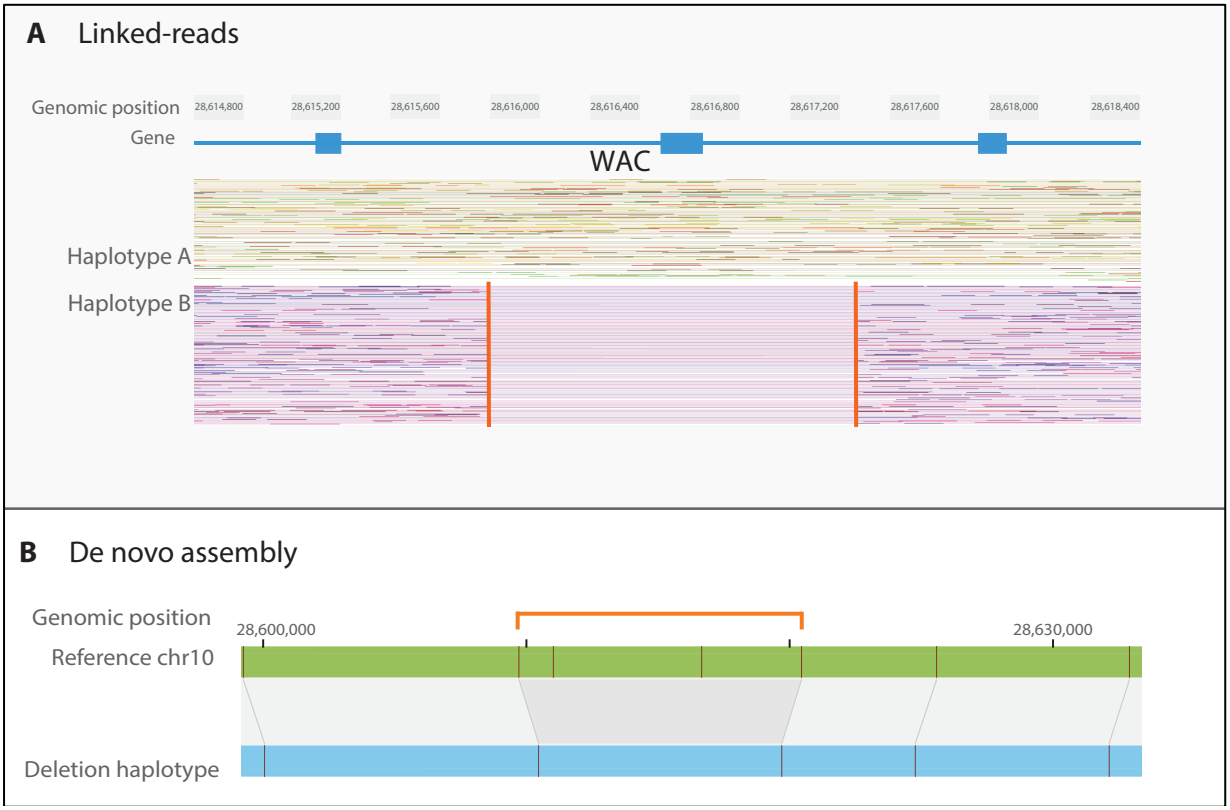

**Supplementary Figure S3.2: Duplicated region in NHEJ1.** Track view from UCSC Genome Browser of duplicated region from non-coding variant (highlighted in light-blue, chr2:219,102,933-219,134,970, GRCh38). The tracks displayed demonstrate how the duplication occurs in a region that interacts with regulatory elements. The H3K4Me1 mark shows where modification of histone proteins is highly suggestive of an enhancer. The GeneHancer track show associations between regulatory elements (grey bars) and their target genes, in this case, *IHH* and *NHEJ1* (purple and red lines).




**Supplementary Figure S3.3. Karyotype and ideogram of genomic rearrangement, 46, XY, t(1;9) (p32.3,p21).** Left panel: Karyotype with arrows points to breakpoint regions. Right panel: Ideogram of translocation.




**Supplementary Figure S3.4. Duplication of 7q11.23, subsequently translocated to 2q37.3 in an inverted orientation** [der(2)t(2;7)(q37.3; q11.23)dup(7)(q11.23;q11.23)]. **Panel A**, top: a de novo scaffold (rearrangement A in light blue) with the translocated region that aligns to chromosome 2 reference (in green) and suddenly chromosome 7 reference (in green) but in an inverted orientation. **Panel A**, bottom: a de novo scaffold (rearrangement A in light blue) of the segmental duplication in 7q11. **Panel C:** matrix view with unexpected barcode overlap between chr2:238,439,963-238,439,968 and chr7:72,774,109-73,100,000 (genome version GRCh38). **Panel D:** An ideogram of the rearrangement. The yellow-orange gradient depicts the inverse orientation. This is a maternally inherited rearrangement.




**Supplementary Figure S3.5. De novo heterozygous 1.5kb deletion disrupting *WAC*** (chr10:28,615,989-28,617.469, genome version GRCh38). **Panel A:** Deletion is seen by drop in coverage haplotype B (orange lines). **Panel B:** De novo assembly (light blue) demonstrates missing sequence labels with respect to reference (green). The orange bracket and gray triangle shows the deleted region.

**Supplementary Tables:**

| | Short-read WGS CNV | Linked-reads | De novo assembly |
|---|---|---|---|
| **Variant calls:** | | | |
| Inter-chromosomal events | 729 | 96 | 18 |
| Filtered inter-chromosomal events | 496 | 0 | 4 |
| **Diagnostic variant:** | | | |
| Identified | yes | yes | yes |
| Correct SV type | n/a | n/a | yes |
| Correct zygosity | yes | yes | no |

**Supplementary Table S3.1.** Comparison of inter-chromosomal events between short-read WGS CNV and genome assembly technologies for proband with a 46,XY,t(1;9)(p32.3;p21) rearrangement.

| | Short-read WGS CNV | Linked-reads | De novo assembly |
|---|---|---|---|
| **Variant calls:** | | | |
| Inter-chromosomal events | 427 | 59 | 19 |
| Filtered inter-chromosomal events | 305 | 2 | 2 |
| **Translocation identified:** | | | |
| Identified | yes | yes | yes |
| Correct SV type | n/a | n/a | yes |
| Correct zygosity | yes | yes | no |
| **Duplication identified:** | | | |
| Identified | no | yes | - |
| Correct SV type | n/a | yes | - |
| Correct zygosity | n/a | yes | - |

**Supplementary Table S3.2.** Comparison of inter-chromosomal events between short-read WGS CNV and genome assembly technologies for a proband with a complex rearrangement. Duplication of 7q11.23, subsequently translocated to 2q37.3 in an inverted orientation [der(2)t(2;7)(q37.3; q11.23)dup(7)(q11.23;q11.23)]

| | Short-read WGS CNV | Linked-reads | De novo assembly |
|---|---|---|---|
| **Variant calls:** | | | |
| Total number | 4510 | 4858 | 8644 |
| High quality | 3760 | 4712 | 1697 |
| Mean size ± SE (bp) | 40,1248 ± 14,166 | 53,827 ± 13,779 | 3,158 ± 671 |
| **Diagnostic variant:** | | | |
| Identified | yes | yes | yes |
| Correct SV type | yes | yes | yes |
| Correct zygosity | no | yes | no |

**Supplementary Table S3.3.** Comparison of deletion calls between short-read WGS CNV and genome assembly technologies. Table includes calls for proband with 36 kb *TANGO2* homozygous deletion.

| | Short-read WGS CNV | Linked-reads | De novo assembly |
|---|---|---|---|
| **Variant calls:** | | | |
| Total number | 4629 | 4649 | 8329 |
| High quality | 3895 | 4490 | 1666 |
| Mean size ± SE (bp) | 66,120 ± 20,869 | 88,137 ± 20,559 | 6,146 ± 1,185 |
| **Diagnostic variant:** | | | |
| Identified | yes | yes | yes |
| Correct SV type | yes | yes | yes |
| Correct zygosity | yes | yes | yes |

**Supplementary Table S3.4.** Comparison of deletion calls between short-read WGS CNV caller and genome assembly technologies. Table includes calls for proband with 1480 bp *WAC* deletion.

# CHAPTER 4: GENOMIC SEQUENCING AND KINDRED ANALYSIS ON THREE PATIENTS WITH HOMOZYGOUS FAMILIAL HYPERCHOLESTEROLEMIA

## 4.1 Abstract

Homozygous Familial Hypercholesterolemia (HoFH) is an inherited recessive condition associated with extremely high levels of low-density lipoprotein (LDL) cholesterol in affected individuals. It is usually caused by homozygous or compound heterozygous functional mutations in the LDL receptor (LDLR). A number of mutations causing FH have been reported in literature and such genetic heterogeneity presents great challenges for disease diagnosis. We aim to determine the likely genetic defects responsible for three cases of pediatric HoFH in two kindreds. We applied whole exome sequencing (WES) on the two probands to determine the likely functional variants among candidate FH genes. We additionally applied 10x Genomics (10xG) Linked-Reads whole genome sequencing (WGS) on one of the kindreds to identify potentially deleterious structural variants (SVs) underlying HoFH. A PCR-based screening assay was also established to detect the LDLR structural variant in a cohort of 641 patients with elevated LDL. In the Caucasian kindred, the FH homozygosity can be attributed to two compound heterozygous LDLR damaging variants, an exon 12 p.G592E missense mutation and a novel 3kb exon 1 deletion. By analyzing the 10xG phased data, we ascertained that this deletion allele was most likely to have originated from a Russian ancestor. In the Mexican kindred, the strikingly elevated LDL cholesterol level can be attributed to a homozygous frameshift LDLR variant p.E113fs. While the application of WES can provide a cost-effective way of identifying the genetic causes of FH, it often

lacks sensitivity for detecting structural variants. Our finding of the LDLR exon 1 deletion highlights the broader utility of Linked-Read WGS in detecting SVs in the clinical setting, especially when HoFH patients remain undiagnosed after WES.

### 4.2 Introduction

Mutations in the low-density lipoprotein (LDL) receptor gene (*LDLR;* OMIM accession: 606945) underlie most cases of familial hypercholesterolemia (FH). This monogenic disorder represents approximately 4% of patients with plasma levels of LDL cholesterol (LDL-C) above the 95$^{th}$ percentile and normal levels of other lipoproteins. The prevalence of heterozygous FH (HeFH) within the general population has been traditionally estimated to be in the range of 1 in 400 to 500 [75], but more contemporary studies suggested that the frequency can be as high as 1 in 200 in the European general population [76-78]. It is typically associated with premature coronary artery disease (CAD) and peripheral vascular disease. Other clinical findings include tendon xanthomas, xanthelasma, and arcus corneae. However, the phenotype of FH can manifest in a much more severe way when individuals harbor homozygous or compound HeFH variants. Individuals with HeFH have a mean LDL-C plasma level of 298 mg/dl whereas those with HoFH have a mean of 625 mg/dl [75]. Among 65 HeFH patients seen at the UCSF Lipid Clinic with known deleterious *LDLR* mutations, the mean LDL-C was 288 mg/dl (unpublished observations: CRP, JPK, MJM). Although deleterious mutations in the *LDLR* gene are the most common causes of severely elevated LDL-C, mutations in other genes account for a few cases. Among these genes are *APOB* (apolipoprotein B*;* OMIM accession: 107730), coding for a ligand for the

LDLR; *PCSK9* (proprotein convertase subtilisin/kexin type 9*;* OMIM accession: 607786); *LDLRAP1* (low density lipoprotein receptor adaptor protein 1*;* OMIM accession: 605747), *STAP1* (signal transducing adaptor family member 1*;* OMIM accession: 604298), and *CYP7A1* (cholesterol 7α-hydroxylase*;* OMIM accession: 118455) [79,80], *LIPA* (lysosomal acid lipase*;* OMIM accession: 613497) [81], *ABCG5* (ATP binding cassette subfamily G member 5*;* OMIM accession: 605459) and *ABCG8* (ATP binding cassette subfamily G member 8*;* OMIM accession: 605460) [80].

Establishing a definitive genetic etiology of FH is important because it is useful in guiding physicians in determining the most effective management on the basis of a specific mutation. For example, if a patient is homozygous, or compound heterozygous, for damaging or null *LDLR* variants, statins alone will have limited effect because these drugs work by increasing the LDL receptor expression on the cell surface thereby removing circulating LDL and remnant lipoproteins from the blood. In the case of defects in *ABCG5* and *ABCG8* transporters the dyslipidemia of FH is compounded by hyperabsorption of sterols. Agents with mechanisms of action not dependent on the presence of competent LDL receptors (such as ezetimibe, niacin, bile acid ion-exchange resins, and lomitapide) are useful. In patients with LIPA deficiency statins are likely to exacerbate the disease by increasing endocytosis of cholesteryl esters.

Here we report two families with 3 patients with rare HoFH. In the first kindred, one homozygous and seven heterozygous patients were identified. The other kindred included two homozygous and three heterozygous patients.

## 4.3 Materials and Methods

*Study Patients*

Two probands with HoFH were referred to the UCSF Pediatric Lipid Clinic. The first was a 5-year old Caucasian girl who presented with a severely elevated level of LDL and tuberous and tendinous xanthomas. The second was a 21-month old girl of Hispanic ancestry who, when first seen, also had a severely elevated level of LDL-C. She had some cutaneous xanthomas at birth. For each kindred, blood samples were also obtained from available family members; 11 relatives for the first kindred and 4 for the second. All study participants gave written informed consent prior to their enrollment in the study, which adhered to the World Medical Association Declaration of Helsinki and were approved by the University of California San Francisco (UCSF) Committee on Human Research Institutional Review Board as part of the UCSF Human Research Protection Program. Children were included with parental consent.

