Genome-wide variation in human germline and postzygotic mutation rates

Michelle Noyes

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Evan E. Eichler, Chair
Philip Green
Kelley Harris

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Genome-wide variation in human germline and postzygotic mutation rates

Michelle Noyes

Chair of the Supervisory Committee:
Evan E. Eichler
Genome Sciences

De novo mutations (DNMs) are new variants that arise in the parental germline or early embryo. In this dissertation, I apply long-read sequencing technology to quads and a multi-generational pedigree to discover DNMs across the genome and quantify the *de novo* mutation rate. First, I demonstrate that long reads enable DNM discovery in previously inaccessible regions of the genome. These newly accessible regions, largely marked by repetitive sequence, have a significantly higher mutation rate than their unique counterparts, including an approximately 66% enrichment in segmental duplications. I was able to trace the origins of DNMs to either the parental germline or early rounds of embryogenesis, revealing that at least 15% of single nucleotide DNMs arise postzygotically, a 50% increase from earlier studies. Further, I found that 60% of postzygotic mutations are transmitted to the next generation, meaning that they contribute to segregating variation in the population. Finally, I estimate the *de novo* mutation rate to be approximately 1.2-1.3$\times 10^{-8}$ substitutions/base pair/generation for 30 year old parents, and the postzygotic mutation rate to be approximately 0.23×10^{-8} substitutions/base pair/generation. My analyses reveal that repetitive regions are in fact hypermutable, and that more variation arises postzygotically than previously thought. This work also lays the foundation for the next frontier in DNM discovery: comparing assembled parent and child genomes to reveal variation in the most complex and mutable parts of the genome.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

As I sit to write this, I'm watching my cats, Taki and Meanie, play together, fighting over a toy bird that they'll abandon as soon as one of them claims ownership. I had cats growing up, and I loved them a lot, don't get me wrong, but the way I felt about them pales in comparison to rush of dopamine I get when I see Taki figure out a new puzzle feeder or Meanie take careful aim and jump to a height I thought impossible.

When I reflect on my time in graduate school, I have so much to be proud of. I grew up, learned how to code (more or less), gave a talk at a big conference, wrote some papers, and survived a pandemic. But my proudest accomplishment? It's building and maintaining a small community that I feel so grateful to be a part of. In a time where it feels like the world is becoming more cynical, the challenges of the last six years have instead made me softer and more big-hearted. Thank you to everyone I've had the honor of loving. Thank you for supporting me and laughing at my jokes (even the bad ones) and listening to me talk about my interest-du-jour even though I know you couldn't care less. Words can't even begin to express the gratitude I feel, but I'll sure give it a try.

To my advisor, Evan: thank you for never giving up on me, for being quick with praise when I did something well, for never hesitating to tell me when I was doing a bad job, for teaching me to think like a scientist, for beers at the College Inn, excellent scientific mentorship, fun stories, and for putting together a lab full of so many wonderful and talented people. Working with you has been a privilege, and I'm not sure anyone else could have gotten me to the finish line. I'm sorry I told you I thought science was boring - I promise it's cool when you talk about it!

To my thesis committee, Kelley Harris, Phil Green, Sharon Browning, Lea Starita, and Debbie Nickerson: thank you for all the time and care you have invested into me and my projects. Your feedback has significantly improved the quality of my work; it has been an honor to receive your wisdom and a pleasure to get to know you.

To the Eichler lab: thank you for your scientific and emotional support. Kendra and Katy, you are both rock stars, and I think we'd all fall apart without you. Thank you for putting up with my questions and always having all the answers. Tonia, you make the world go around. I am infinitely grateful for every paragraph edited, meeting scheduled, form filled out, and everything in between. And to the other students, Mitchell, Phil, Xavi, Taylor, and Lizzie: there isn't anyone else I would have rather shared my days with. Thank you so much for letting me bother you, and I'm sorry for any extra time I may have added to your PhDs. I'm looking forward to the next time we all get together for a beer.

To my cohort: one day we'll win a round of trivia, I'm sure. Is it strange I feel nostalgic for first year, when we edited each other's papers, learned C++, and tried to understand statistics together? I'm so proud of all you've accomplished, and grateful that you stuck with me through all the growing up I had to do.

To my STEMinists: was it fate? I wouldn't say I'm someone with particularly good luck, but living with you was winning the lottery. When we all get together I call it a family reunion, and I feel so incredibly proud seeing us with our pets and our partners and our fancy grad school degrees that we all worked so hard for. Thank you for inspiring me. I can't wait to see what else life has in store for us.

To the only two people from high school that matter: you are my family. You taught me how to be a friend, and though life has put us on very different paths, I feel closer to you than ever. Thank you for being the ultimate hype guys, for making me laugh, for your thoughtful takes that I never would have considered, and for your unwavering support.

To the best part of Talbot House: you suffered through Chicago winters and academic

rigor with me. I'm not exaggerating when I say I would not be here today without you. You have been there for me through so many of my lowest moments, bad haircuts, and wrong decisions, and you've always made me feel so loved. I'm so grateful for every gossip session, movie marathon, intricate meal, and bundt we've had together. I don't call you nearly enough.

To my household, my quarantine family, my co-conspirators, my boyfriend's friends: what an unbelievable journey we've been on together. When we first met, I could have never imagined the friendships we would form. Living with you during the pandemic was a gift. You have changed my life in more ways than you know, and you have been a constant source of joy, even when grad school was really getting me down. Thank you for making space for me in your homes and your lives.

To all the cats I've gotten to know during grad school: meow meow *meow* meow meow meow. Meow *meow meow meow* meow meow mewwwww. Meow meow!

To my parents: this is your achievement. This journey started for me at grandma and grandpa's house, reading Dr. Seuss books. At the dining table, playing with molecule kits. You have given me everything and more, and I hope I've made you proud, even if I'm not the right kind of doctor. I can never truly thank you enough for always supporting and believing in me, encouraging me to try my best, and being there for me when my best wasn't quite good enough. I love you so much.

And finally, to Ryan: you're watching me cry as I write this. You should get an honorary PhD or a medal or a presidential commendation for your support. I am grateful for every problem you've helped me debug, every slide you've looked at, every meal you've cooked while I'm frantically trying to finish something I procrastinated, every moment we've spent together. And, most importantly, I'm grateful that you picked out the two cutest kitties in the whole wide world. Thank you for loving me unconditionally.

DEDICATION

To my parents, Daniela and Jason, and my grandparents, Irene and Arthur.

Chapter 1
INTRODUCTION

Imagine a mutation. What does it look like? Maybe it's the fictional super variant that gave the Teenage Mutant Ninja Turtles their powers and proclivity for pizza. Perhaps you're imagining an albino alligator you saw at the zoo, a tabby with six toes, a giraffe with no spots. Maybe you worked in a fly lab and you're picturing a black fly with legs where its eyes should be. Personally, I think about genetic diseases that children do not share with their parents, like Down syndrome.

With the exception of the Ninja Turtles, who were tweens when they got their powers from a mutagen [Turtlepedia], these mutations are all *de novo*: the result of some change unique to that specific individual. In other words, the individual has a *de novo* mutation (DNM)—any genetic variant that appears in an offspring but not in any of their ancestors. DNMs don't always stay unique, as they can be passed down to the next generation. This process of mutation and inheritance drives evolution: we can thank DNMs for our thumbs, brains, and lack of tails. However, we can also blame DNMs for miscarriages, genetic diseases, and the occasional extra digit or two. The dual ability of mutation to both create and fix problems on an individual and generational scale colors many disciplines of biology.

More than that, the story of mutation is the story of humans. To characterize how we change is to understand both where we come from and where we're going.

That's a little lofty, so let's get into the nitty gritty.

1.1 We're all mutants

So far, I've mostly described mutations with big, visible effects, but the majority of mutations are nearly neutral [74, 133]. Every person is born with approximately 60-80 *de novo*

mutations, with an average of three or fewer of those mutations occurring in protein-coding sequences [149, 11, 50, 134, 157]. While disease-causing mutations are subject to purifying selection, neutral mutations don't face selective pressure and can spread through a population. As they accumulate over time and become segregating variants, *de novo* variants act as the ticks of a molecular clock [199]. The number of neutral differences between the genomes of two closely related species, such as humans and chimpanzees, can suggest how long ago their ancestors diverged [89]. However, because mutation rates are under evolution of their own [10] and vary even between different families in the same population [162], neutral variation by itself cannot fully describe divergence time. So, while DNMs are integral to population genetics, many studies of *de novo* variation are instead nestled in the context of human disease, and my work is no exception.

De novo mutations can cause any number of diseases [147, 195] but go looking for research on DNMs and you'll find many studies on neurodevelopmental disorders, and in particular, autism spectrum disorder (ASD). The link between genetics and ASD was established by Folstein and Rutter in 1977, when they found higher rates of concordant autism in monozygotic (identical) twins than their dizygotic (fraternal) counterparts [49]. Less than two decades later, the first *de novo* mutations were implicated in ASD. These were large structural variants and copy-number changes, both of which were easier to detect with karyotyping and microarray technologies that were common at the time [114, 164]. As whole genome sequencing became affordable at scale, large cohort studies found that DNMs were responsible for approximately 10-30% of simplex ASD cases; [76, 128, 134, 176] in females with ASD that number increased to 45% [76]. Many of these causal DNMs were small gene-disrupting events, like nonsynonymous single nucleotide substitutions (SNVs) or frameshift insertions or deletions (indels), and as our understanding of noncoding variation grows, it is likely that our estimate of the impact of DNMs on ASD will only increase [87, 197].

Still, from a human health perspective, the importance of studying DNMs might not be immediately evident. New mutations are all but guaranteed to occur in every individual, and no scientific intervention can prevent the bad luck of a DNM disrupting a gene or can

repair disease-causing mutations after they've occurred.[1] But the purpose of DNM research is not only preventative: identifying specific mutations can change the lives of both a patient and their relatives. For one, it can inform treatment of the disease or reveal comorbidities that also require medical intervention [58]. In addition, if the disease is *de novo*, parents of the patient can feel absolved knowing there was nothing they could have done to prevent it and comforted that there is no chance of having another child with the same mutation [70]. What's more, families can, and do, build communities specific not just to a disease like ASD, but to the specific mutations responsible for it. *De novo* mutations aren't just a window into understanding our species: they can shape an individual's perception of themself.

Whoops, I'm getting lofty again. I hope you're ready for some biology now.

1.2 On the origin of de novo mutations

When I was taught genetics, there were two distinct classes of mutations to contend with, defined by when and where they occurred. A mutation could be *de novo*, meaning it arose in the parental germline and could be found in a sperm or egg cell before they ever joined forces to create an embryo, or it could be somatic, originating at any point after that sperm and egg cell met, just like that mutation that gave the Ninja Turtles their powers.

I, however, am interested in a third class of mutations that is sandwiched right in between germline *de novo* and somatic events and shares properties with both of them. This class is made up of mutations that arise in the first rounds of embryonic development, right after the sperm and egg come together. These mutations can go by several different names, like gonosomal [162] or early embryonic [177], but because they arise after fertilization creates a zygote, I will side with the studies that refer to them as postzygotic mutations, or PZMs [1, 36, 53, 107]. I also happen to think it's the snappiest name.

Historically, any mutation that occurred sometime after fertilization but before birth was referred to as a PZM [16, 77, 82, 184], but over time the definition has become even less pre-

[1] Well, we can't do anything to fix mutations at the moment. I'm hoping that statement will age poorly, in light of CRISPR and the promises of precision medicine.

cise, and today many studies actually use the terms somatic and postzygotic interchangeably [159, 188, 191]. Although both can definitionally refer to any mutations that occur after conception, I think there is good reason to differentiate between somatic mutations that occur say, in skin cells on a 45-year-old adult, and PZMs that occur early in development, during the first few rounds of cell replication that make a blastocyst and eventually a fetus. From a basic biology perspective, because PZMs occur in the unique cellular conditions of the newly fertilized embryo, they likely reflect mutational biases distinct from later developmental stages or adult tissues [1, 162, 96]. From an evolutionary perspective, somatic mutations are confined to the individual in which they occur, but it is possible for PZMs to be passed on to the next generation. I expect that most PZMs occur early enough in development to be present in the germline, meaning they can be transmitted to offspring and become segregating variants in a population, just like *de novo* mutations.

Before we get ahead of ourselves, it's important to first understand how mutations arise in the first place. Both DNMs and PZMs arise through errors in DNA replication and damage repair, but there are some quirks unique to each class of mutation.

1.2.1 Germline mutations

Let's use spermatogenesis and oogenesis to examine the mutagenic effects of DNA replication and damage repair. Approximately three quarters of all germline DNMs occur in the paternal germline [61, 91, 162], likely as the result of errors in DNA replication. During replication, DNA polymerase can insert the wrong base, creating a mismatch, or slip, resulting in an insertion or deletion (Figure 1.1A). These mistakes should be corrected before the next cell division, but some inevitably fall through the cracks, especially if they occur in regions of the genome that replicate later in S phase [37]. All female germ cell progenitors undergo exactly 24 rounds of DNA replication, while male progenitors replicate 30 times before puberty alone [27]. After puberty, male progenitors replicate about once every sixteen days, which can add up to hundreds of rounds of DNA replication before it ever reaches the egg [118]. These repeated rounds of replication were long believed to cause the paternal age effect, an increase

of 1-2 paternally-derived DNMs per additional year of paternal age [50, 91, 118]. However, mounting evidence suggests that the paternal age effect may result from accumulating DNA damage [53, 192], as male germ cells are particularly vulnerable to certain types of stress [33]. This is still an active area of research, and in all likelihood, both DNA damage repair and replication errors contribute to the high paternal mutation rate.