*DNA Preparation and Biochemical analyses*

Blood was collected from the two probands and their participating family members, after overnight fasting, in tubes containing 0.1% EDTA. When these samples were collected, none of the participants were taking lipid medications. Blood was centrifuged at 3,000 rpm for 20 min at 4°C and plasma separated. An automated chemical analyzer (COBAS Chemistry analyzer) was used to measure levels in plasma of total cholesterol (TC), HDL cholesterol (HDL-C), and triglyceride (TG) as described previously [82,83]. The Friedewald method was used to calculate LDL-C [84]. Genomic DNA was extracted using the Wizard purification kit (Qiagen, Germantown, MD). Genbank reference sequences

used for sanger validations were: *LDLR* NG_009060.1 ; *APOB* NG_011793.1; *MYL5* NM_002477.1; *MSR1* NG_012102.1; *ABCA1* NG_007981.1; *SPTY2D1* NM_194285.3; *LCAT* NG_009778.1; *PCTP* NM_021213.4; *LPIN1* NG_012843.2; *STAB1* NM_015136.3; *ABCC2* NG_011798.2; *PGS1* NM_024419.5; *OSBPL1A* NG_029432.1; *MC4R* NG_016441.1.

*Whole Exome Sequencing*

Genomic DNA derived from the probands of the two kindreds (1ug/sample) was sheared to an average fragment size of 300bp. DNA libraries were prepared using the KAPA DNA library preparation kits for Illumina sequencing platforms. Exons were captured using the Roche NimbleGen SeqCap EZ library probe and the captured libraries were sequenced on the HiSeq2500. Processing of image files was performed using standard protocol. Alignment to Hg19 reference was performed using Burrows-Wheeler Aligner (BWA v0.7.15). Picard v2.5.0 was used to mark duplicate and low-quality reads. Finally, GATK [85] was used for variant calling.

*Whole Exome Sequencing Variant Calling Pipeline*

Resulting variants were annotated with Annovar [86] to evaluate their effect on coding sequences, allele frequency in the general population, and the predicted level of pathogenicity. Synonymous, intronic, intergenic, and UTR variants were removed, along with variants with low read depth support (>20). Variants were intersected with a manually curated list of 594 lipid metabolism candidate genes including those in lipid metabolism pathways and GWAS hits (**Supplementary Table 4.1**; see original

manuscript). Using reported allele frequencies from the GnomAD, TOPMED, ExAC, common variants with a minor allele frequency >1% were also discarded, except for two potentially damaging variants. Minor allele frequencies, prior evidence of disease causality, and pathogenicity scores predicted by SIFT [87] and Polyphen-2 [88] were used to prioritize variants during the curation of the filtered variant list. We put emphasis on variants in genes with strong previous clinical and biochemical evidence. We ranked remaining variants based on SIFT and Polyphen-2 algorithm scores. Chosen variants were then manually verified by Sanger sequencing.

*Final Variant Assessment*

After the above described filtering process, we selected 8 potentially damaging variants for each kindred to study further. These 16 variants were examined by Sanger sequencing using DNA from the appropriate family members. Sanger sequencing was performed using the BigDye Terminator v3.1 and 3730xl DNA Analyzer, Applied Biosystems (Foster City, CA) on PCR products generated from primers designed using MacVector software (MacVector, Inc. Apex, NC).

*10x Genomics Whole Genome Sequencing*

Following WES for the proband of kindred 1 (**Figure 4.1**; subject 1 - 1), 10xG WGS was performed on her father (**Figure 4.1**; subject 2 -3). This was because a fresh blood sample was required here and we wanted to avoid drawing blood again from this child). High molecular weight genomic DNA extraction, sample indexing, and generation of partition barcoded libraries were performed according to the 10x Genomics

(Pleasanton, CA, USA) Chromium Genome User Guide and as published previously [89]. Raw reads were processed and aligned to the reference genome using 10x Genomics' Long Ranger software with the "wgs" pipeline with default settings. Deletion coordinates generated by Long Ranger were overlapped with the disease relevant gene list and the *LDLR* exon 1 deletion rose to the top of our variant candidate list. The *LDLR* deletion mutation observed here was confirmed by Sanger sequencing using primers spanning the breakpoint: forward: 5'-agctcctagaacttgcctatcct-3' and reverse 5'-gaggctgtctctctgcaactaat-3'.

*LDLR 3kb Exon 1 Deletion Screening Assay*

A PCR-based diagnostic assay was established to detect the presence of the *LDLR* 3kb deletion which eliminates the promoter region as well as exon 1. We followed the approach of Simard et al [90] to develop a multiplex PCR assay with 2 primers flanking the deletion breakpoints (as determined by 10xG WGS) and 2 primers within the deleted region to detect the presence of a wild-type allele. The first pair of primers were: 5'- agctcctagaacttgcctatcct-3' and 5'-tcgccacagagcacagcggaa-3', which generate a 254 bp product from the mutant allele. The second pair were: 5'-caacaaatcaagtcgcctgcc-3' and 5'-tgccattaccccacaagtctc-3', which yield a 481 bp product from the wild-type allele. We used this assay to screen all family members in kindred 1, and also a cohort of patients with elevated LDL-C. In this same cohort, using the previously described method [90] we found 3 patients with the French-Canadian *LDLR* exon 1 15.9 kb deletion (unpublished observations: CRP, JPK, MJM).

*Ancestry Determination for the LDLR 3kb Exon 1 Deletion*

To confirm that the patient was of European descent, we leveraged publicly available SNP genotyping data from the Human Genome Diversity Project (HGDP; download link: http://hagsc.org/hgdp/data/hgdp.zip). Tri-allelic variants, sex chromosome variants, and variants with MAF < 0.05 were discarded. Filtered variants were merged with the proband's vcf file followed by LD pruning (plink v1.9 --indep-pairwise 1000 5 0.5 ) using $r^2$=0.5 as the cutoff. Principal components analysis (PCA) was performed using plink and the top two PCs were used for plotting. Super-population labels were assigned based on a previous study [91]. To ascertain the ancestral origin of the *LDLR* exon 1 deletion, we utilized the phased SNPs generated from 10xG longranger. We modeled the expected conditional probability of observing the proband's genotype flanking the deletion events given population allele frequencies based on the HGDP dataset (See original manuscript under Supplementary Methods).

## 4.4 Results

### Clinical and other characteristics of the patients

#### *Kindred 1*

The proband with a clinical diagnosis of HoFH was a 5-year old Caucasian girl (**Figure 4.1**, subject 1-1) who presented in the UCSF Pediatric Lipid Clinic with severely elevated LDL-C (719 mg/dl), tuberous and tendon xanthomas. Also, her plasma level of HDL cholesterol (HDL-C) was abnormally low (36 mg/dl). Her father (subject 2-3) and paternal grandmother (subject 3-3) have elevated LDL-C consistent with HeFH, and each had coronary bypass surgery at ages 43 and 60, respectively. Her mother (subject

2-2), maternal aunt (subject 2-1) and maternal grandfather (subject 3-2) have elevated levels of LDL-C. The proband has an older sister (subject 1-2) and brother (subject 1-3), both with HeFH.

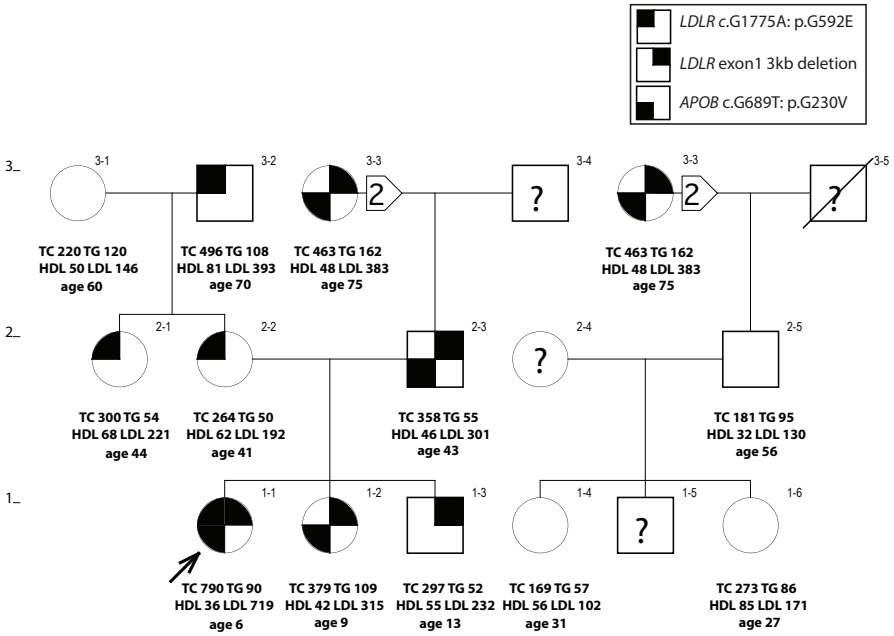



**Figure 4.1: Pedigree of a Caucasian-American family showing the distribution of the APOB (c.G689T: p.G230V) and LDLR (c.G1775A: p.G592E and exon1 3kb deletion) mutations.** Lipid values are in mg/dl. Ages and BMIs are those at time of blood draw.

### *Kindred 2*

In the kindred of Mexican ancestry, the proband (**Figure 4.2**, 1-1) was born with cutaneous xanthomas. She was referred to the UCSF Pediatric Lipid Clinic at age 21 months where a clinical diagnosis of HoFH was made. She had severe hypercholesterolemia with an LDL-C of 672 mg/dl. By age 5 years her LDL-C was 925 mg/dl. Her plasma level of HDL cholesterol (HDL-C) was low (20 mg/dl). A younger sister (subject 1-3) was subsequently seen at the clinic with the same phenotype and more extensive cutaneous xanthomas at birth. At age 7 months her LDL-C was 791

mg/dl, and HDL-C was 25 mg/dl. Another sister (subject 1-2) and the parents (subjects 2-1 and 2-2) all presented with a phenotype of HeFH. Two of the proband's grandparents are first cousins (the paternal grandmother is a first cousin of the maternal grandmother).

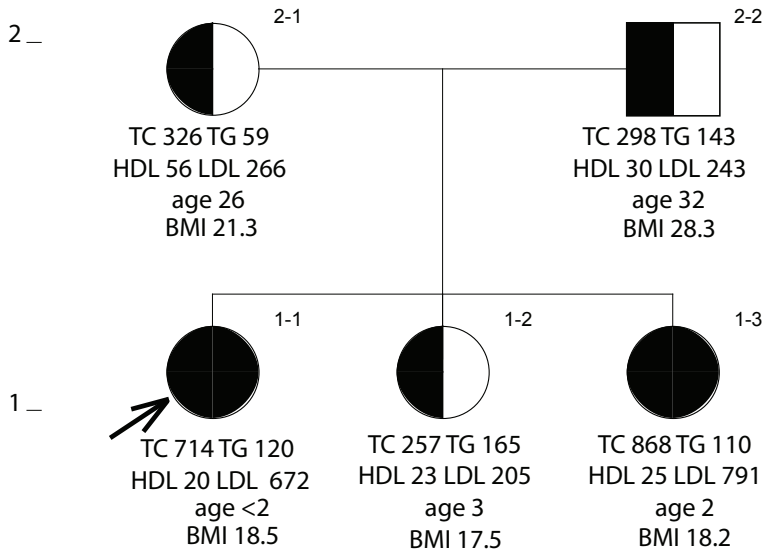



**Figure 4.2: Pedigree of a Mexican-American family showing the distribution of the LDLR frameshift mutation (c.337dupG p.E113fs).** Lipid values are in mg/dl. Ages and BMIs are those at time of blood draw.

**Whole exome sequencing and whole genome sequencing: kindred 1**

Eight potentially deleterious mutations (**Table 4.1**; see original manuscript), revealed by WES of DNA from the proband of kindred 1 (**Figure 4.1**; subject 1-1), were verified and assessed further by Sanger sequencing using all the available DNA samples from the rest of this family. However, the pattern of inheritance of these 8 heterozygous variants did not fully cosegregate with the elevated levels of LDL-C in this family (**Table 4.2**; see original manuscript). Notably, subject 1-3, despite his high level of LDL-C (232 mg/dl), did not carry either of the potentially deleterious *LDLR* (p.G592E) or *APOB* (p.G230V) variants. To evaluate comprehensively and robustly the genome in an unbiased

manner, we utilized Linked-Reads WGS (10x Genomics, Inc. Pleasanton, CA) in an attempt to identify potential disease-causing structural variants that might have gone undetected by WES. 10xG WGS was chosen because it has been reported that the use of virtual long reads allows for a much higher SV detection sensitivity [92]. This revealed a heterozygous 2977 bp deletion that included all of exon 1 (**Figure 4.3**). This mutation was confirmed by Sanger sequencing (**Figure 4.4**) of a PCR product using primers spanning the indicated breakpoint. This sequencing revealed an additional 4 bases (TTCG) between the deletion junction (**Figure 4.4**). To the best of our knowledge, this deletion has not been reported previously. Using the genomic information, we validated

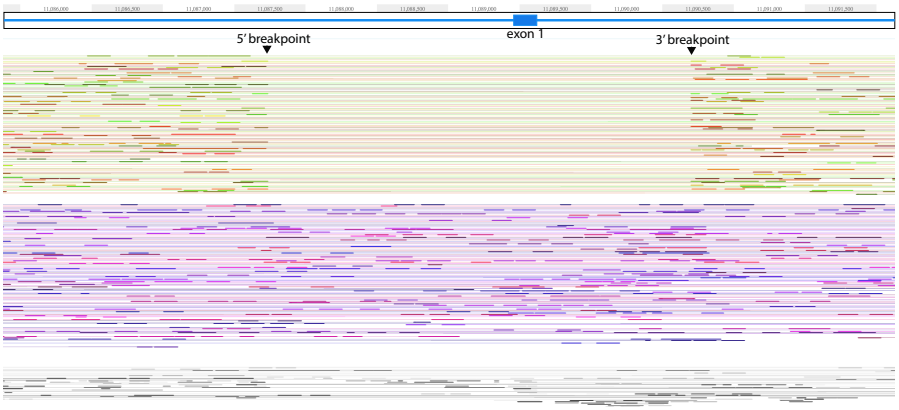



**Figure 4.3. Phased 10xG sequencing results from subject 2-3 in kindred 1 showing the breakpoints of the 3kb *LDLR* exon 1 deletion (GRCh38/hg38 coordinates).**

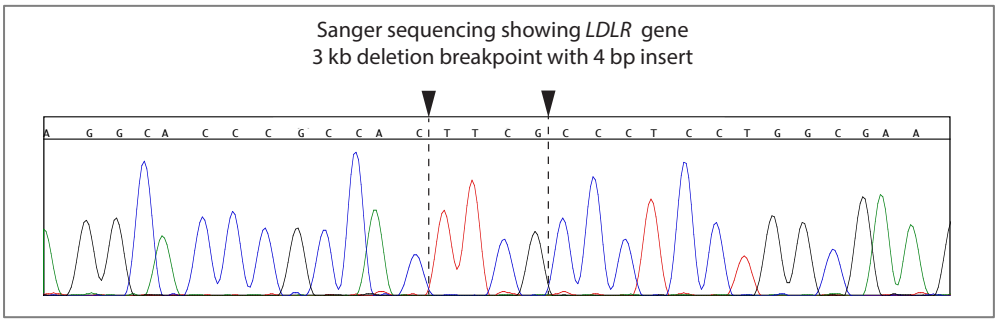



**Figure 4.4. Sanger chromatogram that confirms the breakpoints from 10xG sequencing.** The 4bp insertion lies between the breakpoints in chromosome 19 at nucleotides 11198406 (GRCh37/hg19) and 11201384.

that this proband of kindred 1 was of European descent (**Figure 4.5**). Furthermore, we analyzed the phased SNPs flanking the *LDLR* exon 1 deletion and predicted that this deletion allele was most likely to be originally derived from a Russian ancestor (**Figure 4.6**), which is consistent with the patient's self-reported family migration history.

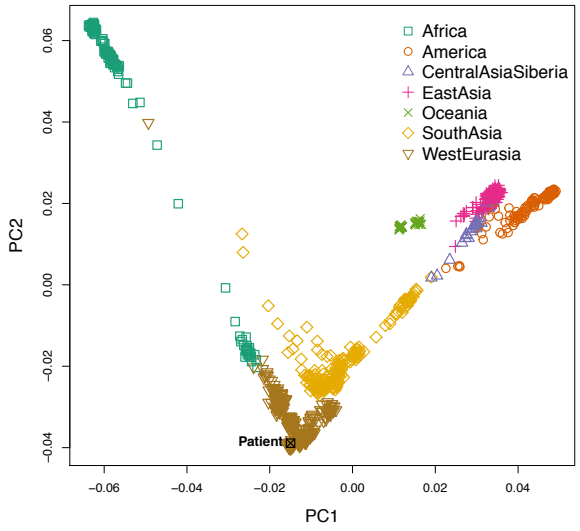



**Figure 4.5. PCA plot illustrating the first two principal components.** HGDP samples are used as the reference dataset and the patient from our study is highlighted in black.

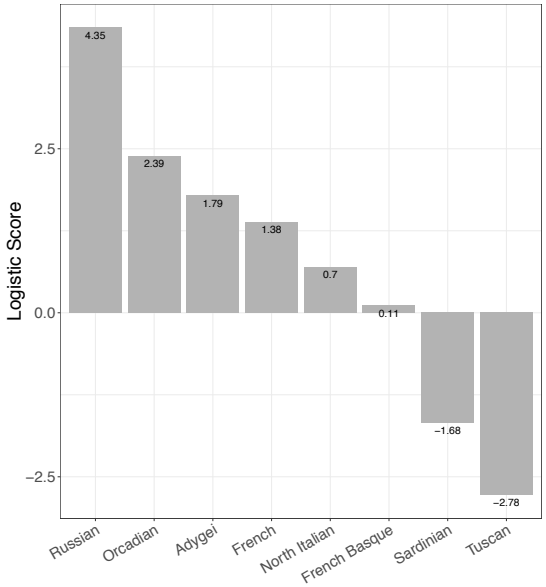



**Figure 4.6. Bar plot showing the distribution of the logistic probability of this entire phased block to be derived from the different European populations.**

To detect this 3kb deletion in other family members, and in a population of patients with elevated levels of LDL-C, we established a multiplex PCR assay similar in design to that described by Simard et al [90]. **Figure 4.7** (an agarose gel) shows the pattern of inheritance in kindred 1 with 5 family members carrying this mutation, including the proband (subject 1-1) and, notably, her brother (subject 1-3). The 481 bp fragment (**Figure 4.7**) is located within the deleted region and is a control for the presence of the

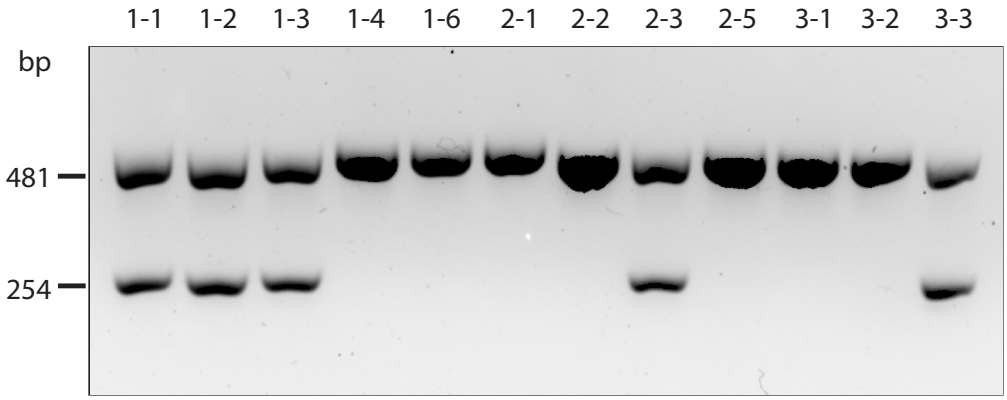



**Figure 4.7. Agarose gel of PCR products showing the carriers in kindred 1of the LDLR exon 1 3kb deletion mutation.** The 481 bp control band for the wild-type allele is from within the deleted region. The 254 bp band is a breakpoint-spanning band demonstrating the presence of the deletion.

wild-type allele, whereas the 254 bp band is a breakpoint-spanning fragment indicating the presence of the 3kb deletion allele. We subsequently screened 641 unrelated patients who were recruited into the UCSF Genomic Resource in Arteriosclerosis (GRA) [82,93], with age and sex-adjusted levels of LDL-C above the 95th percentile, but did not detect any additional patients with this exon 1 3kb deletion.

The pattern of inheritance of the 2 *LDLR* mutations, in addition to the potentially damaging novel *APOB* variant (c.G689T: p.G230V) in kindred 1 is shown in **Figure 4.1** and **Table 4.2** (see original manuscript). The 2 *LDLR* mutations now account for the pattern of LDL-C levels. The degree to which the *APOB* variant contributes an additional

adverse effect is not clear, especially as no one in the kindred carries that variant in the absence of a deleterious LDLR mutation.

The proband, in addition to her high level of LDL-C, has a greatly decreased level of HDL-C (36 mg/dl); less than the 5th percentile for her age [94]. No other member of the kindred has a level of HDL-C that low. Two of the potentially damaging variants listed in **Table 4.1** (see original manuscript) are in genes that have been directly associated with low levels of HDL-C [95]. These are *ABCA1* (ATP binding cassette subfamily A member 1) p.K776N and *LCAT* (lecithin-cholesterol acyltransferase) p.S232T. The first of these substitutions, which occurs in the 5th transmembrane helix of the ABCA1 protein, is a more likely cause of the low HDL-C based on several damaging predictions (**Table 4.1**; see original manuscript), and because the residue is conserved. However, two other carriers (subjects 2-3 and 3-3) do not have notably low HDL-C. Among the other potentially damaging variants, *MYL5* (myosin light chain 5) p.F88S, is in a gene that is one of 10 shown to be regulated by the master trans regulator gene *KLF14* [96], itself associated with levels of HDL-C [95] and type 2 diabetes. An additional carrier (subject 2-3), however, has a normal level of HDL-C.

**Exome and other sequencing: kindred 2**

Eight potentially deleterious variants detected by exome sequencing of DNA from the proband in kindred 2 (**Table 4.3**; see original manuscript) were verified, and the rest of the family were screened for them by Sanger sequencing. **Table 4.4** (see original manuscript) and **Figure 4/2** show the distribution of these variants within the family. *LDLR* mutation (p.E113fs) is the only mutation that is present in the homozygous state

(subjects 1-1 and 1-3). None of the other 7 seem to contribute to the pattern of elevated LDL-C.

The pattern of inheritance of 7 of the variants do not match the low levels of HDL-C seen in 4 of the 5 family members. Only one does, that of the *LPIN1* (lipin 1) p.T249K mutant. This mutation (rs141555457) is extremely rare, and is predicted to be probably pathogenic, but there is no record in ClinVar. The gene codes for an intracellular phosphatidic acid phosphohydrolase important as a regulator of lipid metabolism, especially hepatic VLDL-triglyceride secretion [97]. Deleterious mutations in *Lpin1* can cause fatty liver dystrophy and various dyslipidemias in mice [97]. In humans nonalcoholic fatty liver disease (NAFLD) is often accompanied by low levels of HDL-C. A heterozygous carrier of a predicted pathogenic mutation in *LPIN1* was responsible for statin-induced myopathy [98]. It is not clear whether a heterozygous functional variant in *LPIN1* could be the cause of the low levels of HDL-C seen here.

### 4.5 Discussion

Here we report 3 patients with HoFH from two kindreds. The Caucasian kindred had one homozygous and seven heterozygous individuals caused by two separate deleterious *LDLR* variants (exon 1 3kb deletion and p.G592E), with the possibility that a putatively deleterious novel *APOB* variant (p.G230V) contributes to the severity of FH in this family. The novel 3kb deletion originates from the proband's paternal family and was found among 5 individuals, while the p.G592E mutation is of maternal origin with 4 carriers. The *LDLR* exon 1 3kb deletion removing all of exon 1 and the proximal promotor is a null mutation, *a priori*. The p.G592E mutation occurs in exon 12 in a

spacer domain within the 400-amino acid region that contains 3 cysteine-rich repeats, each having homology to epidermal growth factor (EGF). This spacer region between EGF repeats 2 and 3 constitutes a six-bladed beta-propeller structure with 6 'YWTD' repeats [99]. The p.G592E mutations occurs within the 5$^{th}$ of these repeats at a highly conserved glycine residue [99]. This variant has been reported a number of times among Italian, Polish, German and Spanish patients [100-103]. With this variant, a class 5 mutation, the receptor has some residual activity [104] with the protein product binding and internalizing but failing to release LDL and not recycling to the plasma membrane. The affected residue is highly conserved.

Additionally, in kindred 1 there were 3 other potentially deleterious variants that might affect levels of LDL-C. One was the novel p.G333V variant in the *MSR1* gene (macrophage scavenger receptor 1). It encodes for 3 types of class A macrophage scavenger receptors via alternate splicing. These bind, among numerous other ligands, modified LDL particles. The second variant, p.C63Y, was in the *PCTP* gene (phosphatidylcholine transfer protein). Genetic studies indicate that the PCTP protein plays a role in HDL and VLDL metabolism and can affect LDL particle size [105]. The pattern of inheritance of these 2 variants seems to suggest they cannot account significantly for the elevated level of LDL-C. The *SPT2D1* gene (SPT2 chromatin protein domain containing 1) has been widely reported to be associated with plasma levels of total cholesterol [95,106,107]. The p.K107del, in this gene seems unlikely to be causative here because one carrier (subject 3-1) has a level of LDL-C in the normal range.

The proband in kindred 1 had a significantly decreased level of HDL-C. She is a carrier of the *ABCA1* p.K776N variant, which at first sight appears as a possible cause

here. However, a previous study showed that female carriers did not have lower levels of HDL-C compared to non-carriers [108]. Another report claimed that it didn't demonstrate an "unequivocal segregation" in a family study [109]. ClinVar reports it to be 'likely benign'. Also in kindred 1 there are 2 other carriers of this p.K776N variant, and neither has notably low level of HDL-C.

*LCAT* (lecithin-cholesterol acyltransferase) is another candidate gene associated with levels of HDL [95]. The *LCAT* p.S232T variant MAF (0.0176) makes it not so uncommon. In a study of those with either high (>95$^{th}$ percentile) or low HDL-C (<5$^{th}$ percentile), the p.S232T variant was found only among those with high HDL-C [110]. In kindred 1 it does not show segregation with levels of HDL-C . Likewise, another potential cause of low HDL-C, p.F88S variant in the *MYL5* gene (regulated by *KLF14)*, can probably be ruled out because the father has a normal level of HDL-C and carries this variant as well. Unless there are other undetected damaging mutations, it is likely that the low HDL-C observed in the proband is a consequence of the extremely high level of LDL-C. FH heterozygotes and homozygotes tend to have low levels of HDL with homozygotes having the lowest levels, although reason for this is unknown [75].

In the second kindred, the single *LDLR* frameshift mutant (p.E113fs) is sufficient to explain the strikingly elevated levels of LDL-C. This extremely rare, *a priori*, damaging variant has a pathogenic ClinVar entry. It has been previously reported among Mexican FH heterozygotes [111]. To our knowledge, the two homozygous cases that we report here are the first such cases to be described. Unlike the first kindred, the presence of low HDL-C in the proband of kindred 2 does not seem to be explained simply as a result of the high levels of LDL-C. The *OSBPL1A* (oxysterol-binding protein-like protein 1a)

variant, p.C39fs (rs766116535) has previously been reported to be associated with decreased levels of HDL-C [112,113]. This gene codes for one of a large family of intracellular oxysterol-binding proteins that act as lipid (sterols and anionic phospholipids) receptors that are involved in lipid transfer and signaling. **Table 4.4** clearly shows that the *OSBPL1A* p.C39fs does not explain the very low levels of HDL-C seen in 4 members of this family, with the second carrier (subject 2-1) of this variant having a normal level (56 mg/dl). Without some other cause for the extremely low levels of HDL-C, we believe that the *LPIN1* variant (p.T249K) is the most likely suspect.