While male germ cells continue dividing, oocytes are finished by birth, and remain in a state of suspended meiosis until fertilization. During this time, their DNA can accumulate damage, such as double stranded breaks (DSBs). DSBs are preferentially repaired by homologous recombination [171], where DNA with high sequence identity is used as a template to match missing bases. However, in cases of repeated sequence, like segmental duplications, this can result in interlocus gene conversion, creating a new allele in one sequence that resembles the other (Figure 1.1B) [19, 39]. As female germ cells age, they accumulate more damage over time, resulting in an increase of 1 DNM for every 2-3 additional years of maternal age [53, 162].

The mutagenic mechanisms of replication are not exclusive to the male germline, nor are the mechanisms of repair exclusive to the female germline. Both processes are at play across gametogenesis, resulting in different mutational profiles and rates between male and female germ cells, although the exact contribution of each mutational mechanism is still up for debate [53, 81, 91, 162].

It is also important to note that these mutational processes do not act uniformly across every base pair in the genome. Consider the two types of substitutions: transitions and transversions. Since there are twice as many possible types of transversion, we might expect roughly two transversions for every transition. In fact, transitions occur twice as often as transversions [48, 105], in part due to a bias in the mistakes DNA polymerase makes during DNA replication [45].

Sequence context, or the bases around a site, can also have a significant influence on mutation, so let's go through some common cases. Starting small, sequence context is influential on the scale of a single base, as conversion events are GC-biased, meaning that single

base pair mismatches are more likely to be repaired as a G or a C than an A or a T [40]. Zooming out by just one nucleotide, the neighboring base can have an outsized influence on a site's mutation rate. The classic example is at CpG sites, defined as a cytosine followed by a $3'$ guanine. The majority of these cytosines end up with a methyl group attached to them [13], which can spontaneously deaminate [26]. As a result, cytosine is converted to thymine, creating a mismatched base pair which can be repaired incorrectly, replacing the CpG site with a TpG site (Figure 1.1C). This process significantly increases the mutation rate at CpGs [75], such that these mutations make up approximately 20% of all observed DNMs.

On a larger scale, we see more sequence biased mutations in locally repetitive regions, like tandem repeats. In these regions, DNA polymerase can slip during replication, inserting or deleting copies of the repeat [17]. Even a site's location on a chromosome can have an impact on its mutability. We know, for example, that genomic regions that are replicated later in S phase (the stage of the cell cycle in which a copy of all the DNA is made) have a higher mutation rate than early-replicating regions [169].

1.2.2 Postzygotic mutations

The hallmark of a PZM is that it is mosaic in an individual, meaning the mutation must be incorporated into the genome after fertilization. However, the damage that led to the mutation may have occurred in a germ cell prior to embryogenesis [2]. Sperm cells do not carry a full suite of repair proteins, meaning that any damage incurred from oxidative or other stressors can go unrepaired until the DNA makes its way into an egg [33]. Once a zygote is fertilized, it does not transcribe its genome until the 4-8 cell stage [15], so early embryonic cells must rely on whatever repair proteins are present in the egg cell [124]. This might hinder damage repair, creating mutations that are passed on to one of the zygote's daughter cells. It is more likely, however, that the majority of PZMs arise due to errors in DNA replication, as the first rounds of cell division after fertilization are more error prone than later rounds [14], and an association has been found between PZMs and late-replicating regions of the

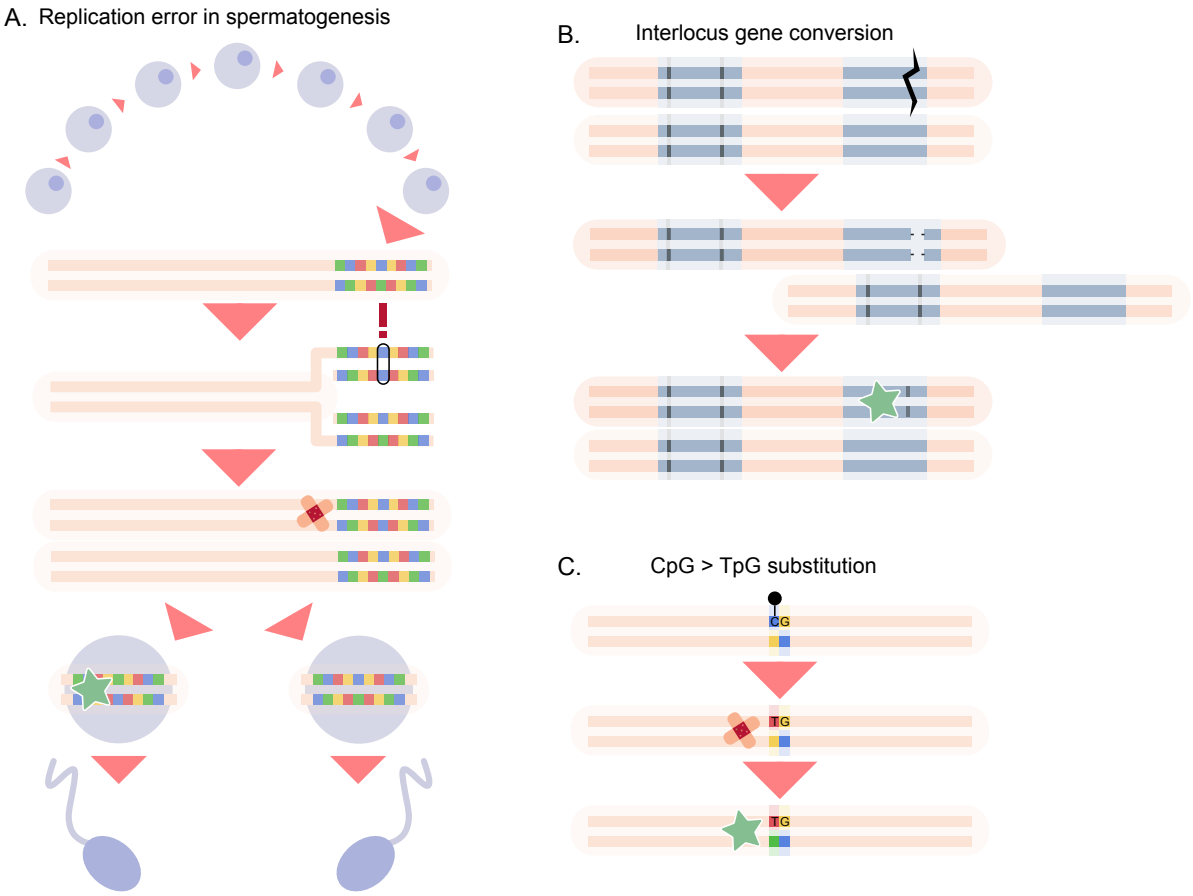


Figure 1.1: **Mutation mechanisms.** A. A toy example of an error in DNA replication giving rise to a DNM during spermatogenesis. An incorrect base is incorporated to the new strand of DNA and is incorrectly repaired by DNA polymerase. B. Interlocus gene conversion gives rise to a new DNM by introducing a variant from paralogous sequence. C. A cytosine at a CpG site is methylated and converted to thymine, creating a DNM.

genome [177]. In total, every cell division in early embryogenesis likely contributes 2-3 PZMs [80], but this may be an underestimate, as PZMs are difficult to identify.

1.3 The de novo mutation rate

The first estimate of the *de novo* mutation rate dates back to 1935, when Haldane attempted to derive the frequency of spontaneous hemophilia cases [64]. He figured there was one hemophilia-causing mutation in every 50,000 "life cycles." Nowadays, we define mutation rates in terms of mutations per base pair per generation (DNMs/bp/gen). For a number of DNMs, m, observed in N samples over G bases of callable genome, the mutation rate μ can be calculated as:

$$\mu = \frac{m}{N \times G}$$

Note that G should be multiplied by 2 in the case of a diploid genome.

Using Haldane's estimate of 1 mutation in every 50,000 generations across the exons of the haploid Factor VIII and Factor IX genes that cause hemophilia, we calculate a mutation rate of about 0.24×10^{-8} DNMs/bp/gen. Considering that Haldane was looking at an X-linked disease, and therefore only estimating mutations in the maternal germline, I would say this estimate holds up remarkably well in light of recent data. Let's put a pin in that.

In 2000, Nachman and Crowell compared chimpanzee and human pseudogenes to determine that the human mutation rate is 2.5×10^{-8} DNMs/bp/gen [126], but Kondrashov quickly revised the estimate down by examining the human mutation rate in 20 disease loci, coming away with a figure of 1.8×10^{-8} DNMs/bp/gen [90]. It wasn't until the 2010s that the first genome-wide assessments of the mutation rate based on familial whole genome sequencing (WGS) data began to emerge. 2010 and 2011 saw two studies, one based on two parents and their two children (a "quad") and one based on two groups of two parents with one child (two "trios"). These studies estimated a mutation rate between 1 and 1.1×10^{-8} DNMs/bp/gen [149, 157].

From there, *de novo* studies have expanded in size to examine thousands of trios, revising estimates of the germline mutation rate up to approximately 1.2-1.4×10^{-8} DNMs/bp/gen

(Table 1.1) [11, 50, 81, 91, 162]. If we remember that maternal DNMs make up less than 25% of all germline events, Haldane's original estimate works out to 1.1-1.2$\times10^{-8}$ DNMs/bp/gen - not bad for someone who didn't even know the structure of DNA!

Study of the postzygotic mutation rate, meanwhile, is a much newer phenomenon, and estimates of early embryonic mutations don't translate perfectly into the number of PZMs you might see in a tissue sample. That said, studies of DNMs report approximately 2-7 PZMs per individual [1], meaning PZMs might comprise up to 10% of all reported DNMs [36, 107, 162, 191]. We can extrapolate that to a postzygotic mutation rate of approximately 0.1×10^{-8} PZMs/bp/gen.

1.4 The art (and science) of DNM discovery

At this point, you may be wondering how we go from sequencing reads to *de novo* variants. Or perhaps the question hasn't occurred to you, because the process is theoretically so simple: just find all the places a child's DNA is different from their parents, count them up, publish a paper, get a Ph.D., and move on with your life. If only. So what does the process actually look like?

Let's say you receive raw sequencing data from a child and their two parents. In order to use these data, you first need to align the sequencing reads to a reference genome, finding each read's unique position. Next, you use variant-calling software to compare each sample's sequence to the reference sequence and identify any differences. There are many variant callers you could choose, or you could opt to use a combination of callers. More on that later, but for now you have a list of parent and child variants, so you can go ahead and select which variants are exclusive to the child to create a list of candidate DNMs. While there are a million parameters you could tweak here, these first few steps have well-established best practices [9], and most studies use pretty similar workflows to start [8, 10, 61, 81, 138, 150, 162, 176, 189, 192]. Finally, you need to filter and validate your candidate DNMs, count them up, and use them to calculate your mutation rate. Here's where the art comes in.

Value	Study	Estimate	CI - 2.5%	CI - 97.5%
germline mutation rate	Sasani et al. 2019	1.10×10^{-8}		
	Turner et al. 2017	1.50×10^{-8}		
	Jonsson et al. 2017	1.29×10^{-8}		
	Rahbari et al. 2015	1.28×10^{-8}	1.13×10^{-8}	1.43×10^{-8}
	Kong et al. 2012	1.20×10^{-8}		
paternal age effect	Sasani et al. 2019	1.44	1.12	1.77
	Turner et al. 2017	1.49	1.32	1.65
	Jonsson et al. 2017	1.51	1.45	1.57
	Goldmann et al. 2016	0.91	0.81	1.02
	Rahbari et al. 2015	2.87	2.11	3.64
	Kong et al. 2012	2.01	standard error: 0.17	
maternal age effect	Sasani et al. 2019	0.38	0.21	0.55
	Jonsson et al. 2017	0.37	0.32	0.43
	Goldmann et al. 2016	0.24	0.15	0.34
paternal:maternal ratio	Sasani et al. 2019	3.96:1		
	Jonsson et al. 2017	4.02:1		
	Goldmann et al. 2016	3.58:1		
	Rahbari et al. 2015	3.70:1		
	Kong et al. 2012	3.90:1		
Y chromosome mutation rate	Helgason et al. 2015	2.30×10^{-8}	2.03×10^{-8}	2.58×10^{-8}
	Xue et al. 2009	3.0×10^{-8}	0.89×10^{-8}	7.0×10^{-8}
	Kuroki et al. 2006	4.77×10^{-8}		

Table 1.1: **Previously observed mutation rates.** Estimates of mutation rates from Illumina-based DNM studies (in colors) [61, 81, 90, 150, 162, 176]. For chrY, Helgason and Kuroki give mutation rates in substitutions per base pair per year whereas here they are scaled by parental age of 30 [68, 95, 194]. CI: confidence interval

1.4.1 Filtering false positive DNMs

As you might remember, we expect to see around 100 DNMs per sample, but an unfiltered set of candidate DNMs may include thousands of DNM calls per sample. Most of these calls are false positives: they are either inherited variants that were not properly represented in parental data or artifactual variants from mistakes in sequencing or alignment. DNM studies often end up using a baroque filtering strategy, leveraging different variant callers, quality metrics, regional annotations, and other samples to separate the wheat from the chaff.

Some of these filters are standard, such as implementing cutoffs for minimum genotype quality or minimum and maximum read depth (too few reads and random errors start to look like real variants, too many reads and you start to question whether the reads are correctly placed) [10, 81, 162]. Another common filter is minimum allele balance, or the fraction of reads that carry the mutation. We expect a new mutation to be present on roughly half of reads, so it is common practice to assume a variant on fewer than, say, 25% or 30% of reads is either somatic in origin or the result of an error [10, 81, 91, 176, 189].