It is important to underline that the clinical management of HoFH patient is challenging, and knowing the underlying genetic defects has significant prognostic and therapeutic implications. Although the application of NGS in clinical settings is becoming more accessible, the number of negative diagnostic results remains high. Here, we were able to pinpoint a clinically relevant structural variant using Linked-Reads in kindred 1. While analyzing SVs in the clinical context is still in its nascent state, the availability of sequencing technology using long-reads or virtual long-reads allows for a much better detection sensitivity for SVs, thus expanding the search space for disease-causing variants that have been previously invisible to geneticists. Subject to consideration of benefit vs cost, the broad implementation of more advanced WGS technology may improve diagnoses for patients with HoFH.

## CHAPTER 5: FUTURE DIRECTIONS

To have gone from reading 3 billion bases for the first time to the implementation of genomic medicine in less than a decade is a stunning trajectory. Although wide adaptation of routine sequencing in healthcare has not yet been fully realized in the US, the National Health Service of the UK has already implemented such a system to sequence every severely ill child, and the plan will be expanded to include every single newborn in the future. As the cost of sequencing declines, it will become more tenable for other countries to pursue similar strategy. Following this trend, more efficient file compression algorithms, data storage solutions, and analytical software must be in place to streamline the process. Meanwhile, the involvement of artificial intelligence will most likely rise in various clinical applications as well.

In terms of elucidating the full picture of the genome, long read technology has been proven to be useful for discovering SVs and resolving sequences in segmental duplications. However, the low sequencing throughput currently limits its potential applications. It is envisioned that with the use of highly accurate long reads coupled with phased *de novo* assembly, comprehensive SV identification on every individual genome will be feasible in the near future. Having the power to unlock every part of the genome will not only accelerate scientific discoveries for understanding health and disease, but it also will also unfold new diagnostic and therapeutic strategies that no one has yet imagined.

## REFERENCES


1 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

2 Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

3 Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**, 849-864, doi:10.1101/gr.213611.116 (2017).

4 Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 1080, doi:10.1016/j.cell.2019.04.032 (2019).

5 Rakocevic, G. *et al.* Fast and accurate genomic analyses using genome graphs. *Nat Genet* **51**, 354-362, doi:10.1038/s41588-018-0316-4 (2019).

6 Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155-1162, doi:10.1038/s41587-019-0217-9 (2019).

7 Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**, 2193-2198, doi:10.1093/bioinformatics/bty841 (2019).

8 Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol* **19**, 90, doi:10.1186/s13059-018-1462-9 (2018).

9 Jain, M. *et al.* MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. *F1000Res* **6**, 760, doi:10.12688/f1000research.11354.1 (2017).

10 Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58, doi:10.1038/nature09298 (2010).

11 Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

12 Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**, 1784, doi:10.1038/s41467-018-08148-z (2019).

13 Aneichyk, T. *et al.* Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell* **172**, 897-909.e821, doi:10.1016/j.cell.2018.02.011 (2018).

14 Ishiura, H. *et al.* Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat Genet* **50**, 581-590, doi:10.1038/s41588-018-0067-2 (2018).

15 Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222, doi:10.1093/bioinformatics/btv710 (2016).

16 Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84, doi:10.1186/gb-2014-15-6-r84 (2014).

17 Miller, D. T. *et al.* Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* **86**, 749-764, doi:10.1016/j.ajhg.2010.04.006 (2010).

18 Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic Medicine-Progress, Pitfalls, and Promise. *Cell* **177**, 45-57, doi:10.1016/j.cell.2019.02.003 (2019).

19 Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* **369**, 1502-1511, doi:10.1056/NEJMoa1306555 (2013).

20 Eldomery, M. K. *et al.* Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med* **9**, 26, doi:10.1186/s13073-017-0412-6 (2017).

21 Clark, M. M. *et al.* Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci Transl Med* **11**, doi:10.1126/scitranslmed.aat6177 (2019).

22 Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663-675.e619, doi:10.1016/j.cell.2018.12.019 (2019).

23 Wong, K., Levy-Sakin, M. & Kwok, P.-Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat Commun.* **In press** (2018).

24 Levy-Sakin, M. *et al.* Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat Commun* **10**, 1025, doi:10.1038/s41467-019-08992-7 (2019).

25 Telenti, A. *et al.* Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci U S A* **113**, 11901-11906, doi:10.1073/pnas.1613365113 (2016).

26 Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206, doi:10.1038/nature18964 (2016).

27 Kehr, B. *et al.* Diversity in non-repetitive human sequences not found in the reference genome. *Nat Genet* **49**, 588-593, doi:10.1038/ng.3801 (2017).

28 Levy-Sakin, M. *et al.* (Manuscript under review, 2018).

29 Demaerel, W. *et al.* The 22q11 low copy repeats are characterized by unprecedented size and structural variability. *Genome Res* **29**, 1389-1401, doi:10.1101/gr.248682.119 (2019).

30 Rakocevic, G. *et al.* Fast and Accurate Genomic Analyses using Genome Graphs. *bioRxiv*, doi:10.1101/194530 (2017).

31 Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res* **27**, 757-767, doi:10.1101/gr.214874.116 (2017).

32 Seo, J. S. *et al.* De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243-247, doi:10.1038/nature20098 (2016).

33 Shi, L. *et al.* Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun* **7**, 12065, doi:10.1038/ncomms12065 (2016).

34 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12, doi:10.1186/gb-2004-5-2-r12 (2004).

35 Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021-3023, doi:10.1093/bioinformatics/btw369 (2016).

36 Smit, A., Hubley, R, Green, P. *RepeatMasker Open-4.0*, 2015).

37 Lassmann, T. & Sonnhammer, E. L. Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* **6**, 298, doi:10.1186/1471-2105-6-298 (2005).

38 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580, doi:10.1093/nar/27.2.573 (1999).

39 Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420-426, doi:10.1126/science.1149504 (2007).

40 Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl regulatory build. *Genome Biol* **16**, 56, doi:10.1186/s13059-015-0621-5 (2015).

41 Chen, S. *et al.* Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol* **20**, 291, doi:10.1186/s13059-019-1909-7 (2019).

42 Ma, W. NUI projection. *Zenodo*, doi:10.5281/zenodo.3496215 (2019).

43 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

44 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).

45 Wiederstein, J. L. *et al.* Skeletal Muscle-Specific Methyltransferase METTL21C Trimethylates p97 and Regulates Autophagy-Associated Protein Breakdown. *Cell Rep* **23**, 1342-1356, doi:10.1016/j.celrep.2018.03.136 (2018).

46 Potkin, S. G. *et al.* Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One* **4**, e6501, doi:10.1371/journal.pone.0006501 (2009).

47 Moessner, R. *et al.* Contribution of SHANK3 mutations to autism spectrum disorder. *Am J Hum Genet* **81**, 1289-1297, doi:10.1086/522590 (2007).

48 Phelan, K. & McDermid, H. E. The 22q13.3 Deletion Syndrome (Phelan-McDermid Syndrome). *Mol Syndromol* **2**, 186-201, doi:000334260 (2012).

49 Betancur, C. & Buxbaum, J. D. SHANK3 haploinsufficiency: a "common" but underdiagnosed highly penetrant monogenic cause of autism spectrum disorders. *Mol Autism* **4**, 17, doi:10.1186/2040-2392-4-17 (2013).

50 Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233, doi:10.1016/j.neuron.2015.09.016 (2015).

51 Zhou, Y. *et al.* Mice with Shank3 Mutations Associated with ASD and Schizophrenia Display Both Shared and Distinct Defects. *Neuron* **89**, 147-162, doi:10.1016/j.neuron.2015.11.023 (2016).

52 Consortium, G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110 (2015).

53 Reid, C. J. & Harris, A. Developmental expression of mucin genes in the human gastrointestinal system. *Gut* **42**, 220-226, doi:10.1136/gut.42.2.220 (1998).

54 Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet* **51**, 30-35, doi:10.1038/s41588-018-0273-y (2019).

55 Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat Biotechnol* **28**, 57-63, doi:10.1038/nbt.1596 (2010).

56 Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat Rev Genet*, doi:10.1038/s41576-020-0210-7 (2020).

57 Biesecker, L. G. & Green, R. C. Diagnostic clinical genome and exome sequencing. *N Engl J Med* **371**, 1170, doi:10.1056/NEJMc1408914 (2014).

58 Levy, S. E. & Myers, R. M. Advancements in Next-Generation Sequencing. *Annu Rev Genomics Hum Genet* **17**, 95-115, doi:10.1146/annurev-genom-083115-022413 (2016).

59 Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405-424, doi:10.1038/gim.2015.30 (2015).

60 Schwarze, K., Buchanan, J., Taylor, J. C. & Wordsworth, S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med* **20**, 1122-1130, doi:10.1038/gim.2017.247 (2018).

61 Nambot, S. *et al.* Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial

interest of prospective annual reanalysis. *Genet Med* **20**, 645-654, doi:10.1038/gim.2017.162 (2018).

62 Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res* **29**, 635-645, doi:10.1101/gr.234443.118 (2019).

63 Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* **20**, 117, doi:10.1186/s13059-019-1720-5 (2019).

64 Mostovoy, Y. *et al.* A hybrid approach for de novo human genome sequence assembly and phasing. *Nat Methods* **13**, 587-590, doi:10.1038/nmeth.3865 (2016).

65 Deisseroth, C. A. *et al.* ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genet Med* **21**, 1585-1593, doi:10.1038/s41436-018-0381-1 (2019).

66 Klopocki, E. *et al.* Copy-number variations involving the IHH locus are associated with syndactyly and craniosynostosis. *Am J Hum Genet* **88**, 70-75, doi:10.1016/j.ajhg.2010.11.006 (2011).

67 Will, A. J. *et al.* Composition and dosage of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). *Nat Genet* **49**, 1539-1545, doi:10.1038/ng.3939 (2017).

68 Peled, A. *et al.* Mutations in TSPEAR, Encoding a Regulator of Notch Signaling, Affect Tooth and Hair Follicle Morphogenesis. *PLoS Genet* **12**, e1006369, doi:10.1371/journal.pgen.1006369 (2016).

69 Hamedani, S. *The common fragile site genes CNTLN and LINGO2 are associated with increased genomic instability in different tumors*. (2010).

70 Barr, E. K. & Applebaum, M. A. Genetic Predisposition to Neuroblastoma. *Children (Basel)* **5**, doi:10.3390/children5090119 (2018).

71 Gambin, T. *et al.* Identification of novel candidate disease genes from de novo exonic copy number variants. *Genome Med* **9**, 83, doi:10.1186/s13073-017-0472-7 (2017).

72 Bonaglia, M. C. *et al.* De novo unbalanced translocations have a complex history/aetiology. *Hum Genet* **137**, 817-829, doi:10.1007/s00439-018-1941-9 (2018).

73 Du, R. *et al.* Identification of likely pathogenic and known variants in TSPEAR, LAMB3, BCOR, and WNT10A in four Turkish families with tooth agenesis. *Hum Genet* **137**, 689-703, doi:10.1007/s00439-018-1907-y (2018).

74 Wong, K. H. Y., Levy-Sakin, M. & Kwok, P. Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat Commun* **9**, 3040, doi:10.1038/s41467-018-05513-w (2018).

75 Brown, M. S., Hobbs, H. H. & Goldstein, J. L. Chapter 120: Familial Hypercholesterolemia. *Metabolic & Molecular Bases of Inherited Disease*, 1-122 (2007).

76 Benn, M., Watts, G. F., Tybjaerg-Hansen, A. & Nordestgaard, B. G. Familial hypercholesterolemia in the danish general population: prevalence, coronary artery disease, and cholesterol-lowering medication. *The Journal of clinical endocrinology and metabolism* **97**, 3956-3964, doi:10.1210/jc.2012-1563 (2012).

77 Benn, M., Watts, G. F., Tybjaerg-Hansen, A. & Nordestgaard, B. G. Mutations causative of familial hypercholesterolaemia: screening of 98 098 individuals from the Copenhagen General Population Study estimated a prevalence of 1 in 217. *European Heart Journal* **37**, 1384-1394, doi:10.1093/eurheartj/ehw028 (2016).

78 Nordestgaard, B. G. *et al.* in *European heart journal* Vol. 34 3478-3490a (2013).

79 Hegele, R. A. *et al.* Targeted next-generation sequencing in monogenic dyslipidemias. *Current opinion in lipidology* **26**, 103-113, doi:10.1097/MOL.0000000000000163 (2015).

80 Pullinger, C. R., Kane, J. P. & Malloy, M. J. Primary hypercholesterolemia: genetic causes and treatment of five monogenic disorders. *Expert review of cardiovascular therapy* **1**, 107-119, doi:10.1586/14779072.1.1.107 (2003).

81 Pullinger, C. R. *et al.* Identification and metabolic profiling of patients with lysosomal acid lipase deficiency. *Journal of Clinical Lipidology* **9**, 716-726.e711, doi:10.1016/j.jacl.2015.07.008 (2015).

82 Pullinger, C. R. *et al.* An apolipoprotein A-V gene SNP is associated with marked hypertriglyceridemia among Asian-American patients. *J Lipid Res* **49**, 1846-1854, doi:P800011-JLR200 [pii] 10.1194/jlr.P800011-JLR200 (2008).

83 Pullinger, C. R. *et al.* Familial ligand-defective apolipoprotein B - Identification of a new mutation that decreases LDL receptor binding affinity. *The Journal of clinical investigation* **95**, 1225-1234 (1995).

84 Friedewald, W. T., Levy, R. I. & Fredrickson, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin. Chem.* **18**, 499-502 (1972).

85 Mckenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

86 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164, doi:10.1093/nar/gkq603 (2010).

87 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31**, 3812-3814 (2003).

88 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L Haines [et al]* **Chapter 7**, Unit7.20, doi:10.1002/0471142905.hg0720s76 (2013).

89 Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Research* **27**, 757-767, doi:10.1101/gr.214874.116 (2017).

90 Simard, L. R. *et al.* The Delta>15 Kb deletion French Canadian founder mutation in familial hypercholesterolemia: rapid polymerase chain reaction-based diagnostic assay and prevalence in Quebec. *Clinical genetics* **65**, 202-208 (2004).

91 Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201-206, doi:10.1038/nature18964 (2016).

92 Wong, K. H. Y., Levy-Sakin, M. & Kwok, P.-y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nature communications* **9**, 3040, doi:10.1038/s41467-018-05513-w (2018).

93 Shiffman, D. *et al.* Identification of four gene variants associated with myocardial infarction. *Am J Hum Genet* **77**, 596-605 (2005).

94 Program., L. R. C. Population Studies Data Book: Vol. I, The Prevalence Study. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health (NIH publication number 80-1527), Washington, DC. (1980).

95 Lange, L. A., Willer, C. J. & Rich, S. S. Recent developments in genome and exome-wide analyses of plasma lipids. *Current opinion in lipidology* **26**, 96-102, doi:10.1097/MOL.0000000000000159 (2015).

96 Consortium, t. M. *et al.* Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nature Genetics*, doi:10.1038/ng.833 (2011).

97 Chen, Z. *et al.* Alterations in hepatic metabolism in fld mice reveal a role for lipin 1 in regulating VLDL-triacylglyceride secretion. *Arteriosclerosis, Thrombosis, and Vascular Biology* **28**, 1738-1744, doi:10.1161/ATVBAHA.108.171538 (2008).

98 Zeharia, A. *et al.* Mutations in LPIN1 cause recurrent acute myoglobinuria in childhood. *American journal of human genetics* **83**, 489-494, doi:10.1016/j.ajhg.2008.09.002 (2008).

99 Jeon, H. *et al.* Implications for familial hypercholesterolemia from the structure of the LDL receptor YWTD-EGF domain pair. *Nature structural biology* **8**, 499-504, doi:10.1038/88556 (2001).

100 Hobbs, H. H., Brown, M. S. & Goldstein, J. L. Molecular genetics of the LDL receptor gene in familial hypercholesterolemia. *Human mutation* **1**, 445-466, doi:10.1002/humu.1380010602 (1992).

101 Górski, B., Kubalska, J., Naruszewicz, M. & Lubiński, J. LDL-R and Apo-B-100 gene mutations in Polish familial hypercholesterolemias. *Human Genetics* **102**, 562-565 (1998).

102 Bochmann, H. *et al.* Eight novel LDL receptor gene mutations among patients under LDL apheresis in Dresden and Leipzig. *Human mutation* **17**, 76-77, doi:10.1002/1098-1004(2001)17:1<76::AID-HUMU18>3.0.CO;2-Y (2001).

103 Mozas, P. *et al.* Molecular characterization of familial hypercholesterolemia in Spain: identification of 39 novel and 77 recurrent mutations in LDLR. *Human mutation* **24**, 187, doi:10.1002/humu.9264 (2004).

104 Romano, M. *et al.* Identification and functional characterization of LDLR mutations in familial hypercholesterolemia patients from Southern Italy. *Atherosclerosis* **210**, 493-496, doi:10.1016/j.atherosclerosis.2009.11.051 (2010).

105 Dolley, G. *et al.* Influences of the phosphatidylcholine transfer protein gene variants on the LDL peak particle size. *Atherosclerosis* **195**, 297-302, doi:10.1016/j.atherosclerosis.2007.01.002 (2007).

106 Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713, doi:10.1038/nature09270 (2010).

107 Asselbergs, F. W. *et al.* Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *American journal of human genetics* **91**, 823-838, doi:10.1016/j.ajhg.2012.08.032 (2012).

108 Frikke-Schmidt, R., Nordestgaard, B. G., Schnohr, P., Steffensen, R. & Tybjaerg-Hansen, A. Mutation in ABCA1 predicted risk of ischemic heart disease in the Copenhagen City Heart Study Population. *Journal of the American College of Cardiology* **46**, 1516-1520, doi:10.1016/j.jacc.2005.06.066 (2005).

109 Alrasadi, K., Ruel, I. L., Marcil, M. & Genest, J. Functional mutations of the ABCA1 gene in subjects of French-Canadian descent with HDL deficiency. *Atherosclerosis* **188**, 281-291, doi:10.1016/j.atherosclerosis.2005.10.048 (2006).

110 Naseri, M., Hedayati, M., Daneshpour, M. S., Bandarian, F. & Azizi, F. Identification of genetic variants of lecithin cholesterol acyltransferase in individuals with high HDL-C levels. *Molecular medicine reports* **10**, 496-502, doi:10.3892/mmr.2014.2177 (2014).

111 Robles-Osorio, L. *et al.* Genetic heterogeneity of autosomal dominant hypercholesterolemia in Mexico. *Archives of medical research* **37**, 102-108, doi:10.1016/j.arcmed.2005.04.018 (2006).

112 Motazacker, M. M. *et al.* Evidence of a polygenic origin of extreme high-density lipoprotein cholesterol levels. *Arteriosclerosis, Thrombosis, and Vascular Biology* **33**, 1521-1528, doi:10.1161/ATVBAHA.113.301505 (2013).

113 Motazacker, M. M. *et al.* A loss-of-function variant in OSBPL1A predisposes to low plasma HDL cholesterol levels and impaired cholesterol efflux capacity. *Atherosclerosis* **249**, 140-147, doi:10.1016/j.atherosclerosis.2016.04.005 (2016).

## APPENDIX

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-05513-w OPEN

# De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations

Karen H.Y. Wong [1], Michal Levy-Sakin[1] & Pui-Yan Kwok [1,2,3]

The human reference genome is used extensively in modern biological research. However, a single consensus representation is inadequate to provide a universal reference structure because it is a haplotype among many in the human population. Using 10× Genomics (10×G) "Linked-Read" technology, we perform whole genome sequencing (WGS) and de novo assembly on 17 individuals across five populations. We identify 1842 breakpoint-resolved non-reference unique insertions (NUIs) that, in aggregate, add up to 2.1 Mb of so far undescribed genomic content. Among these, 64% are considered ancestral to humans since they are found in non-human primate genomes. Furthermore, 37% of the NUIs can be found in the human transcriptome and 14% likely arose from *Alu*-recombination-mediated deletion. Our results underline the need of a set of human reference genomes that includes a comprehensive list of alternative haplotypes to depict the complete spectrum of genetic diversity across populations.

[1] Cardiovascular Research Institute, University of California, San Francisco, San Francisco, 94158 CA, USA. [2] Institute for Human Genetics, University of California, San Francisco, San Francisco, 94143 CA, USA. [3] Department of Dermatology, University of California, San Francisco, San Francisco, 94115 CA, USA. Correspondence and requests for materials should be addressed to P.-Y.K. (email: pui.kwok@ucsf.edu)

Next-generation sequencing (NGS) is being used in numerous ways in both basic and clinical research. At the core of many studies is the human reference genome, which provides a genomic scaffold for read alignment and downstream identification of genomic variants[1–3]. However, despite the tremendous sequencing effort and methodological advances, the creation of a comprehensive human reference genome set that can represent the genetic variations across populations is yet to be realized. While the current human reference genome assembly (hg38.p11) is the most complete version to date, only 261 alternate loci are included to provide representation of haplotype diversity[4]. The remainder of the genome is still being represented as a single consensus haplotype. Many lines of evidence suggest that there are unique insertional sequences not currently represented in the reference genome[3,5–7]. We use the term non-reference unique insertions (NUIs) to describe unique sequences that are found in other individuals but not in the human reference genome. Specifically, NUIs are full-length insertions that harbor at least 50 bp of non-repetitive sequences not found in the hg38 reference set, including alternative haplotypes and patches. Translocation events are also excluded from the dataset. The exact definition and selection procedures used in this paper are described in Methods.

Accumulating evidence has shown that structural variations (SVs) contribute significantly to genome diversity but SVs are not identified in routine NGS[3,7–10]. While numerous deletion algorithms can pinpoint deletion breakpoints with high sensitivity because they can be detected in a single short read, it is challenging to detect and localize long insertions comprehensively because whole genome de novo assembly is required. Several studies have identified long, non-reference insertions and shed new light on the complexity of the human genome. The 1000 Genomes Project (1000GP) structural variation consortium discovered 128 non-reference insertions in their pilot phase SV release set[11]. Other SV detection studies like the Genome of the Netherlands (GoNL) project[3], the Simon Genome Diversity project[12], and the deep sequencing of 10,000 individuals[6] have, respectively, found 7718, 950, and 4876 genomic segments not mapped to the human reference genomes. However, no breakpoint information was provided by any of these groups. Recently, deCODE genetics/Amgen discovered 3791 breakpoint-resolved non-reference sequences from 15,219 Icelandic individuals[5]. Despite the large sample size, this study involves a homogenous population and does not capture the global genetic variation. Furthermore, all the published studies exclusively used Illumina WGS with <45X coverage, without any long-range sequence information necessary to bridge repetitive elements for accurate NUI placement.

In this study, we analyze in silico phased (haplotype-resolved), de novo human genome assemblies generated with the 10×G "Linked-Read" technology. Using a custom pipeline, we identify 1842 breakpoint-resolved NUIs in 17 1000GP individuals originating from 5 different super-populations. Also, we find that these NUIs follow a population-specific pattern, which is consistent with the previous studies using single nucleotide polymorphisms[13,14]. Over half of the NUIs are also found in non-human primate genomes, suggesting that they are ancestral to humans. Furthermore, some NUIs align uniquely to entries in the Expressed Sequence Tag database (dbEST)[15] or reads from RNA sequencing (RNA-Seq) experiments, indicating that they are part of the transcriptome. Our results underline the need of a set of human reference genomes that fully incorporates the diversity of sequences across populations as sequencing of individuals across the world becomes routine.

## Results

**Genome assemblies of 17 individuals from 5 populations**. Based on the 1000GP, 14 individuals representing populations most distinctive from one another were selected for 10×G WGS using "Linked-Read" technology. We additionally downloaded 10×G WGS data of 3 other 1000GP samples from the 10×G website. This dataset includes 5 Africans (AFR), 3 Americans (AMR), 4 East Asians (EAS), 3 Europeans (EUR), and 2 South Asians (SAS). All samples were sequenced to ~60X mean read depth (except for HG00733 and NA19240, that were sequenced to 79X and 89X, respectively) with a median molecule length of 103 kb. Sequencing reads were aligned to the hg38 human reference genome using the software Long Ranger. De novo assemblies of these samples were also generated using Supernova to yield diploid pseudo-haplotypes. The average scaffold and phased block N50s for these assemblies are 18 and 3 Mb, respectively. Full summary statistics of the assemblies are found in Supplementary Table 1.

**NUI discovery pipeline and validation**. An alignment-based and de novo assembly-based custom pipeline was built to identify NUIs. The pipeline first extracted high-quality sequence reads that did not align well to the human reference genome hg38 (Supplementary Fig. 1 and Methods). These reads were aligned to the individual's diploid de novo assembly and the regions containing clusters of the reads were identified. The corresponding assembled sequences and their flanking regions were extracted and realigned to the reference genome to compute the exact breakpoints. NUIs aligning to any alternative haplotypes or patches were removed from downstream analysis. NUIs from all 17 samples were merged to generate a unified, non-redundant call set. To investigate whether the NUI count per individual was sensitive to the quality of the sequencing reads and the sequencing depth, we computed the Pearson correlations between these parameters. The Pearson correlations ($r$) were 0.37 ($p = 0.14$; Supplementary Fig. 2a) and 0.26 ($p = 0.32$; Supplementary Fig. 2b), respectively, indicating that there was no such evidence for technical bias.

To validate our call set, we used an orthogonal approach to detect insertional sequences. We generated optical genome maps with fluorescent labels marking Nt.BspQI nicking endonuclease recognition sites. SVs could be inferred by comparing the distances between two adjacent labels. Due to the inherent constraints of optical mapping, this strategy only allowed us to detect large insertions (>2 kb in size). We applied two SV calling pipelines: one from BioNano Genomics and one from a modified version of OMSV[16]. Detailed methods used to validate the NUI call set are described in Methods. Of all the NUIs over 2 kb in length, the average precision rate is 88.4% (Supplementary Table 2), corresponding to an average of 61 validated insertions out of 69 called NUIs per sample. Most of the discordant NUIs are between 2 and 3 kb in size, a size range where the BioNano SV calling algorithm is known to have a higher false positive rate due to sizing errors, especially in regions with sparse label density.

**The structure of genetic diversity across populations**. Overall, the unified, non-redundant NUI call set includes 1842 variants. They add up to 2.1 Mb genomic sequences not found in the human reference genome or the alternative haplotypes and patches (Table 1; Supplementary Data 1). Each individual has an average of 690 NUIs (Table 1; Supplementary Data 2) that represent 711 kb of so far undescribed genomic content. Of the NUIs identified, 32% are shared across all five populations while 5% are found in all 17 individuals. The NUIs are found on all human chromosomes (Fig. 1a), with 25% are <131 bp in size, half are <450 bp, and 75% are <1260 bp (Fig. 1b).