Other filters are more study-specific. For example, although DNMs can and do occur in clusters, groups of nearby DNMs can also be caused by sequencing or alignment errors, and the majority of clustered calls are false positives. Accordingly, some studies remove all DNMs within a certain distance of each other [192], and other studies set these mutations aside for their own treatment [10, 61, 176]. Speaking of alignment errors, read alignments can be ambiguous in repetitive regions of the genome, such as segmental duplications, as there are multiple sequences to which they might correspond. False events due to mismapped reads can be removed by applying a mapping quality filter, but some studies may opt to entirely exclude these regions instead [176, 189]. On a smaller scale, local sequence context can decrease the accuracy of sequencing data, like in homopolymers, strings of a repeated base which are plagued by spurious insertions or deletions. A study may opt to omit variants in these regions or apply additional scrutiny to them [162, 189].

A more straightforward approach is to exclude any DNM calls that have been observed

in other studies, or that are observed in more than one sample in a dataset [10, 192]. Over the 3 billion base pairs in the genome, the probability of the same mutation occurring in two samples is vanishingly low, especially in a small dataset. However, omitting variants that are observed in multiple samples might exclude recurrent mutations, unfairly penalizing the most mutable regions of the genome [1]. This strategy is most successful if you happen to have sibling information, as any variant that appears *de novo* in two siblings is most assuredly inherited from a parent, or so Occam would say [81, 162, 189, 192].

Finally, and this is a strategy the Eichler lab loves to employ, using multiple variant callers can increase the specificity of a callset [176, 189, 131]. The false positive rate for variants identified by multiple calling algorithms is meaningfully lower than for variants identified by just one caller [131]. However, using multiple callers alone is not sufficient to generate a perfectly accurate callset, so this strategy must be used in tandem with other quality filters.

1.4.2 Validating variants - wait, how is this different from filtering?

Philosophically, it might be impossible to prove a negative, but in the world of DNM discovery, I think it's a lot harder to prove a positive. Sure, we can use all sorts of filters to rule out DNMs as false positives, but when we have whittled down our final callset, how do we know the remaining calls are real? There are three main techniques to validate variants, each a little more reliable but more difficult than the last.

The simplest strategy is to manually evaluate variants by visualizing sequencing reads with a tool like the Integrative Genomics Viewer (IGV) [8, 138, 158]. This method does not require the generation of any additional data, which might leave you asking how it adds any additional information. No matter how specific variant filters are, there are always some variants that, when visualized, just *look wrong*. It may seem unscientific to rule out a variant based on intuition or feel, but actually seeing read data can be illuminating. For example, you might notice that the reads are quite noisy with lots of variants nearby, or that there are stray reads from a parent that weren't considered by a variant caller. While imprecise, checking variants in IGV is a great first step to building confidence in DNM calls.

If there is still DNA left from the samples, you can go back and evaluate it with a different sequencing method. Most commonly, studies will use targeted Sanger sequencing to make sure that a variant is present in a child, but the parents can also be resequenced to check that they don't have the variant [8, 192]. This method is great for confirming that a variant isn't just a sequencing artifact, but it relies on the ability to revisit the DNA. What's more, Sanger sequencing can't be applied uniformly throughout the genome, as repetitive regions cannot be uniquely targeted.

Perhaps the most convincing way to validate a DNM is to show that it is transmitted to the next generation. Once a sample has children, if they truly carry a DNM, we would expect it to be passed on 50% of the time, just like any other variant in the genome. Studies that have access to larger pedigrees can sequence the children of a sample and confirm that every DNM is present in at least one of the children – any DNM that isn't passed on is assumed to be a false positive or a somatic mutation masquerading as *de novo* [81, 162, 192]. However, most studies do not have the luxury of working with a three-generation family, and this strategy could not be applied to children or individuals without offspring.

1.4.3 Common denominator

Once you have assembled your final DNM callset and assessed your false positive rate, you can finally calculate your mutation rate. Well, almost. You've found the numerator in the equation, but now you have to determine what the denominator is. In a perfect world, the denominator would be every site in the genome. However, given the limitations of your sequencing data and variant filters, you can only truly assess variation at a subset of sites. Any site that is not callable should not be represented in the denominator, or it will serve to deflate your estimate of the mutation rate.

In order to determine which sites are callable and which should be excluded from the denominator, studies typically exclude any sites in the genome where at least one parent has a variant [10, 81, 192]. While there is no biological effect it should have on DNMs, parental heterozygosity often complicates variant calling pipelines, so any heterozygous sites

just aren't assessed from the jump. Further, any site that doesn't pass minimum depth or other sample-based quality filters in both parents and a child should be excluded, as DNMs there cannot be validated [8, 138, 162, 192]. If a study chooses to exclude variants in regions like segmental duplications or sequences like homopolymers, those sites must also be removed from the denominator. Most studies use some version of this logic to determine their denominator [8, 81, 138, 162, 192], and while it works well enough, I suspect it overestimates the amount of the genome that is truly callable.

1.5 It that it?

As I mentioned, recent *de novo* mutation studies have relied on Illumina WGS sequencing of trios and quads. These data expanded the mutational search space from a number of disease-causing loci to the scale of the whole genome, but they are insufficient for examining the most complex regions. The reason for their shortcoming is twofold. First, reference genomes of the 2010s omitted repetitive sequences like centromeres and the highest identity segmental duplications [163]. Second, many repetitive regions that were represented in the reference could not accurately have short-read data aligned to them, restricting the *de novo* search space to approximately 84% of the genome [174]. Because they couldn't align repetitive sequences, many *de novo* studies omitted repetitive regions altogether, or designed filters for mutations in repeats to fail [176, 189].

With the advent of highly accurate long-read sequencing and a complete human reference genome [132], the quest for *de novo* variation can begin anew. In these previously inaccessible, repetitive regions of the genome, I suspect that the mutation rate is higher than in unique regions due to mutational mechanisms like interlocus gene conversion [180]. By identifying DNMs in newly accessible, hypermutable regions of the genome, I'd wager that the overall germline mutation rate can yet again be revised upward.

Not only can long reads help to shine a spotlight on germline DNMs in high sequence identity regions, they can also identify postzygotic mutations as well. Typically, PZMs can only be assessed in three-generation families, where they are recognizable by their incomplete

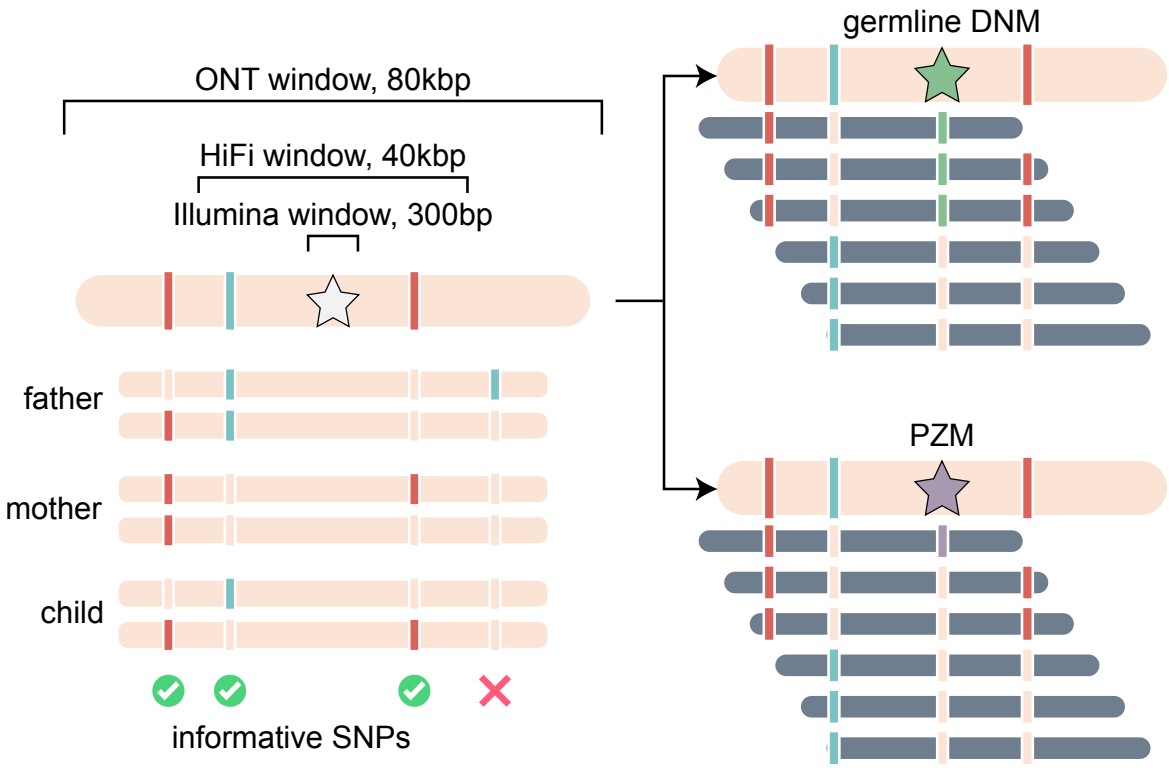


Figure 1.2: **PZM identification strategy.** We can use informative SNPs, or variants that can be uniquely assigned to one parent, to assign a DNM to a parental haplotype. A germline DNM should be present on every read on a parental haplotype, while a PZM is only present on a fraction of the reads from a haplotype.

transmission pattern to the third generation [162]. With long-read data, the majority of mutations should fall in sequencing reads with an informative variant on them, such that they can be assigned to a paternal or maternal haplotype (Figure 1.2). PZMs can be defined as *de novo* variants that are not represented on every read from a parental haplotype, making it possible to distinguish all postzygotic from germline mutations. If PZMs truly do make up a tenth of all discovered DNMs, then I might expect the germline *de novo* mutation rate to be 10% lower than predicted.

By this point, you must be dying to know: does the germline mutation rate go up or

down when you use long reads to identify DNMs? I'm so glad you asked.

1.6 The contents of this dissertation

In this work, I will describe three efforts to bring the mutation rate into the long-read era.

1.6.1 You have to walk before you can run

In Chapter 2, you'll become very closely acquainted with one family, a quad from the Simons Simplex Collection [47]. We deeply sequenced this family with every technology available to us, including Illumina short reads and PacBio High Fidelity (HiFi) and Oxford Nanopore Technologies (ONT) long reads, with the goal of discovering every *de novo* variant in the two children. Using reads aligned to both the old GRCh38 reference and the more complete T2T-CHM13 reference, I compared four variant calling methods: two previously described Illumina-based methods and two HiFi-based variant callers. I found a total of 195 germline *de novo* mutations across both children, in addition to 23 potential postzygotic mutations.

By using HiFi data for discovery instead of Illumina, we expanded our search space to at least 88% of the reference genome. Accordingly, HiFi data enabled a 20% increase in SNV discovery over Illumina, and the callset size grew by an additional 5% when reads were aligned to the T2T genome. As I expected, some of these gains were in segmental duplications and centromeres, where I found 4 and 3 DNMs, respectively, exclusively in T2T-aligned long reads. Based on this family, I calculated a germline *de novo* mutation of 1.41×10^{-8} DNMs/bp/gen.

1.6.2 42: the answer to life, the universe, and the mutation rate?

In Chapter 3, I expand my short- and long-read dataset to include 42 samples, including the 2 samples from Chapter 2, all sequenced with Illumina, HiFi, and ONT. With more samples, I hypothesized I would find a significant enrichment of DNMs in repetitive regions, like segmental duplications. To do so, I further refined my HiFi calling approach, including

a new strategy to identify postzygotic mutations by assigning them to parental haplotypes and a new sex-chromosome specific calling approach. My new method was able to evaluate 91% of the genome, and recovered a total of 3,703 *de novo* SNVs genome-wide across all 42 samples. Of those SNVs, approximately 15% were postzygotic in origin, a 1.5-fold increase from what has been reported in previous studies.

I found significant enrichments of both DNMs (38% increase) and PZMs (3-fold increase) in segmental duplications, driven by the highest-identity repeats. In addition, I was able to discern differences in germline and postzygotic mutational profiles, including a postzygotic depletion in CpG>TpG substitutions, the most commonly observed *de novo* mutation type. With this dataset, I calculated a genome-wide germline mutation rate of 1.31×10^{-8} DNMs/bp/gen and postzygotic rate of 0.20×10^{-8} PZMs/bp/gen.

1.6.3 Going platinum

In Chapter 4, I move to an entirely new dataset, a four-generation pedigree with 28 members, 10 of whom I used for DNM discovery. Like the previous dataset, these samples were sequenced with HiFi, ONT, and Illumina, but the pedigree structure provides a unique advantage: it allowed me see DNMs transmitted to the next generation. This confirms that DNMs are, in fact, present in the sample's germline and not merely tissue-specific mutations or calling errors. I applied a similar *de novo* discovery method to the one used in Chapter 3, with some modifications to leverage the power of the pedigree. In total, I found 745 *de novo* SNVs, 14.6% of which were postzygotic in origin.

Four samples had offspring with HiFi data that I could check for DNM transmission. I found that 97% of germline DNMs calls were transmitted to the next generation, compared to 60% of PZMs. In this family, I also observed a significant enrichment of DNMs and PZMs in segmental duplications, again driven by the highest identity repeats, confirming my hypothesis from Chapter 3. I calculated a germline mutation rate of 1.17×10^{-8} DNMs/bp/gen and postzygotic rate of 0.22×10^{-8} PZMs/bp/gen.

Chapter 2

FAMILIAL LONG-READ SEQUENCING INCREASES YIELD OF *DE NOVO* MUTATIONS.