As expected, Africans have the most NUIs while Europeans have the fewest (Fig. 1c; ANOVA $F(4,12) = 5.643$, $p = 0.0086$ followed by Tukey [AFR-EAS] $p = 0.0389033$; [AFR-EUR] $p = 0.0067794$). Significant difference in NUI count is observed

**Table 1 Non-reference unique insertions summary**

| Sample | Super population | NUI count | Total bp | Median NUI count in each population |
|---|---|---|---|---|
| HG02623 | AFR | 747 | 751,291 | 747 |
| HG03115 | | 727 | 742,200 | |
| NA19240 | | 762 | 777,935 | |
| NA19440 | | 784 | 829,736 | |
| NA19921 | | 703 | 763,748 | |
| HG00733 | AMR | 657 | 687,094 | 670 |
| HG01971 | | 670 | 677,269 | |
| NA19789 | | 709 | 724,723 | |
| HG00512 | EAS | 648 | 651,655 | 665.5 |
| HG00851 | | 667 | 692,452 | |
| NA18552 | | 708 | 726,857 | |
| NA19068 | | 664 | 705,625 | |
| HG00250 | EUR | 639 | 652,081 | 639 |
| HG00353 | | 663 | 696,507 | |
| NA20587 | | 621 | 622,096 | |
| HG03838 | SAS | 635 | 662,632 | 682.5 |
| NA21125 | | 730 | 728,111 | |
| Total (non-redundant) | | 1842 | 2,107,893 | |

between Africans and East Asians but this difference is much more striking between Africans and Europeans. Principal component analysis (PCA) of NUIs shows a population-specific pattern (Fig. 1d). Specifically, PC1 clusters African samples away from other populations, while PC2 separates the East Asians from the Europeans, the South Asians, and the Americans.

**NUI origin**. To determine the origin of the NUIs, we aligned them to four different non-human primate genomes: chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), and bonobo (*Pan paniscus*). Over half of the NUIs aligned to the chimpanzee (1059; 57%), the gorilla (1017; 55%), and the bonobo (916; 50%) genomes (Table 2). Just a quarter of the NUIs aligned to the orangutan genome (498; 27%). This trend correlates well with the evolutionary divergence times inferred based on mutational profiling[17–19]. In aggregate, 1175 (64%) NUIs can be aligned to at least one primate genome and 451 (24%) are present in all four non-human primate genomes (Fig. 2a). The large number of ancestral sequences implies that the donors of the human reference genome came from human-specific lineages where these sequences were deleted after the human–chimpanzee split.

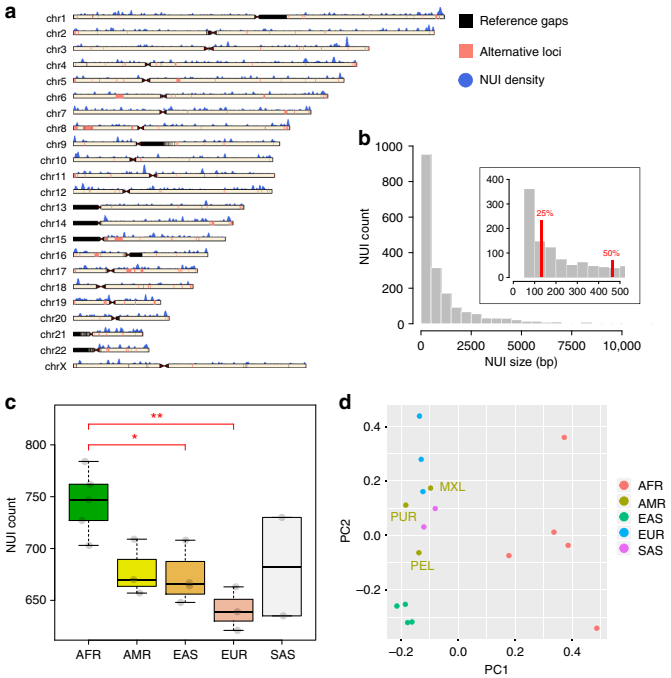



**Fig. 1** Overview of the non-reference unique insertions and their distributions across populations. **a** Ideogram depicting NUI occurrences across all chromosomes. The blue histogram above each chromosome describes the number of NUI using a sliding window of 1 Mb with a 10 kb step size. Alternative loci incorporated in hg38.p11 are shown in pink, and for display purpose, the sizes of these loci are extended by 100 kb on both sides. **b** NUI size distribution using a bin size of 500 bp in the main plot and 50 bp in the zoomed area. The 25th and 50th percentiles are labeled in red. NUIs longer than 10 kb in length were removed before plotting. **c** The number of NUIs across all five super-populations. Each gray dot shows the actual NUI number per individual. ANOVA $F(4,12) = 5.643$, $p = 0.0086$ followed by Tukey [AFR-EAS] $p = 0.0389033$; [AFR-EUR] $p = 0.0067794$. The box plot illustrates the median, the upper and lower quartiles for each population. Since no points exceed the 1.5 X interquartile range, the whiskers correspond to the minimum and maximum values in each group. ($^*p \le 0.05$; $^{**}p \le 0.01$). **d** The first two principal components based on the NUI occurrence matrix. The sub-populations of the American samples were labeled on the plot. AFR Africans, AMR Americans, EAS East Asians, EUR Europeans, SAS South Asians. American sub-populations: MXL Mexican Ancestry in Los Angeles, CA, USA; PEL Peruvian in Lima, Peru; PUR Puerto Rican in Puerto Rico

**Table 2 Non-reference unique insertions identification in non-human primates**

| | Number of aligned sequences | Aligned sequence percentage | Number of sequences found in all 5 populations | Number of sequences found in all 17 individuals |
|---|---|---|---|---|
| Chimpanzee (panTro5) | 1059 | 57% | 475 | 85 |
| Gorilla (gorGor5) | 1017 | 55% | 445 | 85 |
| Bonobo (panPan2) | 916 | 50% | 407 | 78 |
| Orangutan (ponAbe2) | 498 | 27% | 229 | 51 |
| Union of all non-human primates | 1175 | 64% | 509 | 91 |
| Total sequences in the dataset | 1842 | 100% | 584 | 95 |

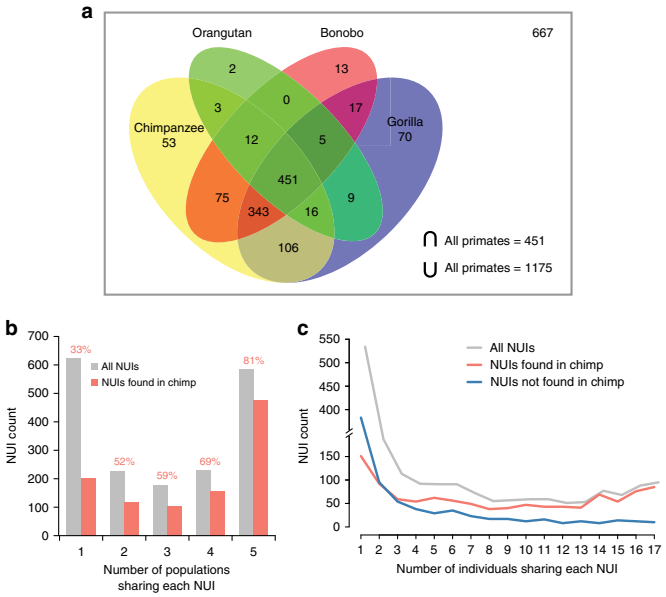



**Fig. 2** Non-reference unique insertions in non-human primates. **a** Venn diagram illustrating the number of NUIs found in non-human primate genomes. The gray box represents the universe, which includes 658 NUIs that were not found in any non-human primate genomes. **b** NUI frequency stratified based on the number of populations sharing these sequences. Frequencies are further broken down depending on whether they aligned to the chimpanzee genome. The numbers on top of the bars represent the percentages of NUIs identified in the chimpanzee genome ($\chi^2 = 309.5$, $p = 9.71e{-}66$). **c** As in **b** but at the individual-level

Next, we assessed whether the more common NUIs were enriched for ancestral sequences. Our analysis shows that the NUIs shared across all five human populations are significantly enriched in the chimpanzee genome (Fig. 2b; $\chi^2 = 310.31$, $p = 6.46e{-}66$). Remarkably, 81% of the NUIs found in all five human populations are also present in chimpanzee genome whereas only 33% of the NUIs found in a single population are present in the chimpanzee. A similar trend is observed at the individual level, where NUIs found in at least four individuals are much more likely to be found in the chimpanzee genome (Fig. 2c).

**Analysis of repeat sequences associated with NUIs**. We ran RepeatMasker[20,21] on the entire NUI call set to analyze the composition of transposable elements (TEs) (Supplementary Table 3). We found that 21.4% and 23.4%, respectively, of the overall NUIs were short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs). In contrast to the genome-wide repeat content of SINEs and LINEs (13% and 21%)[22], only SINEs were significantly enriched in this dataset. To explore the distributions of TEs, we sorted the major types of TEs into NUIs of difference sizes (Fig. 3a). NUIs under 200 bp are mostly unique sequences while longer NUIs associate mostly with SINEs, especially *Alu* elements. LINEs are the next most abundant TE found in NUIs longer than 200 bp. Long terminal repeats (LTRs) and DNA transposons are found at lower frequency in all size ranges. We additionally characterize the flanking repetitive sequences to identify potential mechanisms mediating the formation of NUIs. We found that 63% of the NUIs were flanked by a TE on at least one end (Table 3).

When aligning NUIs to the human reference genome, we observed that the two ends of the NUIs occasionally contained homologous sequences that collapsed into one overlapping copy in the reference sequence (Fig. 3b). This type of NUI account for

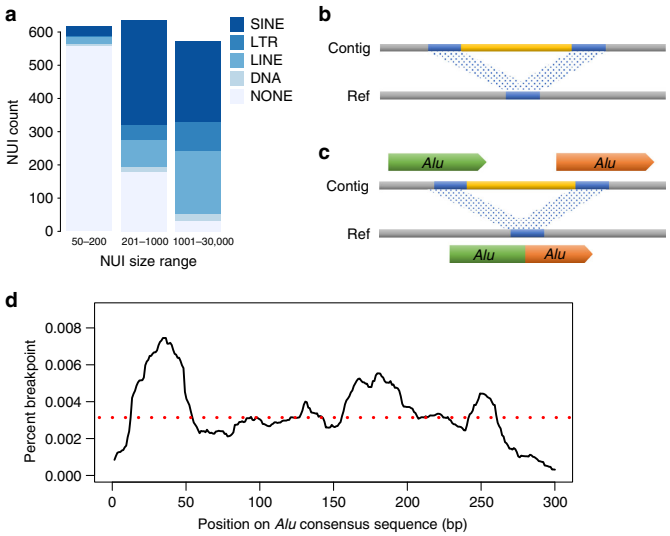



**Fig. 3** Characteristics of non-reference unique insertions. **a** Major transposable element in each NUI split by three different size ranges. Stacked bars represent the total number of NUI in each size range. SINE short interspersed nuclear element, LTR long terminal repeat, LINE long interspersed nuclear element, DNA DNA transposon, NONE no interspersed repeat detected. **b** Example of an overlapping alignment in the reference where both ends of the NUI are homologous. Yellow block, NUI; blue blocks, homologous sequences; contig, a stretch of sequence containing the NUI; ref, reference. **c** When two *Alus* recombine ectopically, the resulting alignment, as shown in the reference strand, becomes a chimeric *Alu*. **d** Percentage of breakpoint along an *Alu* consensus sequence. The red dashed line represents the average breakpoint percentage (0.0033%)

**Table 3 Distribution of transposable elements flanking all non-reference unique insertions**

| | Count | Percent |
|---|---|---|
| Alu/Alu | 336 | 18% |
| Alu/LINE | 49 | 3% |
| LINE/LINE | 105 | 6% |
| Other TE | 673 | 37% |
| Non-TE | 679 | 37% |
| Total NUIs | 1842 | 100% |

**Table 4 Distribution of transposable elements flanking non-reference unique insertions with at least 10 bp homologous sequences on both ends**

| | Count | Percent | Found in chimp | Percent in chimp |
|---|---|---|---|---|
| Alu/Alu | 265 | 52% | 163 | 62% |
| Alu/LINE | 5 | 1% | 4 | 80% |
| LINE/LINE | 13 | 3% | 4 | 31% |
| Other TE | 135 | 26% | 40 | 30% |
| Non-TE | 92 | 18% | 25 | 27% |
| Total NUIs | 510 | 100% | 236 | 46% |

at least 28% of the entire call set (510 out of 1842 NUIs have at least 10 bp overlap). Shorter overlaps are also observed but 10 bp is used as a threshold for our analysis. Using RepeatMasker on this NUI subset, we identified 418 (82% of this subset) NUIs flanked by a TE on at least one end (Table 4), with 52% flanked by *Alu* elements on both ends.

This observation, where two *Alu* elements flank many NUIs on both ends, suggests that *Alu* recombination-mediated deletion (ARMD)[23,24] is responsible for their formation. Specifically, recombination between two different *Alu* elements not in equivalent positions can give rise to ARMD and leave behind a single *Alu* chimera (Fig. 3c). In other words, the deleted version is found in the reference genome while the NUIs represent the ancestral sequences. Of the 265 candidate ARMDs identified in this dataset, 163, 167, 102, and 38, respectively, are also identified in the genomes of the chimpanzee, gorilla, bonobo, and orangutan (Table 4; Supplementary Fig. 3). The union of the ARMDs found across the four non-human primates yield a total of 195 events, which is equivalent to 74% of the total candidate ARMDs found in the NUIs. Among those, *AluSx* and *AluY* are the two most abundant *Alu* sub-species responsible for these events. This observation is congruent with the genome-wide copy number of these two *Alu* sub-family in the primate genome[25]. Moreover, we identified two putative recombination hotspots at the 12–52 and 156–205 bp positions of *Alu* elements (Fig. 3d). The first recombination hotspot is in accordance with a previously published report[23], bolstering the idea that there might be short highly conserved sequences that allow for frequent ARMDs.

**Transcription potential of NUIs**. Although 99% of the NUIs are located in the intergenic and intronic regions of the genome, there is a possibility that they are previously unannotated exons or regulatory elements with transcription potential. We used two orthogonal approaches to determine whether any of the NUIs were transcribed. First, we aligned the NUIs to the human Expressed Sequence Tag database (dbEST) and found 129 NUIs (7%) uniquely aligning to the human dbEST.

To extend this line of analysis, we also aligned the NUI call set to the high-quality RNA-Seq reads from the Geuvadis project[26].

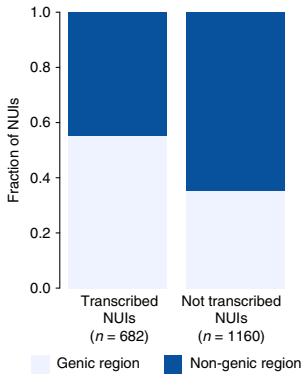



**Fig. 4** Characterization of non-reference unique insertions in the transcriptome. Stacked bars representing the proportions of NUIs that are either in the genic or non-genic regions of the genome (Fisher $p = 8.12e-17$, OR = 2.26)

We randomly selected 50 individuals representing two super-populations (Europeans and Africans) and extracted RNA-Seq reads not mapped to the human reference genome. We realigned the unmapped RNA-Seq reads to our dataset to identify NUI bearing transcribed sequences (Methods). From the Geuvadis RNA-Seq data, we identified 643 trancribed NUIs (35%), among which 90 overlapped with the result from the human dbEST. This overlap of transcribed NUIs is statistically significant (Fisher's $p = 8.018409e-17$; OR = 4.83). In total, 682 NUIs are transcribed and more than half are located in the genic region of the human genome (Fig. 4; Fisher $p = 8.12e-17$, OR = 2.26).

**Comparison with the Icelandic deCODE Genetics dataset.** Recently, deCODE Genetics/Amgen analyzed 15,219 Icelanders to identify a set of 3791 breakpoint-resolved non-reference sequences not found in the human reference genome[5]. To compare our dataset with the published results, we aligned all the NUIs to the sequences reported by deCODE Genetics in a reciprocal manner (Methods). We found that 578 of our 1842 NUIs (31%) intersected with the deCODE dataset. After filtering out the singleton NUIs, 1308 remain in our dataset and 516 (39%) are also found by deCODE Genetics. Of these 516 NUIs, 258 (50%) are present across all five populations and 463 (90%) are found in non-human primates. Given the fact that Icelanders are of British and Norwegian origin, we then assessed whether the Europeans in our study cohort shared more NUIs with this deCODE dataset than other populations. We found that among the shared NUIs, Africans had the most shared NUIs with the deCODE dataset while Europeans had the fewest (ANOVA $F(4,12) = 6.216$; $p = 0.006$; followed by Tukey [AFR-AMR] $p = 0.0349966$; [AFR-EAS] $p = 0.0462222$; [AFR-EUR] $p = 0.0058295$). This observation suggests that most of the shared NUIs are not European-specific but are mostly ancestral sequences. Additionally, the size distributions of the non-reference sequences identified by deCODE Genetics are smaller than the ones in our call set, with only 238 (6%) of the deCODE non-reference sequences >2 kb and 29 (0.8%) > 5 kb. In contrast, our call set consists of 310 (17%) NUIs > 2 kb and 73 (4%) > 5 kb.

## Discussion

In this study, we identified a large set of NUIs using phased, high-quality de novo human genome assemblies from 17 individuals with diverse genetic backgrounds. This is, to our knowledge, the most diverse set of high-quality whole genome assemblies to date using technologies that can infer long-range sequence information. In addition, the reconstruction of NUIs in this study confirms the highly variable population patterns of genomic inserts. The high abundance of NUIs found in Africans is not surprising as it is well-known that Africans have the highest level of genetic diversity among populations[27–30]. Europeans have the fewest NUIs, which is presumably due to the fact that about 70% of the human reference genome sequences came from a single donor of likely African–European admixed ancestry[4]. Despite having a small sample size for principal component analysis, the clustering pattern accurately recapitulates previously established population pattern and admixture. The fact that most of the NUIs are found across two or more populations indicates that the human reference genome has the minor alleles at these loci. The inclusion of minor alleles in the reference can interfere with every stage of variant discovery and downstream annotation[31–33]. This is particularly problematic as NUIs usually involve long stretches of DNA that differ from the reference. These missing sequences can hamper efficient read mapping. By integrating NUIs to the human reference genome, a more complete genomic scaffold can be used to annotate biological data.

Our observation that a significant subset of NUIs is flanked by *Alu* repeats and other TEs suggests that ARMD may be the mechanism by which they are deleted from the reference genome. *Alu* repeats and other TEs are among the major driving forces of genomic variation and evolution[34,35]. The over-representation of SINEs in this NUI dataset maybe attributable to the high frequency of *Alu* elements involved in ARMD, which accounts for 14% of the NUIs in this dataset. This mechanism has been described previously in the context of chimpanzee versus the human genomes[23,24]. However, this is the first time such ARMDs are assessed genome-wide within the human population. Here, we have shown that ARMDs can be polymorphic between individuals or populations, and some events classified as "human-specific" are in fact highly variable even within the human populations.

In identifying transcription potential of NUIs, we found a subset of sequences aligning to either the dbEST or reads from RNA-Seq experiments, suggesting that some of the NUIs contribute directly to transcriptomic diversity that has not been accounted for in the past. These sequences may include previously unannotated exons or non-coding regulatory sequences that can alter the rate of transcription of targeted genes. The inclusion of these NUIs is essential to our understanding of the human transcriptome and the transcriptional regulation of non-coding sequences. It is also important to note that only RNA-Seq from two populations were used in the analysis. The number of transcribed NUIs would certainly increase when data from more individuals or additional populations are also analyzed.

Finally, when we evaluated the non-reference sequences reported by deCODE Genetics/Amgen[5], we found that 39% of the more common NUIs were also present in Icelanders. Even with the singletons removed, 516 NUIs (89% of the initial 578 shared NUIs) still remained. This result suggests that the sequences reported by deCODE Genetics are not exclusive to Icelanders. However, since the genome of the Icelanders is relatively homogenous[36], we expect to see a small proportion of genomic alterations that are unique to them due to an elevated level of genetic drift and the potential for founder effect[37].

Overall, our diverse dataset allows us to perform a cross-population survey of NUIs that are typically difficult to detect when one relies only on short-read sequencing data, thereby expanding the NUI catalog and sequences specific to populations that are often underrepresented. The current human reference genome represents a single composite haplotype at any given locus. In this linear form, and even with the inclusion of a limited set of alternate haplotypes, it cannot capture much of the genetic

diversity across the different continental groups. Several studies have already demonstrated that the use of an ethnicity-matched reference allows for increased accuracy in imputation, and ultimately, disease susceptibility prediction[38–41]. It is therefore critical to produce a set of human reference genomes that includes NUIs from many populations in order to depict the unbiased landscape and diversity of the human genome.

## Methods

**Collection of 10×G Linked-Read data**. Linked-Read data for 14 samples were generated on site using cell lines purchased from the Coriell Institute. Another three samples, namely HG00512, HG00733, and NA19240, were obtained from the 10×G website in the format of FASTA and FASTQ files (https://support.10xgenomics.com/de-novo-assembly/datasets/1.1.0/msHG00512; https://support.10xgenomics.com/de-novo-assembly/datasets/1.1.0/msHG00733; https://support.10xgenomics.com/de-novo-assembly/datasets/1.1.0/msNA19240). The FASTA files corresponded to the pseudo-haplotypes that were generated with Supernova v1.1 while the FASTQ files were downloaded so that we could generate the alignment BAM files using Long Ranger v2.1 in-house.

**Collection of unaligned/poorly aligned read pairs**. The NUI discovery pipeline initially accepted a BAM file generated from 10×G Long Ranger v2.1 and identified reads that did not align well to the human reference genome (core hg38). This was defined as reads fulfilling at least one of the following criteria:

- Reads with an unaligned SAM flag, and their mates
- Read pairs not mapped within insert sizes (BWA-MEM[42] estimated the insert size based on the bulk read pair distributions and this information was used by the Long Ranger alignment software)
- Read pairs with wrong read orientations
- Reads with an alignment score ≤−80, and their mates (this corresponds to a minimum of 40 mismatches or a combination of other penalties according to the Lariat scoring parameters: AS:f = −2 * mismatches −3 * indels −5 * clipped −0.5 * clipped_bases −4 * improper_pair)
- Reads with more than 40 bases clipped off, and their mates

Samblaster v0.1.24[43] was used to extract reads with clipped and unaligned reads while sambamba v0.5.9[44] and samtools v1.2[45] were used to collect other poorly aligned reads.

The raw FASTQs of this collection of unaligned/poorly aligned read pairs were extracted using seqtk v.1.0[46] and processed by trimming off the first 23 bp of the first mate of each pair to remove the 16 bp 10×G barcode and the 7 bp low accuracy sequence from an N-mer oligo. This collection of reads was then filtered based on their base qualities using the fastq_quality_filter utility provided in the FASTX Toolkit v0.0.14[47]. The entire reads, along with their mates, were removed if less than 70% of their bases had a quality score of 30 or above.

**Alignment of unmapped reads to 10×G pseudo-haplotypes**. BWA-MEM[42] paired-end mode was used initially to align all the unmapped/poorly mapped read pairs to the hg38 reference genome and the sample-matched pseudo-haplotypes generated from 10×G Supernova v1.1. Starting from this alignment step onward, all the procedures were repeated for both pseudo-haplotypes. After read alignment, read pairs that mapped well to the hg38 reference genome were discarded from the other two alignment outputs, which were further filtered based on the following stringent criteria:

- Read1 with an alignment score of at least 90
- Read2 with an alignment score of at least 113 (read1 was shorter due to trimming of the barcode and the N-mer oligo)
- Read pairs in the proper orientation
- Read pairs mapped within insert size
- Mapping quality was at least 30

The alignment scores used in this analysis (90 and 113) were determined based on the BWA-MEM alignment scoring scheme. Each mismatch gets assigned a penalty of −4, and hence 90 and 113 roughly correspond to 9 mismatches in the sequence alignment (or a combination of penalties).

**Identification of read clusters**. Keeping only reads that aligned well to the pseudo-haplotypes, the read coverage at every position was calculated using bedtools[48]. Read clusters with coverages between 8 and 100 were located. The pseudo-haplotype sequences corresponding to the read clusters were extended by 7000 bp on each end, or until the ends of the assembled sequences, to serve as the alignment anchors for the following step. If the ends between two different contigs were overlapping or separated by less than 200 bp, the entire sequence from the upstream left anchor to the downstream right anchor was extracted for Lastz[49] alignment. Contigs with more than 10 N's were removed from downstream analysis to ensure high accuracy.

**Breakpoint computation**. Extended contigs were aligned to the hg38 core reference genome using Lastz with the following parameters: –step = 20 –seed = match15 –notransition –exact = 400 –identity = 99 –match = 1,5. We then computed the precise breakpoint in each contig by locating where the sequence alignment broke off and realigned. Ideally, one contig should produce two alignments as each anchor should align uniquely to the reference genome, separated by one or two breakpoints. This is, however, not always the case. When part of an anchor aligned to multiple places on the reference, we selected the one with the longest alignment and the highest sequence identity. If an anchor aligned to more than five genomic loci, we ensured that the anchor alignment was at least 3500 bp in length before choosing the best alignment candidate. When the two anchors from the same contig partially overlapped on the reference, we filtered them out if the overlapping sequence was larger than 800 bp in size to ensure the accuracy of our call set. Alignments that created an overlap of 800 bp had poor concordance rate with the BioNano insertion calls. Even when BioNano makes an insertion call at the genomic locus overlapping an NUI candidate, the median size difference between BioNano-predicted size and the length of the NUI was about 3 kb. In contrast, the median size difference for all other alignments were usually between 300 and 600 bp. While BioNano cannot determine the precise insertion size due to its inherent resolution limits, this difference suggested that the sequences with larger overlap were not as reliable. To ensure the high quality of our call set, these sequences were discarded prior to the analysis. At this point, only insertional sequences with gap size ≥ 50 bp were kept for downstream analysis. Contigs whose ends were immediately flanked by N-gaps were discarded. Output from the two pseudo-haplotypes were combined at this point. A unified list was generated by combining homozygous contigs, that was, if their breakpoints were less than 10 bp apart on each side. The unaligned breakpoint-to-breakpoint sequences of all of the contigs were extracted for further analysis.

**Definition of NUI variants**. To determine whether these contigs fulfilled the definition of an NUI, we ran RepeatMasker v4.0.7 and dustmasker[50] to determine the extent of interspersed repeats and low complexity sequences in each contig. RepeatMasker was run with -species human and dustmasker was run with the default settings. The number of unmasked base pairs was counted for each contig and sequences with ≥ 50 unique bases were kept. To ensure that these contigs were not included in the human reference genome including all the alternate haplotypes, fix patches, and novel patches, we used BLAST to align these NUIs—extended by 50 bp on both ends—to the hg38.p11 human reference genome. NUIs aligning to hg38.p11 with ≥ 95% identity and 100% coverage were removed from the call set. We also removed NUIs resulting from translocation events. To identify translocated sequences, we used BLAST to align NUIs—breakpoint-to-breakpoint—to the human reference genome with –task megablast and –dust no. Alignments with ≥ 95% identity and 100% coverage were removed from the call set. To reduce false-positive calls, we filtered out NUIs whose breakpoints overlap assembly gaps, segmental duplications, and other problematic regions as defined by the 10×G SV filter criteria: https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/sv-blacklist. Two files: sv_blacklist.bed and segdups.bedpe were used for this filtering step. Finally, we merged all the NUIs across 17 individuals to make a unified, non-redundant call set by collapsing NUIs if the inserted sequences shared one identical breakpoint with another sequence or if both breakpoints (start and end) were within 50 bp of those from another sequence. The sequences in FASTA format and the NUI occurrence matrix can be found in Supplementary Data 1 and 2, respectively. The reported NUI occurrence matrix is encoded as 0, 1, and 2 (2 meaning the individual harbors the NUI in both pseudo-haplotypes). However, since the homozygous calls were not definitive, all the NUIs with a genotype of 2 were recoded to 1 to ensure high data quality. The recoded binary matrix was used for the rest of the analysis. We believe that some of the homozygous calls were incorrect due to the observations that many of the NUI singletons are homozygous calls rather than heterozygous calls. In these cases, the Supernova assembler places the NUIs in both pseudo-haplotypes inappropriately.