This chapter is adapted with minimal modification from:

> **Noyes MD**, Harvey WT, Porubsky D, Sulovari A, Li R, Rose NR, Audano PA, Munson KM, Lewis AP, Hoekzema K, Mantere, T, Graves-Lindsay, TA, Sanders, AD, Goodwin, S, Kramer, M, Mokrab, Y, Zody, MC, Hoischen, A, Korbel, JO, McCombie, WR, Eicheler, EE. Familial long-read sequencing increases yield of *de novo* mutations. Am J Hum Genet 109: 631-646. (2022)

2.1 Abstract

Studies of *de novo* mutation (DNM) have typically excluded some of the most repetitive and complex regions of the genome because these regions cannot be unambiguously mapped with short-read sequencing data. To better understand the genome-wide pattern of DNM, we generated long-read sequence data from an autism parent-child quad with an affected female where no pathogenic variant had been discovered in short-read Illumina sequence data. We deeply sequenced all four individuals using three sequencing platforms (Illumina, Oxford Nanopore, and Pacific Biosciences) and three complementary technologies (Strand-seq, optical mapping, and 10X genomics). Using long-read sequencing, we initially discovered and validated 171 DNMs across two children—a 20% increase in the number of *de novo* single-nucleotide variants (SNVs) and indels when compared to short-read callsets. The number of DNMs further increased by 5% when considering a more complete human reference (T2T-CHM13) due to the recovery of events in regions absent from GRCh38 (e.g., three DNMs in heterochromatic satellites). In total, we validated 195 *de novo* germline mutations and 23

potential post-zygotic mosaic mutations across both children; the overall true substitution rate based on this integrated callset is at least 1.41×10^{-8} substitutions per nucleotide per generation. We also identified six *de novo* insertions and deletions in tandem repeats, two of which represent structural variants. We demonstrate that long-read sequencing and assembly, especially when combined with a more complete reference genome, increases the number of DNMs by >25% compared to previous studies, providing a more complete catalog of DNM compared to short-read data alone.

2.2 Introduction

de novo mutations (DNMs) are spontaneous germline mutations that arise through a myriad of mechanisms, such as replication error, DNA damage repair, and non-allelic homologous recombination. Different mechanisms give rise to different types of mutation, the most common of which are small single-base substitutions (single-nucleotide variants [SNVs]) and insertions and deletions of a small number of bases (indels); *de novo* SNVs and indels have been reported at an average rate of approximately 70 DNMs per individual [81, 91, 176]. Other classes of mutation, like expansions of tandem repeats, have been estimated to be very common as well (>50 events per individual) but are currently incompletely ascertained [122]. Larger mutations, such as structural variants (SVs), which affect more than 50 bp, are significantly rarer and have been observed at a rate of approximately one in every six individuals [7, 136]. All three classes of mutations have been implicated in autism, and it is estimated that more than 30% of all ASD cases may arise as a result of DNM in a protein-coding sequence or a *de novo* SV [76]. These estimates are based almost solely on the analysis of thousands of families using short-read whole-genome sequencing (WGS) datasets. Since long-read WGS methods have greatly increased sensitivity for SVs and large indels as well as all variant classes in repetitive loci [18, 43], we expect *de novo* rates may have been systematically underestimated.

Mapping Illumina sequence data can successfully access approximately 84% of the genome [174]. Repetitive regions, where the same 150 bp long read maps to multiple locations, are

Genomic Technology	**PacBio CLR**	**PacBio HiFi**	**Illumina**	**ONT**	**10X**	**Strand-seq**	**Bionano OGM**
Source	Blood	Blood/Cells[a]	Blood	Blood	Cells[a]	Cells[a]	Cells[a]
Platform	Sequel	Sequel 2	Hi Seq X Ten	PromethION	Chromium		
Metric	Coverage	Coverage	Coverage	Coverage	Mean depth	Number of cells	Effective coverage of reference
Father	55.5	47.1	37.5	27.5	73.1	66	273.5
Mother	54.5	43.8	33.5	29.0	61.4	63	256.2
Proband	74.4	34.0	41.8	30.8	64.1	56	246.2
Sibling	63.2	30.6	32.9	34.3	41.8	48	294.0
Center	UW	UW	NYGC	CSHL	WU	EMBL	Radboud

Table 2.1: **14455 sequencing summary** For each genomic technology – PacBio continuous long-read (CLR) and high-fidelity (HiFi) sequencing, Oxford Nanopore Technologies (ONT) sequencing, Bionano optical genome mapping (OGM) – the depth of sequencing is given for each member of the family. Coverage is based on genome size of 3.1 Gbp.
[a]Cells are EBV-transformed lymphoblasts.

typically excluded, potentially underestimating the true mutation rate [41]. In addition, Illumina sequencing is insensitive to large SVs where it is estimated that 75% of events (especially insertions) are missed in callsets generated from short-read sequencing technology [18]. Previous efforts to identify *de novo* variation using long-read sequencing were able to call *de novo* SNVs and indels in five individuals, but even their highest confidence candidate set only had a true positive rate of 79% and failed to recover any *de novo* structural variation [136]. Other studies have successfully identified *de novo* SVs using long-read sequencing but did not address SNVs or indels [117, 152]. In this study, we wished to measure the full extent of human *de novo* variation that exists in a family. To that end, we deeply sequenced DNA derived from blood from a family composed of two parents and their dizygotic twins with multiple long- and short-read technologies, including Pacific Biosciences (PacBio) continuous long-read (CLR) and high-fidelity (HiFi) sequencing, Oxford Nanopore Technology (ONT) sequencing, Chromium 10X genomic sequencing (10X), and single-cell DNA template strand

sequencing (Strand-seq) [46], and complemented with Bionano Genomics optical genome mapping (OGM) (Table 2.1). This family was selected because one of the children, the female proband, is affected with autism, and no genetic cause has been identified. This is despite extensive study by both whole-exome [92] and whole-genome Illumina sequencing [176] and the twofold increased likelihood of discovering a genetic event in females with autism [97]. Here, we investigate and quantify the difference in DNM detection that can be reliably identified between short- and long-read data as well as the effect of a more complete reference genome for variant discovery. The use of multiple orthogonal sequencing technologies allows all events to be validated, producing a rigorous truth set with the potential to improve DNM detection and estimates of DNM rates.

2.3 Results

2.3.1 Detection of de novo variation using PacBio HiFi sequencing

In order to identify *de novo* SNVs and small indels (<20 bp), we initially applied three orthogonal sequencing technologies to blood-derived DNA obtained from each member of a simplex autism family (two parents and a dizygotic twin pair) where the daughter had been diagnosed with autism while the son was unaffected (Figure 2.1). We used Illumina and HiFi sequencing for variant discovery, while ONT, owing to its higher error rate, was used strictly for validation purposes (Figure A.1). Illumina WGS short-read data were generated and aligned to the reference GRCh38, and variants were called using both GATK HaplotypeCaller [140] and FreeBayes [56] using previously described best practices (Turner et al. [176]). *De novo* variants identified by both callers were included in the two-caller callset, containing 171 total DNM calls across the proband and sibling (true positive rate = 78.4%). To generate a more sensitive Illumina callset, we also included the variant callers Strelka2 [88] and Platypus [156] as described previously (Wilfert et al. [189]). *De novo* variants identified by at least three of the four Illumina callers were included in the four-caller callset, containing 144 total DNM calls across the proband and sibling (true positive rate = 91.7%).

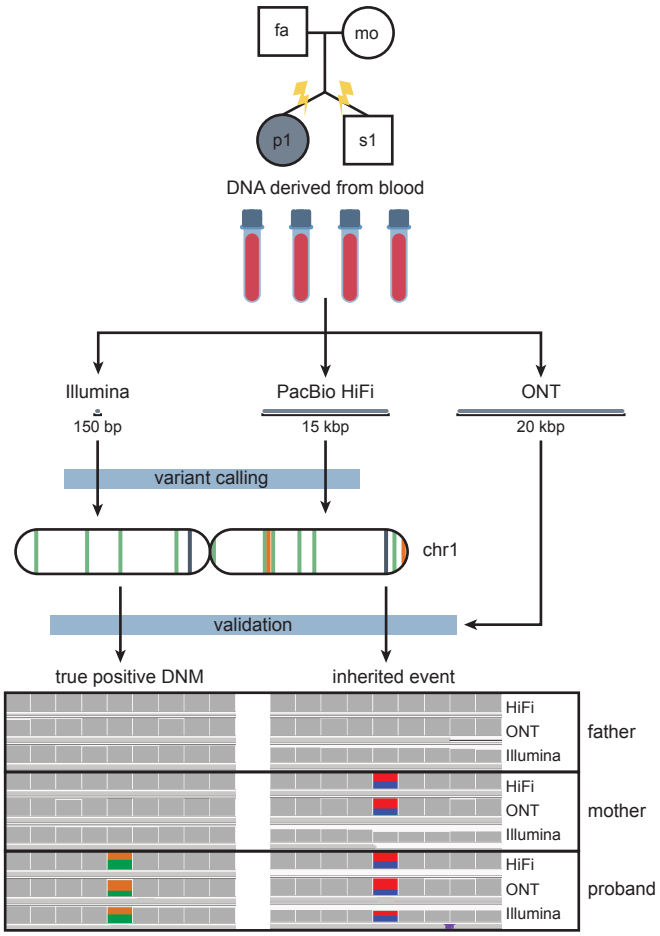


Figure 2.1: ***de novo* SNV calling and validation method.** The pedigree of the quad, which consists of two parents, aged 35.53 and 35.20 (father and mother) at the time of their children's birth, and their dizygotic twins. The proband, a female, is affected with autism, and her sibling is not. In this simplex case of autism in a female, there is an increased likelihood of finding a *de novo* mutation (DNM) to explain the cause of her autism. DNA derived from blood was sequenced using three different platforms: Illumina, PacBio HiFi, and ONT. Illumina and HiFi data were used for *de novo* discovery, and variants were validated by examining the sites across all three sequencing technologies. True positive *de novo* events are exclusive to the child, whereas misclassified inherited events can be seen in at least one parent.

Between the two- and four-caller callsets, we identified 180 candidate DNMs. These DNMs were validated by examining the sites in both HiFi and ONT sequencing data—any sites where the variant was absent in the parents and present in the child in the orthogonal data were designated as true positive events. After validation, the Illumina callsets identified 62 total *de novo* SNVs and indels in the proband and 80 in the sibling (true positive rate = 78.9%), setting a lower bound for the number of DNMs present in the children. The limited number of validated *de novo* variants in the Illumina callsets was due in part to the exclusion of variants in repetitive sequence, such as repeats with greater than 90% identity (recent repeats), including low-complexity regions (LCRs) [100] and centromeres, effectively restricting the callable genome to 78.6%. To identify variation missed by Illumina, we used HiFi sequencing, which generates long reads (median 15 kbp) that can unambiguously align to 88.1% of the genome. HiFi reads were aligned to GRCh38, and two variant callers were used to naïvely identify *de novo* SNVs and indels. The first caller, GATK [140], identified 211 DNM calls across the proband and sibling (true positive rate 80.6%). The second caller, DeepVariant [139, 196], identified 194 DNM calls across the proband and sibling (true positive rate 87.1%). Between both callsets, 217 candidate DNMs were identified and validated by examining the sites in both Illumina and ONT sequencing data. After validation, the HiFi callsets recovered 80 *de novo* SNVs and indels in the probands and 91 in the sibling—a 20.4% increase in the number of DNMs identified by Illumina (Figure 2.2A-C).

There were 75 DNM calls in HiFi not identified by the Illumina callers, 37 of which had support in ONT and retrospective analysis of the underlying Illumina sequence. As expected, most true DNMs exclusive to HiFi (23/37) were located in regions excluded by the Illumina pipelines; 82.6% of such calls were removed from Illumina callsets based on this mask. However, removing variants in masked regions is a crucial part of the Illumina calling pipelines, as it eliminates more than 300 false positive DNM calls across both samples. Conversely, 38 false DNM calls were made only by HiFi callers, nearly three quarters of which (n=27) were inherited variants incorrectly classified as *de novo* due to sequence coverage issues in one of the two parents. More than half of the inherited events (16/27) were located