**NUI validation using BioNano optical maps**. We used BioNano optical maps insertion calls to validate our NUI call set. Our SV calling strategies involved two pipelines: BioNano pipeline 4618/4555 and a modified version of OMSV[16]. Insertions called by either algorithm are merged together for downstream analysis. To validate, we first identified a list of NUIs that were greater than 2 kb in length. SVs smaller than 2 kb are prone to sizing error caused by DNA fragments that are either not completely linearized or over-stretched.

Next, we overlapped NUIs with the individual-matched optical maps and filtered out insertions where the optical maps had zero-coverage, or if they were within 10 kb of these zero-coverage regions. Finally, we filtered out NUIs if another SV was reported within 10 kb, including deletions, complex inversions, or simple insertions that did not fulfill the definitions of NUIs. Exclusion of these NUIs were necessary since the BioNano optical maps only show the overall size change between two labels. In other words, BioNano SV calling is error-prone if multiple SVs occur in tandem.

Taking this set of filtered NUIs, we computed the precision rate for each individual. The precision rate was calculated as follows:

$$\text{precision} = \frac{\text{NUIs supported by BioNano}}{\text{NUIs supported by BioNano} + \text{NUIs not supported by BioNano}}$$

**Contamination screen**. We did not expect sequencing contaminants from bacteria, virus, plants, and yeast to be present in this NUI call set because they should not have strong anchor to the human reference genome. Additionally, contaminants would not have significant barcode sharing with nearby sequences, and hence they would not form long contigs. To verify these assumptions, we used BLAST to align NUIs to the bacteria database (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/), the human microbiome database (ftp://ftp.ncbi.nih.gov/genomes/HUMAN_MICROBIOM/Bacteria/all.fna.tar.gz), the univec database (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/UniVec), and the virus database (ftp://ftp.ncbi.nih.gov/genomes/Viruses/all.fna.tar.gz). None of the NUIs produced significant alignment with these common contaminants using 90% identity and 90% query coverage filter thresholds.

**Principal component analysis**. The NUI occurrence matrix was used as the input for the principal component analysis. To clean up the matrix, NUIs with 1, 2, and 17 occurrences were removed from the matrix before analysis. NUIs with 1 or 2 occurrences might increase ambiguity in the dataset while NUIs with 17 occurrences (found in every individual) had no variance. Overall, 1026 NUIs remained in the matrix for analysis.

**Aligning NUIs to non-human primates**. Four non-human primates were used in this study and their reference genomes were downloaded from either UCSC or NCBI. The specific versions used in this study were chimpanzee (panTro5), gorilla (gorGor5), bonobo (panPan2), and orangutan (ponAbe2). We used BLAST to align the NUIs, including 50 bp flanking sequences, against each of these non-human primate genomes using -task megablast and -dust no. Only sequences that aligned with at least 95% identity and 95% query coverage were considered as real hits.

**Identifying the major transposable element in NUIs**. To identify the major TE in each NUI, we ran RepeatMasker individually to compute the sequence composition for each TE. For each entry, the TE that makes up the highest percentage of that sequence was deemed the major TE. RepeatMasker was run with –species human, -xsmall, and -nolow.

**Determining transposable elements flanking NUIs**. To determine whether the two ends of the NUIs were flanked by TE, we extracted 300 bp upstream and 300 bp downstream of each end and ran RepeatMasker to determine the composition of these sequences. We specifically filtered for TEs containing the NUI breakpoints to determine the potential mechanisms mediating these insertions.

**Identifying breakpoint frequency in an *Alu* element**. We assessed the breakpoint frequency along an *Alu* element consensus sequence to identify potential hotspots for the ARMD. First, we identified *Alu*-flanking NUIs that shared homologous sequences on both ends. These sequences form overlapping alignments on the reference and 10 bp sequence homology was required. Based on the RepeatMasker output from before, we determined the position of the *Alu* element corresponding to the breakpoint, and we subtracted the length of overlap from that breakpoint position to obtain a breakpoint range for that particular NUI. Since the breakpoints are ambiguous over a range of positions, only $\frac{1}{\text{length of range}}$ was added per each position to compute the breakpoint frequency.

**ARMD size distributions**. Similar to the last section, we first extracted all the NUIs flanked by *Alus* on both ends and we further filtered to keep the ones where the two *Alus* came together to form a single *Alu* chimera on the reference. This final set of NUIs was used to create Supplementary Fig. 2.

**Aligning NUIs to the expressed sequence tag database**. We used BLAST to align the NUIs to the human dbEST (ftp://ftp.ncbi.nlm.nih.gov/blast/db/est_human.[number].tar.gz). Regions with 95% sequence identity, regardless of the query coverage, were extracted for further assessment. We adjusted the filter criteria here because entries in the dbEST are short and the sequences usually only represent the ends of expressed genes. Next, the regions that aligned to the dbEST were realigned to hg38 with BLAST, to ensure these regions do not align to anywhere else on the genome. Any of these regions that aligned to the reference genome were discarded. Otherwise, these sequences were considered to be transcribed.

**Aligning RNA sequencing reads to the NUI call set**. We gathered RNA-sequencing reads from the Geuvadis project[26] and the raw sequencing data is publicly available. We randomly selected 50 individuals for the analysis—10 from each sub-population included in the Geuvadis project (CEU-EUR, FIN-EUR, GBR-EUR, TSI-EUR, and YRI-AFR). Raw FASTQ files for the RNA-Seq reads were downloaded from https://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-3/. To align RNA-Seq reads, we first generated the genome index using the hg38 primary reference genome and the GTF file downloaded from the ensembl website (ftp://ftp.ensembl.org/pub/release-92/gtf/homo_sapiens/Homo_sapiens.GRCh38.91.gtf.gz) with the following parameters: STAR –runThreadN 32 –runMode genomeGenerate –genomeDir /path/to/STAR_genome –genomeFastaFiles /path/to/hg38/fa –s–jdbGTFfile /path/to/emsembl_gtf –sjdbOverhang 74. We aligned all RNA-Seq reads to this indexed genome, one sample at a time, as follows: STAR –runThreadN 32 –outReadsUnmapped Fastx –genomeDir /path/to/STAR_genome –outFileNamePrefix /out/path/prefix –outFilterMultimapNmax 100000 –outSAMunmapped Within KeepPairs –limitOutSAMoneReadBytes 1000000 –readFilesIn read_1.fastq read_2.fastq. By default, STAR does not output unmapped reads in the SAM alignment file. This feature had to be turned on with –outSAMunmapped and –outReadsUnmapped. The second parameter automatically wrote all the unmapped FASTQs into separate files that could be used directly for the subsequent steps. Since STAR aligner considers reads unmapped if they are mapped to more than 20 loci by default, this feature was turned off by setting this threshold to 100000 using –outFilterMultimapNmax. We then collected all of the unmapped reads and their mates from the 50 individuals. Specifically, unmapped reads were defined by the SAM flag 4. We merged the FASTQ files from all samples into a single read1 file and a single read2 file.

Since RNA-Seq aligners behave differently when aligning to a reference with and without the GTF annotations[51], we realigned the collection of unmapped reads to the human reference genome without supplying the GTF file. All unmapped reads from this second alignment step were collected for the final alignment procedure, which involved mapping these reads to the NUI call set in which every NUI was extended by 300 bp on both ends. The NUI extension step was essential to facilitate reliable alignments since some of the NUIs were too short for mapping. The genome index for the NUI was generated using the same parameter described above, except we did not provide a GTF file. The unmapped reads were passed on to the STAR aligner using the following command: STAR –runThreadN 32 –genomeDir /path/to/ STAR_genome_NUI –outFileNamePrefix /out/path/prefix –readFilesIn unmapped_read1.fq unmapped_read2.fq. Reads were then filtered based on the following criteria:

- Both reads in a read pair were mapped
- Both reads in a read pair could not fully reside in the 300 bp flanking sequences
- Read pairs mapped in correct orientation and with correct insert size
- Reads with alignment scores ≥ 140 (this is equivalent to four total mismatches for a read pair)
- Reads with a mapping quality score of 255, which is indicative of unique mapping based on the STAR scoring scheme

NUIs aligning to at least any one of these filtered read pairs were considered to be transcribed.

**Comparison with the deCODE genetics (Icelandic) dataset**. The non-reference sequences generated by deCODE Genetics/Amgen were obtained from supplementary data 1 of the publication[5]. We first aligned all the NUIs to the Icelandic contigs using BLAST. Alignments with 95% sequence identity and 95% query coverage were considered as real hits. Next, we reciprocally aligned the Icelandic contigs to the NUIs also with BLAST, and the alignment resu7lts were filtered using the same criterial described above. Results from the two alignment steps were merged and a non-redundant list of NUIs was used for the reported statistics. Reciprocal sequence alignment as performed since the reported breakpoints between the two pipelines might not be consistent. After the initial analysis, we also filtered out uncommon NUIs from our NUI call set. Namely, NUIs with one occurrence in the study cohort were discarded. These sequences are likely to be population-specific and we would not expect them to be shared with the Icelandic genomes.

**Code availability**. The NUI pipeline source code can be accessed via GitHub (https://github.com/wongkarenhy/NUI_pipeline.git).

**Data availability**. The 10×G read alignments (BAMs), the Supernova assemblies (FASTAs), and the BioNano assemblies (CMAPs) can be accessed via NCBI BioProjects PRJNA418343 and PRJNA435626. All NUIs were deposited to NCBI GenBank nucleotide database with accession numbers MH533022-MH534863. Three additional samples, HG00512, HG00733, NA19240, were obtained from the 10×Genomics website. The Geuvadis RNA sequencing data set was obtained from accession number E-GEUV-3.

Received: 16 March 2018 Accepted: 11 July 2018
Published online: 02 August 2018

## References

1. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
3. Hehir-Kwa, J. Y. et al. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7**, 12989 (2016).

4. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
5. Kehr, B. et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nat. Genet.* **49**, 588–593 (2017).
6. Telenti, A. et al. Deep sequencing of 10,000 human genomes. *Proc. Natl Acad. Sci. USA* **113**, 11901–11906 (2016).
7. Iafrate, A. J. et al. Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
8. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
9. Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
10. Chaisson, M. J. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
11. Mills, R. E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
12. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
13. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
14. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
15. Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. dbEST–database for "expressed sequence tags". *Nat. Genet.* **4**, 332–333 (1993).
16. Li, L. et al. OMSV enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. *Genome Biol.* **18**, 230 (2017).
17. Prüfer, K. et al. The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527–531 (2012).
18. Consortium, C. Sa A. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
19. Scally, A. et al. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012).
20. Smit, A., Hubley, R., Green, P. RepeatMasker Open-4.0. http://www.repeatmasker.org/ (2015).
21. Hubley, R. et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
22. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
23. Han, K. et al. Alu recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genet.* **3**, 1939–1949 (2007).
24. Sen, S. K. et al. Human genomic deletions mediated by recombination between Alu elements. *Am. J. Hum. Genet.* **79**, 41–53 (2006).
25. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**, 370–379 (2002).
26. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
27. Tishkoff, S. A. & Williams, S. M. Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* **3**, 611–621 (2002).
28. Tishkoff, S. A. & Verrelli, B. C. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genom. Hum. Genet.* **4**, 293–340 (2003).
29. Hammer, M. F. et al. Hierarchical patterns of global human Y-chromosome diversity. *Mol. Biol. Evol.* **18**, 1189–1203 (2001).
30. Tishkoff, S. A. et al. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387 (1996).
31. Barbitoff, Y. A. et al. Catching hidden variation: systematic correction of reference minor allele annotation in clinical variant calling. *Genet. Med.* **20**, 360–364 (2017).
32. Dewey, F. E. et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS Genet.* **7**, e1002280 (2011).
33. Magi, A. et al. Characterization and identification of hidden rare variants in the human genome. *BMC Genom.* **16**, 340 (2015).
34. Britten, R. J. Transposable element insertions have strongly affected human evolution. *Proc. Natl Acad. Sci. USA* **107**, 19945–19948 (2010).
35. Hedges, D. J. & Belancio, V. P. Restless genomes humans as a model organism for understanding host-retrotransposable element dynamics. *Adv. Genet.* **73**, 219–262 (2011).
36. Helgason, A., Sigureth ardóttir, S., Gulcher, J. R., Ward, R. & Stefánsson, K. mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am. J. Hum. Genet.* **66**, 999–1016 (2000).
37. Helgason, A., Nicholson, G., Stefánsson, K. & Donnelly, P. A reassessment of genetic diversity in Icelanders: strong evidence from multiple loci for relative homogeneity caused by genetic drift. *Ann. Hum. Genet.* **67**, 281–297 (2003).
38. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
39. van Leeuwen, E. M. et al. Genome of The Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nat. Commun.* **6**, 6065 (2015).
40. Nagasaki, M. et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**, 8018 (2015).
41. Martin, A. R. et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
42. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
43. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
44. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
45. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. Li, H. Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences. https://github.com/lh3/seqtk/ (2013).
47. Hannon lab. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/ (2009).
48. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
49. Harris, R. S. *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, Pennsylvania State University (2007).
50. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.* **13**, 1028–1040 (2006).
51. Engström, P. G. et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10**, 1185–1191 (2013).


## Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health (NIH) under award R01 HG005946 to P.-Y.K. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Yulia Mostovoy for her intellectual contributions to this work, Le Li for sharing the SV call set generated from the BioNano optical maps, and Ernest Lam for providing technical guidance for the optical mapping analysis.

## Author contributions

Pipeline development: K.H.Y.W. Data analysis: K.H.Y.W. and M.L.-S. Manuscript writing: K.H.Y.W., M.L.-S., and P.-Y.K.

## Additional information

**Supplementary Information** accompanies this paper at https://doi.org/10.1038/s41467-018-05513-w.

**Competing interests:** The authors declare no competing interests.

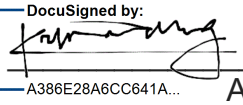
**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Publishing Agreement**

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

A386E28A6CC641A...
Author Signature

3/16/2020
Date