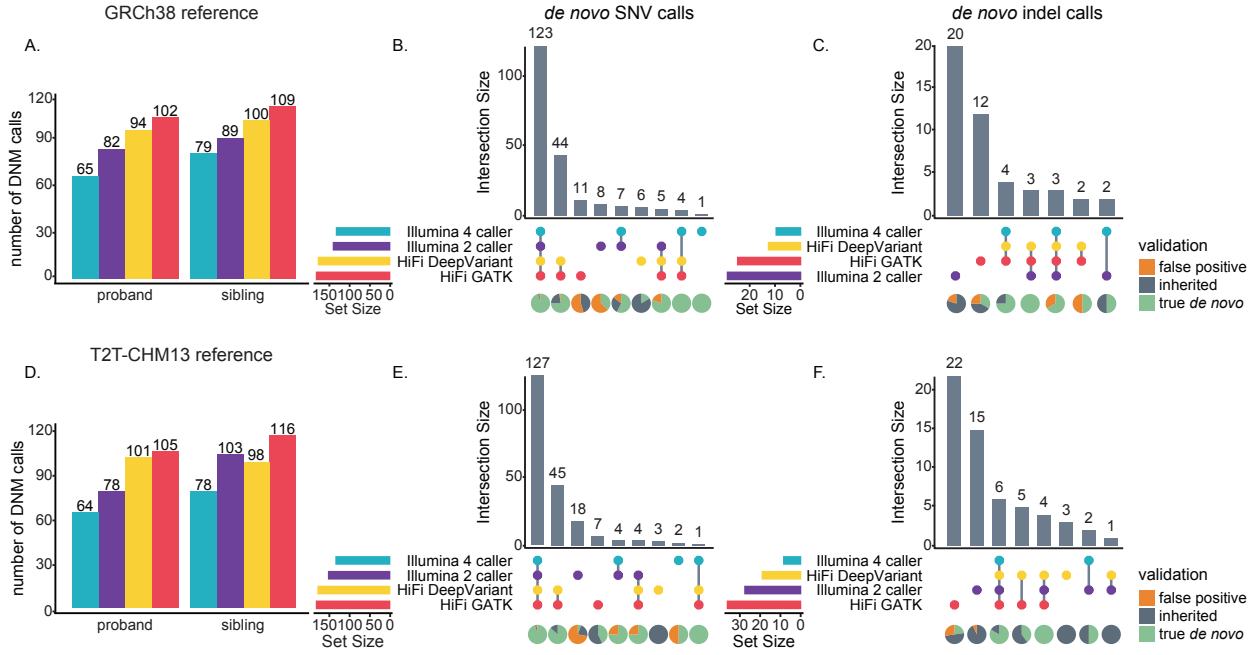


Figure 2.2: **Comparison of DNM recall across short- and long-read callsets.** A. Number of DNM calls in reads aligned to GRCh38, based on the Illumina four-caller (blue), Illumina two-caller (purple), HiFi DeepVariant (yellow), and HiFi GATK (red) pipelines. B. Upset plot showing the concordance of *de novo* SNV calls across GRCh38 callsets, with the proportion of true positive (green), false positive (orange), and inherited (gray) events shown below each category. Validation status was assigned by examining the variants in ONT and HiFi or Illumina sequences, as described in the Methods. C. Upset plot showing the concordance of *de novo* indel calls across GRCh38 callsets. D-F. The same analysis repeated on reads aligned to T2T-CHM13.

in clusters of less than 1 kbp. In most cases, inherited miscalls were the result of failure to sequence one of the parents' haplotypes to sufficient coverage and could be resolved by sequencing to higher depth. For example, the mean paternal and maternal read depth in false HiFi calls is significantly lower than in validated *de novo* events ($p = 1.01 \times 10^{-6}$ and $p = 7.75 \times 10^{-8}$, Welch two-sample t-test) (Figure A.2). Thus, more permissive discovery or subsequent genotyping of the parents may further reduce the number of false calls. There were also eight DNM calls identified in Illumina missing from the HiFi callsets—all but two of these calls were identified by at least one HiFi caller but were excluded because they were

either not sequenced to sufficient coverage or failed other quality filters (e.g., allele balance). In total, 179 true *de novo* SNVs and indels were discovered in GRCh38 aligned reads (82 in the proband and 97 in the sibling).

The reference genome GRCh38 is incomplete, missing regions such as centromeres and some highly identical segmental duplications. The newly assembled T2T-CHM13 genome [132] contains more than 240 Mbp of additional sequence. In order to discover *de novo* variation in these regions, we aligned the same Illumina and HiFi reads to the T2T-CHM13 assembly and used the same *de novo* calling pipelines to identify variation. Across the Illumina two- and four-caller callsets, we identified 184 DNM calls in the proband and sibling (true positive rate = 76.1%). Predictably, HiFi variant callers outperformed Illumina callers, as the HiFi GATK and DeepVariant callsets collectively identified 228 DNM calls in the proband and sibling (true positive rate = 80.3%). In total, of the 269 DNM calls made by the Illumina and HiFi callers using the T2T-aligned reads, 188 *de novo* SNVs and small indels had support in ONT and Illumina or HiFi sequence (Figure 2.2D-F)—a 5% increase in the number of DNMs identified when compared to GRCh38 aligned reads, and only seven sites were missing as seen exclusively by callers on GRCh38 aligned reads (Figure 2.3). Both HiFi callers performed better on T2T-aligned reads, not only identifying more *de novo* sites, but also generating callsets with true positive rates greater than those of the corresponding GRCh38 callsets (Figure 2.4A). By applying additional filters, we can improve the true positive rate by further reducing the number of false calls made in T2T-aligned data by 63%, at the expense of only three true DNM calls. First, by applying a parental genotype quality filter (GQ>25 for both parents) to the HiFi GATK callset, we can eliminate a total of 12 incorrect calls and 1 true call. Next, by applying a mapping quality filter (mapq>59) to the HiFi DeepVariant callset, we can further remove 7 incorrect calls and 1 true call. The biggest source of error, however, comes from the Illumina two-caller callset. There was a total of 33 calls identified by only the two-caller callset, 32 of which were false. By completely excluding the two-caller set from the analysis, we remove 32 incorrect calls and 1 true call. There is no overlap in the sites affected by these three methods—resulting in 54 sites removed

from the total callset—a loss of 51/81 incorrect calls and 3/188 real DNMs. In the remaining T2T callset, there are only 7 false positives and 23 inherited calls.

Figure 2.3: ***de novo* variant calling by technology and genomic region.** Regions of the genome with mapping quality <57 are highlighted in blue. HiFi mapping quality (mapq) is at the top of each chromosome and Illumina is on the bottom, along with all regions removed from Illumina callsets, including low-complexity regions, centromeres, and recent repeats highlighted in orange. The total accessible genome for GRCh38-aligned HiFi reads is 2.88×10^9 bp and for Illumina reads is 2.36×10^9 bp. The total accessible genome for T2T-CHM13-aligned HiFi reads is 3.07×10^9 bp and for Illumina reads is 2.31×10^9 bp. DNMs identified by each of the HiFi and Illumina callers are plotted in their respective locations. The chromosome 4 popout shows four true DNM calls that were made by HiFi callers but not Illumina callers and includes their annotations. The chromosome 14 popout shows three true DNM calls made by T2T callers but not GRCh38 callers.

2.3.2 Detection of de novo variation using a more complete reference genome

Most true *de novo* events were supported in both T2T and GRCh38 aligned sequence, with the exception of two events that could not be lifted from GRCh38 to T2T coordinates, and six events that had support in ONT and HiFi or Illumina reads aligned to GRCh38, but

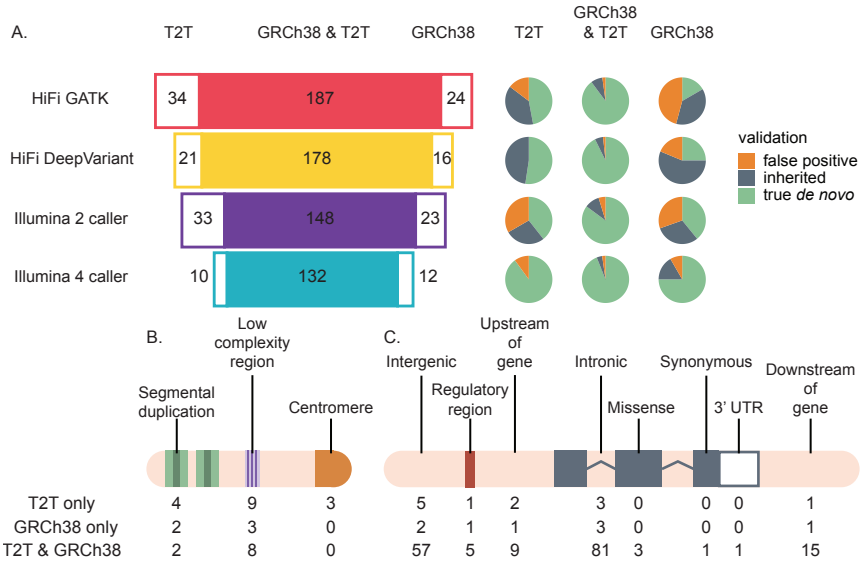


Figure 2.4: **Human genome reference comparisons.** A. Concordance between T2T-CHM13 and GRCh38 callsets for each caller used. The overlap between callsets is on the left, and the proportion of true positives, false positives, and inherited events is on the right. B-C. Functional annotation of DNM calls across T2T and GRCh38 callsets. The three categories in B (repetitive regions) are separate annotations from those in C; for example, a site that is in both a segmental duplication and a regulatory region would be included in both counts.

not when the same reads were aligned to T2T. Excluding those events, across both T2T and GRCh38 callsets there were 88 *de novo* SNVs and small indels in the proband and 107 in the sibling with support in ONT and either Illumina or HiFi sequence. Eight percent of DNM calls were identified exclusively in T2T aligned reads—3 were found in centromeres, 4 in segmental duplications, and 12 in LCRs or recent repeats (Figure 2.4B). A total of four *de novo* SNVs were initially called in centromeres and were manually inspected in IGV (Figure A.3), three of which were strongly supported in HiFi and ONT data; the fourth call was supported as *de novo* but the level of SNV heterozygosity for that portion of the alpha-satellite was much higher than anticipated, so we do not report it as a true *de novo* event.

In addition to DNM calling in repeats, variant calling sensitivity was also increased in functionally important regions, and T2T aligned reads were able to identify nearly all calls made in GRCh38 (Figure 2.4C). In total, we identified three missense mutations (in XPO1 in the proband, and USP49 and SEMA6B in the sibling), seven regulatory variants, and one 3' UTR variant. Combining all the data, we identify 195 DNMs in the proband (n=88) and sibling (n=107) where the DNM status has been validated by ONT and the variant is absent in parental data, 185 of which had support in HiFi, Illumina, and ONT. Taking advantage of the available ONT and HiFi long-read data (Methods), we successfully phased 99.5% (194/195) of the variants by considering informative single-nucleotide polymorphisms extending 20 kbp on either site of the variant position. Predictably, 72.2% (140/194) of the phased events originated in the male germline. In addition to these 195 events, another 14 sites (9 SNVs and 5 indels) had support in HiFi and Illumina data but not in ONT. Of those sites, five appeared to be false negatives in ONT (the affected child had no ONT reads with the alternate allele), despite having support in Illumina and HiFi data in both T2T- and GRCh38-aligned reads—these sites were not included in our final *de novo* callset. Another nine sites appeared inherited in ONT (one or both parents had more than one ONT read with the alternate allele).

In order to provide an estimate of false negatives and to maximize sensitivity, we measured the total number of *de novo* SNVs and small indels in the children by examining all of the candidate *de novo* calls made by DeepVariant and GATK on T2T-aligned HiFi reads. In this analysis, we removed the allele balance filter requiring the alternate allele to be present in at least 30% of the child's reads (while retaining other filters for minimum quality and read depth) and examined all of the remaining calls in ONT and Illumina. This set, which would include mosaic DNMs, contained 209 DNMs with support across all three sequencing platforms. It included all of the 195 DNMs except 12, which were identified by only Illumina callers. If we add these 12 to the total, we identify 221 DNMs across the proband and sibling, predicting a false negative rate of 11.7% and an upper bound on the total number of *de novo* SNVs and small indels. Of the 26 variants missed in our *de novo* calling analysis (Table 2.2),

Child	ID	HiFi	HiFi AB	ONT	ONT AB	Illumina	Illumina AB	Parental Haplotype
14455.p1	chr2_94618830_C_A	3 / 48	0.06	2 / 14	0.14	1 / 66	0.02	paternal hap1
14455.p1	chr4_1518905_G_C	5 / 16	0.31	2 / 13	0.15	1 / 25	0.04	maternal hap1
14455.p1	chr5_124981762_T_C	12 / 54	0.22	2 / 16	0.13	6 / 64	0.09	maternal hap1
14455.p1	chr9_42454095_C_T	4 / 33	0.12	3 / 9	0.33	1 / 37	0.03	maternal hap2
14455.p1	chr17_81586404_G_C	3 / 24	0.13	3 / 22	0.14	8 / 43	0.19	maternal hap1
14455.p1	chr18_15654268_G_T	8 / 40	0.20	2 / 21	0.10	1 / 62	0.02	paternal hap2
14455.p1	chrX_114777954_G_A	3 / 31	0.10	3 / 26	0.12	1 / 40	0.03	maternal hap1
14455.s1	chr1_2104522_A_G	2 / 20	0.10	1 / 26	0.04	0 / 23	0.00	paternal hap2
14455.s1	chr2_91095600_G_T	3 / 30	0.10	2 / 14	0.14	2 / 126	0.02	maternal hap1
14455.s1	chr3_96412796_A_C	2 / 33	0.06	1 / 40	0.03	0 / 34	0.00	maternal hap1
14455.s1	chr6_62484584_G_A	6 / 34	0.18	3 / 24	0.13	1 / 39	0.03	maternal hap2
14455.s1	chr6_70745885_CAT_C	4 / 22	0.18	2 / 31	0.06	1 / 7	0.14	unknown
14455.s1	chr6_127695258_G_A	5 / 35	0.14	1 / 35	0.03	2 / 32	0.06	paternal hap2
14455.s1	chr6_163529274_T_C	4 / 28	0.14	2 / 28	0.07	2 / 15	0.13	maternal hap1
14455.s1	chr7_2981075_ TATATAG_T	6 / 30	0.20	1 / 39	0.03	1 / 36	0.03	maternal hap1
14455.s1	chr7_58404995_T_A	4 / 30	0.13	1 / 29	0.03	1 / 50	0.02	paternal hap2
14455.s1	chr7_58405316_C_T	4 / 31	0.13	1 / 27	0.04	1 / 56	0.02	paternal hap2
14455.s1	chr7_156505941_C_A	2 / 30	0.07	1 / 29	0.03	3 / 22	0.14	unknown
14455.s1	chr11_50917416_C_A	3 / 28	0.11	5 / 17	0.29	1 / 28	0.04	paternal hap1
14455.s1	chr14_3924466_C_ CATTCCATTCCATTCT	1 / 23	0.04	2 / 3	0.67	1 / 54	0.02	unknown
14455.s1	chr14_10160460_G_T[a]	3 / 29	0.10	1 / 15	0.07	1 / 13	0.08	maternal hap2
14455.s1	chr19_5409128_C_T	6 / 33	0.18	1 / 26	0.04	8 / 55	0.15	paternal hap1
14455.s1	chr22_20777172_G_T[a]	8 / 42	0.19	1 / 27	0.04	1 / 62	0.02	maternal hap2

Table 2.2: **14455 potential mosaic mutations.** DNMs identified after removing the allele balance filter for HiFi long-read data aligned to T2T-CHM13. The number of reads with the alternate allele, total number of reads, and allele balance (AB) ratio for PacBio HiFi, ONT, and Illumina.
[a]All variants were identified by GATK with the exception of the two variants identified by DeepVariant, denoted with the superscript.

all but three are AB<0.35 in the affected child across all three sequencing platforms. Because of this consistently low allele balance, we suspect that these 23 variants may, in fact, represent potentially mosaic sites in the children (Figure A.4). None of these 23 sites were reported in previous *de novo* studies of this family [176, 189]. This results in a mosaic mutation rate of 2.39×10^{-9} mutations per nucleotide per individual, likely underestimating the true mosaic mutation rate because we are selecting only the highest frequency variants [125]. Of the 23 potential mosaic variants, 12 were assigned to maternal haplotypes and 8 were assigned to paternal haplotypes, resulting in a paternal:maternal ratio of 0.66:1, significantly different from the 2.59:1 ratio observed in the *de novo* germline variants ($p = 0.0067$, two-sample Z-test). Although the mosaic sample is small, this observation is consistent with the expectation that there is no parent-of-origin bias for post-zygotic mutations.

2.3.3 Detection of de novo STR and VNTR events

Candidate *de novo* expansions of short tandem repeats (STRs) and variable number tandem repeats (VNTRs) are particularly challenging using standard calling pipelines. We applied three orthogonal approaches to detected *de novo* events in the WGS data (Figure A.5). The first approach leveraged the targeted phased assembly from Sulovari et al. [173]; sequence resolved and phased STR and VNTR sites with larger repeats were examined for *de novo* variation in the proband and sibling, resulting in 10 candidate events. The second approach used ExpansionHunter Denovo [35] to identify repeat expansions that were present in the children but not their parents, identifying three candidate events. The third approach used a custom pipeline to compare the number of uniquely mappable 30-mers in the parents and their children (after controlling for GC-adjusted read depth using the same genomic control regions as Sudmant et al. [172]), selecting sites for subsequent analysis with a higher number of k-mers in the child relative to its parents. Using these three approaches, we identified a total 15 candidate *de novo* STR and VNTR events, none of which was initially observed by more than one approach (Table A.1). All 15 candidate events had support when validated with phased assembly generated by CLR reads haplotagged by the integrated 10x

Chromium and Strand-seq phased variant data. The events were further validated with Sanger sequencing: six failed to sequence, five were shown to be inherited variants, but four represented true *de novo* events. Of the true positive events, one was a VNTR expansion in the proband (Figure 2.5), one was an STR deletion in the proband, one was an STR expansion in the sibling, and one appeared to be an STR expansion in both the proband and the sibling (Figure A.6). The VNTR was identified by ExpansionHunter Denovo (true positive rate = 50%), one STR was identified by the 30-mer approach (true positive rate = 33%), and the remaining two STRs were identified by the approach from Sulovari et al. [173] (true positive rate = 20%). All three approaches had low true positive rates and identified far fewer *de novo* STR and VNTR events than expected based on previous reports [122].

In an effort to increase yield, we applied the same assembly-driven methodology used to detect structural variation to discover indels greater than or equal to 20 bp and less than 50 bp. Variants of this size disproportionately (86% of deletions and 64% of insertions) reside in short tandem repeats in GRCh38 coordinates. We started with a set of 12,284 deletions and 13,226 insertions in the proband in addition to 12,284 and 13,124 for deletions and insertions in the unaffected sibling. We then filtered this set down to 179 deletions and 276 insertions in the proband and 167 deletion and 219 insertion events in the unaffected sibling, but not in the parents. Automatic inspection of raw parental long-read alignment validation yields 7 potential *de novo* deletions and 14 potential *de novo* insertions in the proband. For the unaffected sibling this estimate is 3 and 15 for deletions and insertions, respectively. Manual inspection of the raw reads overlapping these calls yielded 2 confident *de novo* indels in the proband and 1 in the unaffected sibling, which were not seen in previous analyses (Figure A.7, Table A.2). However, one of the indels in the proband, originally identified as a 24 bp insertion with respect to the reference, was revealed to be an expansion of an 8 bp paternal allele, yielding a total *de novo* insertion length of 16 bp (Figure A.7B).

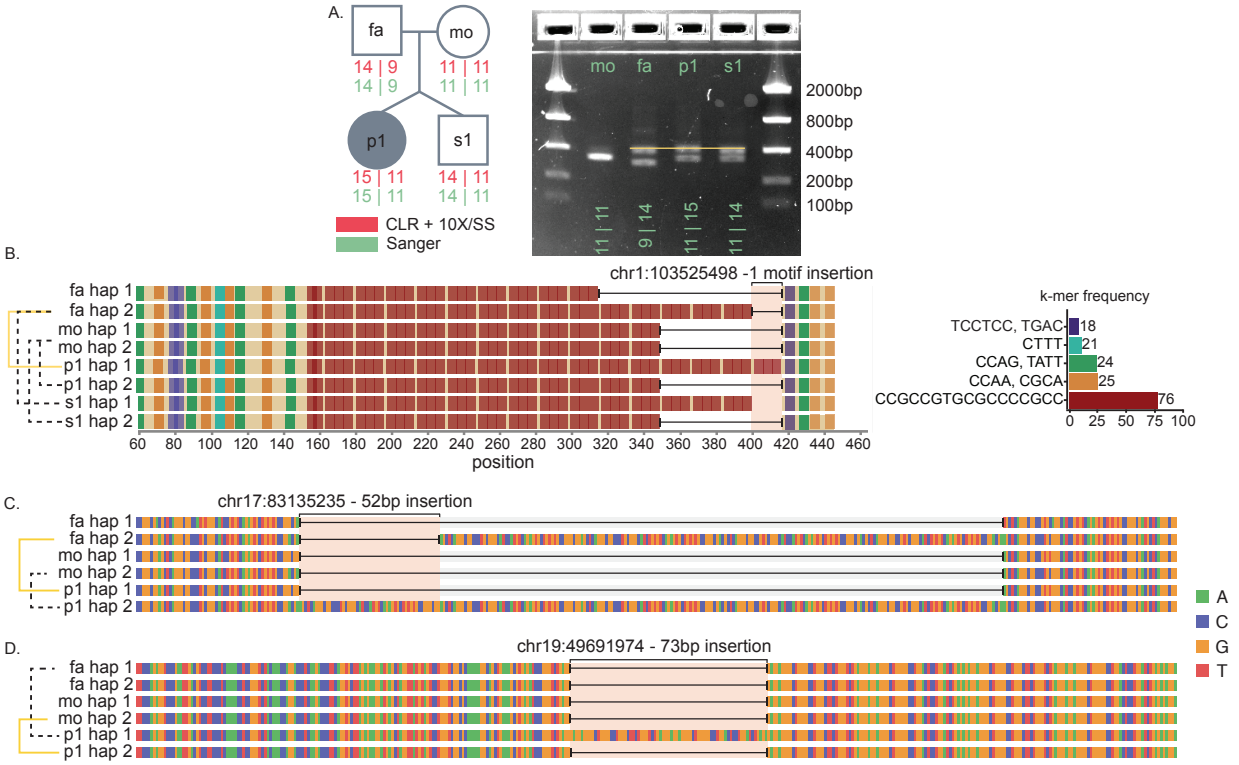


Figure 2.5: ***De novo* VNTRs.** A. The quad structure annotated with the number of repeats seen in the VNTR, as determined by PacBio CLR, 10X, and Strand-seq data in addition to Sanger sequencing, and represented on a gel. B. Haplotypes for every individual in the quad based on HiFi sequencing clearly show an extra copy of the motif (in red) in the proband. C. The 52 bp insertion in the proband compared to the parental haplotypes. D. The 73 bp insertion in the proband.

2.3.4 De novo structural variation detection

We also applied both assembly- and read-based approaches to discover SVs (events >50 bp in length). We initially generated haplotype-resolved assemblies using a combination of PacBio HiFi and Illumina short-read data using hifiasm v0.12 and applied the PAV caller to create a set of 9,982 deletions and 16,815 insertions in the proband [21, 43]. Similarly, for the sibling we started with a set of 9,999 deletions and 16,879 insertions. We used PBSV [43], subseq [43], and DeepVariant [139] to provide further support in addition to a secondary

analysis of PAV using lra [153]. We selected all variants detected by one or more of these additional callers for our SV callset for a total of 222 candidate *de novo* SVs consisting of 57 deletions and 165 insertions in the proband and 74 deletions and 121 insertions in the sibling. We initially validated these events by examining read-based support using subseq to analyze parental read data, resulting in 48 potential DNMs that were visually validated by examining both PacBio HiFi and ONT alignments over the regions in IGV. Of these events, 28 were clearly inherited, 8 appeared to be false positives, and the remaining 12 were absent in parental data but present in at least one read in both technologies for the proband (7) or sibling (5). These 12 candidates were finally validated by examining the haplotype-resolved assemblies for the parents and child, inspecting realigned contigs of the 6 kbp surrounding the site. Of the 12 candidate SV events, only 3 appeared to be true *de novo* events, with 2 in the proband and 1 in the sibling (Figure A.8).

In an effort to minimize the extent of manual curation, we automated this process and developed a novel pipeline that implements some of the approaches made during manual inspection (Figure A.9). By using a combination of subseq, callable regions from parental PAV calls, and multiple sequence alignment of familial haplotypes, we were able to validate the same 2 *de novo* events (both insertions) in the proband in an automated fashion, but we were not able to increase sensitivity. The single *de novo* candidate in the unaffected sibling did not validate with the automated pipeline, as multiple haplotypes were discovered overlapping this variant. Accordingly, we reclassified this variant as a low-confidence potential *de novo* event. This automated pipeline, named dnSVal, is available on GitHub. The 2 true *de novo* SVs that occurred in the proband were VNTR expansions but had not been identified using our STR/VNTR-specific approaches. Both *de novo* events in the proband map to genic regions (CPT1C intron and TEC exon), but neither have been functionally implicated in autism.

Because discovery of *de novo* SVs is still challenging, we finally considered the potential of applying both Strand-seq and Bionano Genomics as standalone technologies to increase discovery sensitivity. For Strand-seq, we used the procedure described in Ebert et al. [43]

to detect and phase 127 nonredundant inverted sites (median size: 38.9 kbp, min: 2.3 kbp, max: 4.3 Mbp). Because Strand-seq can unambiguously split short sequencing reads by haplotype, it makes possible the detection and phasing of large heterozygous deletions; we identified 62 redundant heterozygous deletions with respect to GRCh38 with a median size of 55.8 kbp (min: 10.1 kbp, max: 550 kbp). Considering parental genotypes, we initially identified 2 and 4 candidate inversion *de novo* events in the proband and sibling, respectively. However, after manual inspection of these automated inversion calls, these were determined to likely represent false positives, as they fall into regions where short Strand-seq reads map with lower confidence, such as centromeres and segmental duplications. In addition to the inversions, we detected 2 potentially large heterozygous deletions using Strand-seq as an orthogonal method. However, due to lack of support in phased HiFi reads, we were unable to validate these events.

Similarly, we used a Bionano coverage-depth-based algorithm to discover three *de novo* SV candidates in the proband (2 DEL, 1 INS) and five (3 DEL, 2 INS) in the unaffected sibling (Table A.3) [113]. With the exception of the deletions in the unaffected sibling, these calls are seen with relatively high frequency in the population according to Bionano controls. None of these events intersect with our read-based or assembly-driven callsets, nor do they contain any support in a manual inspection of the reads underlying this region. Given that none of the Bionano calls are supported by other data, we failed to identify any true *de novo* events using Bionano as a sole discovery tool. Since both the Strand-seq and Bionano Genomics data were derived from cell lines (as opposed to primary material), we consider the possibility that these invalidated events may also represent additional cell line artifacts.

2.3.5 Meiotic breakpoints and DNMs

Since recombination has been shown to be mutagenic in the human population [5, 50, 67], we reassessed our validated set of DNMs with respect to meiotic crossover positions in the parental haplotype. Leveraging the inherent phasing data present in Strand-seq along with the long-read PacBio sequencing data allows one to define crossover breakpoints at a fine-

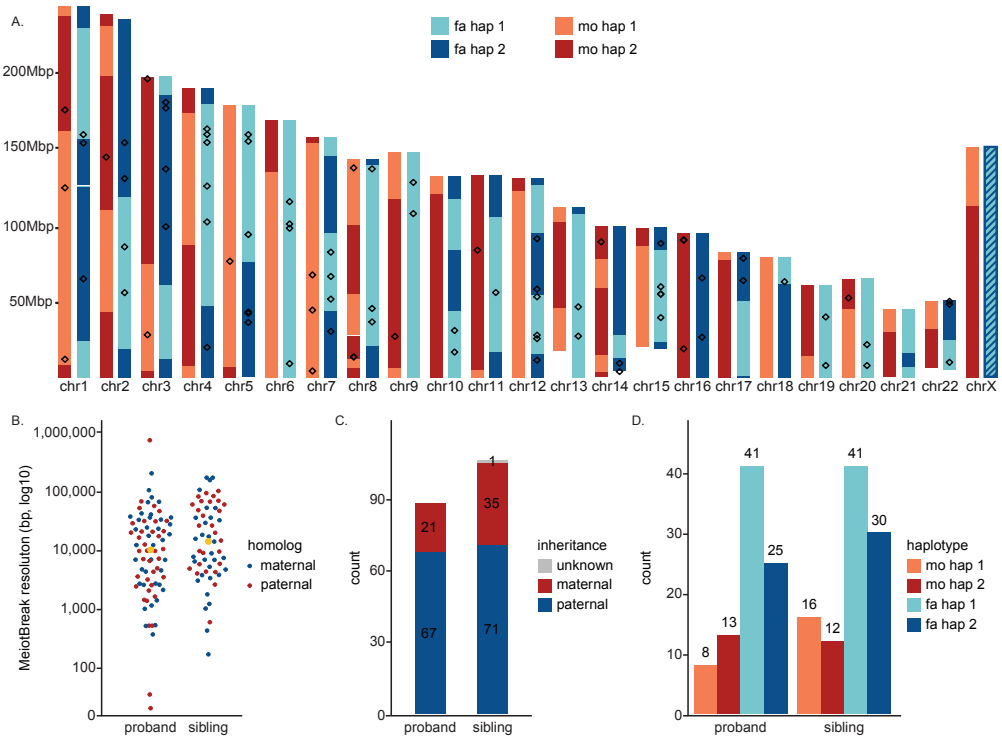


Figure 2.6: **Meiotic recombination and DNM.** A. A genome-wide overview of detected meiotic recombination breakpoints for the proband. Inherited segments of maternal homologs (H1-light red, H2-dark red) appear on the left side of each chromosome while inherited segments of paternal homologs (H1-light blue, H2-dark blue) appear on the right side of each chromosome. Recombination breakpoints are visible as changes from H1 to H2 segments and vice versa. Detected DNMs that could have been assigned to a single parental homolog (n=89) are shown as empty boxes over maternal (left) and paternal (right) homologs. This individual is a female meaning that paternal chromosome X does not recombine (empty blue box). B. Size distribution of detected meiotic recombination breakpoints for both the proband (n=76) and sibling (n=59). Median value is shown as an orange dot for both distributions. C. Total number of DNMs assigned to paternal (dark blue) and maternal (dark red) homologs, separately for proband (14455.p1) and sibling (14455.s1). D. Total number of DNMs that occurred on paternal homologs H1 (light blue) or H2 (dark blue). The same results are shown for maternal homologs H1 (light red) and H2 (dark red). Counts are reported separately for proband (14455.p1) and sibling (14455.s1). We could not determine inherited parental segments for one DNM in proband and for seven DNMs in sibling.

scale level of resolution without the need for grandparental sequence data [146, 143]. We defined 135 total crossover events in the proband (Figure 2.6A) and sibling (Figure A.10) at a fine-scale of resolution (median: 12.6 kbp) (Figure 2.6B, Table A.2). Among the children, maternal and paternal crossover events were equally distributed (69 maternal and 66 paternal) with no particular bias towards certain genomic regions. We then projected all DNMs and measured the distance of a DNM to the nearest crossover event in the maternal or paternal lineage (Figure 2.6C) as well as assigning the DNM to grandparental haplotypes (Figure 2.6D). We performed a simulation based on the observed distance distribution and found no enrichment between DNM and inferred positions of meiotic recombination events (Figure A.11). This study provides a near-complete picture of the occurrence of DNMs with respect to parental homolog and meiotic recombination.

2.4 Discussion

We identified 195 *de novo* SNVs and indels in the quad: 88 in the proband and 107 in the sibling—a 35% increase from the 65 and 79 DNMs identified from our previous analysis of this family [189], which was optimized for specificity and similar to the original estimates of Kong et al. [91]. This contrasts with a second analysis we performed on the same family using a two-caller Illumina approach, which was geared toward increased sensitivity and reported 131 and 138 DNMs for the proband and sibling, respectively [176]. A comparison with both ONT and HiFi data for the same family, however, shows that this two-caller method demonstrates only a 78% true positive rate. Thus, while the overall numbers appear similar, the long-read data are discovering a new subset of DNMs traditionally excluded or filtered by the short-read data. Notably, this study also widens the gap in the number of DNMs present in the proband and sibling: the total 19 DNM differential would place this quad in the 91st percentile among twins in Wilfert et al. [189] and the 60th percentile in Turner et al. [176]. (Figure A.12)

Of the 195 true DNMs that we identified, exactly seven lie on the X chromosome. Calling on the X chromosome presented a unique challenge, as we failed to identify a single DNM

on the female proband's X chromosomes. Despite there being many potential sites on the proband's X, most failed to reach the minimum allele balance threshold, and those that did fail to meet our read depth requirements, resulting in only four calls in our final callsets, none of which were true *de novo* events. The male sibling, on the other hand, showed the opposite trend, with 20 DNM calls in our final callset, seven of which were validated based on our criteria. The high number of calls in the sibling were the result of HiFi GATK, which alone contributed 11 sites on the X chromosome. It is likely that the increased sensitivity on the male X chromosome is an artifact of the lower average read depth—with lower read depth, one or two sequencing errors would pass the allele balance threshold, allowing a variant call to be made, as we did not set a higher threshold for AB on the male X chromosome.

In the past, we relied on both the two- and four-caller pipelines to identify *de novo* variation [176, 189] from autism WGS datasets in an effort to balance both specificity and sensitivity. However, this study revealed that both pipelines suffer from limitations that can make it more difficult to identify potential disease-causing variation. Across GRCh38 and T2T-CHM13, the four-caller pipeline identified 141 true DNMs with a 91.6% true positive rate, which is the highest of all callsets, but lower than our previous validations had estimated. The four-caller pipeline underestimated the number of DNMs in this family by at least 30%. The two-caller pipeline, on the other hand, suffers from both a low true positive rate (72.1%) and when restricted to just true positive sites, underestimates the number of DNMs in this family by at least 25%. The high false positive rate in the two-caller pipeline makes it a poor choice for studies aimed to identify disease-causing mutations. Going forward, the Illumina four-caller pipeline could be optimized by removing the mask used to exclude variants in repetitive sequence and developing a new filtering schema that does not remove variants based on region alone. By removing this mask, the size of the four-caller callset could be increased by at least 20% and would provide a better snapshot of the true *de novo* variation present in a genome.

The increase in the total number of true mutations relative to previous studies indicates that the DNM rate for SNVs and indels is likely higher than current estimates suggest. In

addition, we were able to document DNMs in centromeres and segmental duplications, two regions that are just beginning to become accessible with long-read WGS platforms. If we set the total accessible genome of our study to be the total number of 10 kbp regions with high mapping quality by HiFi (mapq $\geq$ 57) in T2T-CHM13, the total genome size is 3.07 billion bp. Based on this, we can estimate the *de novo* substitution rate from this one family to be approximately 1.41×10^{-8} per nucleotide per generation, which is on the high end of projected mutation rates [61, 81, 85, 149]. While the rate is higher than most previous genome estimates, a larger number of samples will be needed to determine if the mutation rate is particularly elevated in these regions of the genome. It should be stressed, however, that the methods we used to identify *de novo* SNVs with long reads were still stringent, requiring an allele balance of between 0.3-0.7 and confirmation across two sequencing platforms. Different sequencing platform biases even among the long-read technologies will tend to underestimate variant calling for specific regions and classes of variation. Based on our false negative analysis, which revealed that there are 23 true *de novo* (likely post-zygotic) SNVs that eluded our calling pipelines, the mutation rate in this family is likely closer to 1.59×10^{-8} per nucleotide per generation.

In order to replicate these results, more families need to be sequenced with orthogonal long-read sequencing approaches. Despite the additional cost, long-read sequencing enabled us to increase our DNM discovery by 35%, granted us access to new regions of the genome, and allowed us to search for and verify larger mutations as well. While this study focused on DNMs, we also performed a preliminary analysis of inherited rare variants (<0.1%) by re-filtering the GATK callsets to search for variants that were present in a child and exactly one of the parents and confirmed by ONT (Figure A.13). Although comparison of larger number of long-read genomes will be required to determine true allele frequencies, this analysis suggests a comparable increase (26.4% in GRCh38) of inherited rare variants throughout the genome (Figure A.14) as result of greater access to more complex and repeat rich regions of the genome. It should be noted that the number of sequencing platforms used in this study is not necessary to extensively catalog the DNM load in a trio—a combination of

Illumina, ONT, and HiFi would be sufficient, and most potential mutations could be validated with much more affordable Sanger sequencing. DNM calling can be further optimized by aligning both short and long reads to the more complete T2T-CHM13 genome, which will be invaluable for estimating the mutation rate in repetitive regions of the genome.

In addition to our large *de novo* SNV and indel callset, we discovered three *de novo* STR expansions and one VNTR expansion (<50 bp) (Table A.4). This is fewer STR events than we would expect based on previous projections that predict greater than 50 STR expansions per transmission [122]. We also identified two small *de novo* SVs, insertions of 52 and 73 bp but did not observe any larger germline SVs. Despite their importance in neurodevelopmental disease, this more comprehensive DNM analysis did not reveal any new candidate mutations to better explain the proband's autism status. It could be the case that the underlying etiology is inherited or polygenic as the upper bound for DNM underlying autism has been estimated to account for 30% of cases [76]. Alternatively, if the causal variant is *de novo*, this may mean that despite our many orthogonal methods to identify DNMs in the proband, sensitivity is not yet optimized. This is especially the case for *de novo* structural variation where methods for *de novo* SV calling among VNTRs/STRs remains a challenge despite the dramatic advances in SV detection using long reads [18, 136]. Another possibility, albeit less likely, is that we missed a causative mutation in a region that we are still not able to sequence and assemble, such as in the highest identity repeats. Even when we align HiFi reads to the T2T-CHM13 genome, there is still approximately 200 Mbp of sequence with coverage more than 2 standard deviations above or below the mean and with mapq less than 57; until we can accurately assign sequencing reads to these regions of the genome, we will not be able to fully catalog all classes of variation in a genome.

An important aspect of this work was that all DNM candidates were obtained from primary tissue, in this case peripheral lymphocytes from blood. It is worthwhile noting, however, that apparent *de novo* SVs were initially identified using other technologies including Strand-seq, 10X Genomics and Bionano Genomics where lymphoblastoid cell culture instead of primary blood were used to obtain larger amounts of DNA or actively dividing

cells for the assay. The most striking was a 147 kbp deletion event in the sibling removing all but the first exon of SETD2 (chr3-47014726-DEL-147102)—a gene previously associated with autism. This deletion was discovered from Bionano SV calls and confirmed by 10X data, but in both cases, calls originated from DNA prepared from cell line material. We found no evidence for the variant in any sequence data derived from blood DNA, including read-depth changes in Illumina and PacBio (Figure A.15A). Similarly, a second event, a 158 kbp deletion in the sibling (chr7-142644005-DEL-158289), deleted several small genes (PRSS1 and PRSS2) and had support from all three cell-line-derived datasets (Bionano, 10X, and Strand-seq), but no datasets derived from blood DNA (Figure A.15B). While undetectable low levels of somatic mosaicism in the blood DNA may underlie this, it is more likely that such potentially impactful deletion events occurred in cell culture after only a few passages. This emphasizes the importance of discovery and validation of DNM variants from primary source material.

2.5 Methods

Illumina sequencing and microarray data

Illumina WGS was performed on the Simons Simplex Collection (SSC) samples of a family (14455) at the New York Genome Center (NYGC) using 1 μg of DNA, an Illumina PCR-free library protocol, and sequencing on the Illumina X Ten platform. The father, mother, proband, and sibling were sequenced to an average depth of 37.47, 33.50, 41.78, and 32.90, respectively. Post-sequencing, reads were initially aligned to the reference genome (GRCh38). We applied two different single-nucleotide variant (SNV) callers—FreeBayes and GATK. We also applied a suite of structural variant (SV) callers (details in Turner et al., 2017) to maximize sensitivity for *de novo* SV mutation detection of various size ranges [176]. This family was selected because we found only two variants of interest (likely gene-disrupting, missense (CADD>30), 3' UTR, or pNCR-TFBS) and no *de novo* or inherited SVs in exonic regions. The two *de novo* variants were a 3' UTR event in ATP9A in the proband and a 3'

UTR in OLFM3 in the sibling (neither are candidate autism genes at this time). In addition, Illumina whole-genome sequencing (WGS), whole-exome sequencing (WES) [92], and single-nucleotide polymorphism array data [161] were generated as part of the SSC phase 1 study [47]. No *de novo* variants were identified as likely pathogenic [176]. This family, then, was selected for long-read sequencing for two reasons: 1) the autism case was unsolved for rare variants of large effect and 2) the unaffected and affected individuals represent fraternal twins where the female sibling was affected with autism.

Pacific Biosciences continuous long-read (CLR) sequencing.

DNA from blood was sheared with a Megaruptor (Diagenode) on the 50 kbp setting. Material was prepared for PacBio sequencing using the SMRTbell Template Prep Kit 1.0 (PacBio P/N 100-259-100) (TPK1) or SMRTbell Express Template Prep Kit (P/N 101-357-000) (ExV1) following the recommended protocols. Briefly, the sample is treated to remove single-stranded overhangs, damage repaired, end prepared, and ligated to PacBio SMRTbell adapters. TPK1 libraries have an additional step to remove imperfect SMRTbell templates, which is omitted in the ExV1 protocol. SMRTbell libraries were size selected on the BluePippin system (Sage Science) at 30 kbp or 40 kbp high pass settings. Libraries were quantified with Qubit (Thermo Fisher Scientific) and sized with FEMTO Pulse (Agilent Technologies) instruments before loading on the PacBio Sequel System using version 2.1 or 3.0 chemistries with 10-hour movie acquisition times.

Pacific Biosciences high-fidelity (HiFi) sequencing

DNA from blood (all family members) or cell culture (mother, father for additional coverage) was sheared to a tight distribution with peak size of 10 or 20 kbp using gTUBEs (Covaris). SMRTbell libraries were prepared with TPK1 as described above (proband, sibling, blood; mother, father, cells), or SMRTbell Express Template Prep Kit 2.0 (P/N 100-938-900) and SMRTbell Enzyme Clean up Kit (P/N 101-746-400) (mother, father, blood), and size fractionated on the SageELF (Sage Science) to generate tightly sized fractions. The fraction

sized at 13 kbp (proband, sibling, blood), 15 kbp (mother, father, cells) or 20 kbp (mother, father, blood) was chosen for sequencing on the Sequel II system with version 1.0 (proband, sibling, blood), version 2.0EA (mother, father, cells), or version 2.0 (mother, father, blood) chemistries and 30-hour movies. Raw subreads were processed through the CCS workflow (PacBio SMRTLink version 7.1) to generate HiFi reads with a minimum estimated quality value (QV) of 20 (phred scaled, corresponding to an accuracy of 99%).

Oxford Nanopore Technology (ONT) sequencing

DNA from the same blood-derived DNA aliquot used for HiFi sequencing was sheared to 50 kbp using a Diagenode Megarupter following manufacturer's recommendations. DNA was size selected via the Circulomics small read eliminator 25 kbp kit. DNA was prepared for Nanopore sequencing using the ONT 1D sequencing by ligation kit (SQK-LSK109). Briefly, 1-1.5 μg of fragmented DNA was repaired with the NEB FFPE repair kit, followed by end repair and A-tailing with the NEB Ultra II end-prep kit. After an Ampure clean-up step, prepared fragments were ligated to ONT-specific adapters via the NEB blunt/TA master mix kit. The library underwent a final clean-up and was loaded onto a PromethION PRO0002 flowcell per manufacturer's instructions. The flowcell was sequenced with standard parameters for 3 days and generated a Read N50 >32 kbp for all samples. Base calling was performed with Guppy version 5.0.7 super accuracy model.

Strand-seq

Strand-seq libraries were prepared from four lymphoblast cell lines: SSC11453 (father), SSC11165 (mother), SSC11168 (sibling), and SSC11163 (proband). All lines were maintained in RPMI-1640 with 10% FBS, 1% Glutamax and 1% penicillin/streptomycin. BrdU (Bromodeoxyuridine; Sigma, B5002) was added to log-phase cell cultures at 40 μM or 100 μM concentrations for a period of 18 or 24 hours. Single nuclei were prepared and sorted using the BD FACSMelody cell sorter into 96-well plates for Strand-seq library production, as described (Falconer et al. [46]; Sanders et al. [160]). The Strand-seq protocol was im-

plemented on a Biomek FXP liquid handling robotic system, and pooled single-cell libraries were sequenced on the NextSeq500 platform (MID-mode, 75 bp paired-end protocol). After demultiplexing, Strand-seq sequencing reads were aligned to the human reference assembly GRCh38 (GCA_000001405.15_GRCh38_no_alt_analysis_set.fna) using the default parameters of BWA-MEM (version 0.7.15-r1140). Aligned BAM files were sorted by genomic position using SAMtools (version 1.7) and duplicated reads marked using sambamba (version 0.6.6). After alignment, each single library was evaluated to select only high-quality Strand-seq data for downstream analyses. Specifically, libraries with visible background reads (i.e., reads mapped to opposite direction on chromosomes that inherited template strands with the same directionality) and libraries with low (<50,000 reads) or uneven coverage were excluded, as detailed previously (Sanders et al. [160]; Porubský et al. [146]).

Bionano Genomics

Optical genome mapping was performed as described previously (Mantere et al. [113]). Briefly, ultra-high molecular weight (UHMW) gDNA was isolated from frozen cell pellets, harvested from (EBV)-immortalized lymphocyte cell lines, following the manufacturer's guidelines (Bionano Prep SP Frozen Cell Pellet DNA Isolation Protocol, Bionano Genomics #30268). For each sample, 750 ng of purified UHMW gDNA was labelled with DL-green fluorophores using the Direct Labeling Enzyme (DLE-1) chemistry and cleaned up with membrane adsorption [Bionano Prep Direct Label and Stain (DLS) Protocol, Bionano Genomics, #30206]. Labelled gDNA samples were loaded on the Saphyr chips for linearization and imaging on the Saphyr instrument. Each flowcell was run on the maximum capacity to generate 1300 Gbp of data per sample using Hg38 as the reference. The *de novo* assembly and SV annotation pipeline were executed with Bionano Solve software v.3.4. Fractional copy number estimates were based on the coverage-based CNV-tool and visual inspection of the events.

10X Genomics linked-read sequencing

High molecular weight (HMW) DNA was extracted from 1 million cells following a protocol outlined by 10X Genomics utilizing the salting out method. DNA was isolated with a Qiagen MagAttract HMW DNA kit, resulting in >80 kbp DNA fragments. The HMW DNA was diluted to 1 ng/μL prior to the v2 Chromium™ Genome Library prep (10X Genomics). Approximately 10-15 DNA molecules were encapsulated into nanoliter droplets. DNA molecules within each droplet were tagged with a 16nt 10X barcode and 6nt unique molecular identifier during an isothermal incubation. The resulting barcoded fragments were converted into a sequence-ready Illumina library with an average insert size of 500 bp. The concentration of each 10X WGS library was accurately determined through qPCR (Kapa Biosystems) to produce cluster counts appropriate for the NovaSeq6000 platform (Illumina). Paired end-sequence (2x150) data were generated on a S4 300 cycle kit utilizing the XP workflow (Illumina) targeting 60X coverage (190 Gbp) providing long linked reads across the length of individual DNA molecules.

Accessible genome

The total accessible genome is based on the number of 10 kbp regions with an average mapq >57 in both Illumina and HiFi. These calculations are based on previously described methods (Nurk et al. [132]), aligning CHM13 Illumina and HiFi reads to both GRCh38 and T2T-CHM13 assemblies. This results in a haploid genome size of 3.02 billion bases in T2T-CHM13 based on HiFi read alignment vs. 2.63 billion bases in T2T-CHM13 based on Illumina sequence read alignment.

SNV and indel variant calling with HiFi

HiFi reads were aligned to the reference (either GRCh38 or T2T-CHM13) using minimap2 to generate a BAM file [101]. These BAM files were then used in a bifurcated pipeline to call *de novo* variants. For the DeepVariant portion of the pipeline, we applied DeepVariant

v1.0.0 to call variants for each individual, and then merged these variant files using GLnexus [139, 196]. After calling, variants were filtered using GATK VariantFiltration, removing all calls with Phred-scaled quality score (QUAL) < 30.0. In addition, BCFtools was used to left-align and normalize indels. From this filtered set of variants, potential *de novo* variants were initially identified based on genotype (father and mother genotypes were equal to $0/0$ and the child's genotype was equal to $0/1$ or $1/1$). For each *de novo* call, Pysam was used to count the number of reads with reference and alternate alleles in the BAM files, in order to ensure the depth and allele balance were correct for each individual. Lastly, we used the following sample-level filters: father depth >10, mother depth >10, child depth >10, and child genotype quality (GQ) >20. Any remaining indels with length greater than 20 bp were also removed. For the GATK portion of this pipeline, we applied GATK HaplotypeCaller v4.0.0 to call variants for each individual, and then jointly genotyped the output using GATK GenotypeGVCFs [140]. After calling, variants were separated into three groups used GATK VariantFiltration to filter for quality of depth (QD), QUAL, and the z-score for the Mann-Whitney Rank Sum Test for the position of the alternate allele on reads (ReadPosRankSum). Specifically, the three groups were: SNVs (QD<2, QUAL<30, ReadPosRankSum<-8.0), 1-2 bp indels (QD<8, QUAL<30, ReadPosRankSum<-20.0), and 3+ bp indels (QD<2, QUAL<30, ReadPosRankSum<-20.0). The three groups of variants were then merged, and BCFtools was used to left-align and normalize indels. From this filtered set of variants, potential *de novo* variants were initially identified based on genotype (father and mother genotypes were equal to $0/0$ and the child's genotype was equal to $0/1$ or $1/1$). For each *de novo* call, Pysam was used to count the number of reads with reference and alternate alleles in the BAM files, in order to ensure the depth and allele balance were correct for each individual. Lastly, we used the following sample-level filters: father depth >10, mother depth >10, child depth >10, child allele balance >0.25, and child GQ >20. Any remaining indels with length greater than 20 bp were also removed. For reads aligned to T2T-CHM13, a final filtering step was applied to remove all GATK calls within 2-100 bp of each other. This filter removed 24 calls (17 from the proband, 7 from the sibling) from the final T2T

GATK callset.

SNV and indel calling with Illumina WGS

We called SNVs and indels in families using four different callers and two different pipelines that used two (GATK and FreeBayes) or all four of the callers as previously described [176, 189]. Specifically, we applied GATK HaplotypeCaller v.3.5.0 FreeBayes v1.1.0, Platypus v0.8.1, and Strelka2 v2.9.2 [56, 88, 140, 156]. In addition, MNVs were called using FreeBayes and Platypus. Post-calling BCFtools (version 1.3.1) norm was used to left-align and normalize indels. We partitioned the genome into the high-quality (HQ) regions, consisting of unique space as well as ancient repeats, and the recent repeat (RR) regions, which consisted of repeats <10% diverged from the consensus in RepeatMasker. Variants were only assessed in HQ portions of the genome and the RR region variants were removed from the study. Qualities of the callsets were assessed using KING for relationship checks, a variant per chromosome counter, and concordance checks for individuals with available array data [112].

De novo variants were called using a custom pipeline. First, variants that were *de novo* based on genotype (father and mother genotypes were equal to 0/0 and the genotype in the child was 0/1 or 1/1) were retained for further assessment. Second, variants from Platypus with a filter of LowGQX or NoPassedVariantGTs were removed and Strelka2 variants had to have the filter field equal to PASS. Third, variants needed to have the support of at least two of the four callers. Fourth, variants were regenotyped with FreeBayes using default settings and needed to remain as *de novo*. Fifth, variants in a homopolymer A or T of length 10 or greater were removed. Sixth, we removed all variants in low-complexity regions, recent repeats, or centromeres. Finally, we applied the following sample level filters: the father alternate allele count = 0, mother alternate allele count = 0, child allele balance >0.25, father depth >9, mother depth >9, child depth >9, and either child GQ >20 (GATK) or sum of quality of the alternate observations (QA) >20 (FreeBayes). For variants on the X chromosome, we separately considered variants in the pseudoautosomal regions

(chrX:10000-2781479, chrX:155701382-156030895) and the X/Y duplicatively transposed region (chrX:89201803-93120510).

SNV and indel validation

Previously, we performed random Sanger validation of both the four-caller and two-caller DNM callset and combined this data with published validations to look at a total of 3,233 sites in a conditional inference analysis [189]. We estimated our validation rate in this dataset at 99.5% and our false negative rate at 3.5%.

In this study, SNVs and indels were validated by examining the site across three different sequencing platforms—ONT, HiFi, and Illumina—aligned to the T2T-CHM13 reference (Figure A.1). We used Pysam to calculate the number of reads with the reference and the alternate allele from the BAMs of reads aligned to the reference and used it to make the DNM call for both the parents and the child with the mutation. ONT reads were filtered to exclude any reads with base call QV <10 at the site of the *de novo* variant—any site with more than one ONT read with the alternate allele in a parent was deemed inherited, and any site with fewer than one ONT read with the alternate allele in the child was deemed a false positive. Sites were further examined across the other two sequencing platforms—in regions of read depth within two standard deviations of average, any site with one or more HiFi reads, or any site with several Illumina reads, with the alternate allele in a parent was deemed inherited.

SNV and indel phasing

SNVs and indels were phased by applying WhatsHap v1.0 to aligned reads generated by PacBio HiFi, ONT, and Illumina sequencing [115]. SNVs were phased in 40 kbp windows around a DNM of interest. We then used a Python script to select nearby "informative" SNVs of unambiguous parental inheritance (for example with genotype 0/0 mother, 0/1 father, and 0|1 in the child), and omit all SNVs that could not be phased or assigned to a parent. These informative SNVs were then used to determine a maternal or paternal haplotype around the

DNM of interest by using an average haplotype score weighted inversely proportional to the distance from the DNM (nearby sites were weighted more highly in cases of disagreement between haplotype inheritance). This method was able to phase 194/195 (99.5%) germline DNMs in our dataset, as well as 23/26 (88.5%) potential mosaic likely postzygotic DNM sites.

Assembly-driven detection of de novo structural variation

Assemblies for each member of this family were generated using hifiasm v0.12 [21], which leverages PacBio HiFi reads generated from blood to produce haplotype-resolved assemblies. In the case of the children, Illumina short-read data was used to assign these haplotypes to a parent of origin. Phased assembly variant (PAV) caller [43] was used to detect SVs, indels, and SNVs in these assemblies. Detection of variants is driven by assembly to reference alignments using minimap2 v2.17 (CIGAR) and analysis of the CIGAR string. SVs were then supported by an additional run of PAV using Long Read Aligner (lra) [153], PBSV [43], subseq [43], and DeepVariant [139]. Variants with detection from two or more callers (PAV [minimap2] + support caller) were then carried through to the final callset. Using this procedure, 222 candidate *de novo* SVs were identified consisting of 57 deletions and 165 insertions. After these candidates were identified, alignment of parental data was computationally analyzed using subseq over the SV region in order to determine read-based support for these events, which might have been missed in the assemblies.

STR/VNTR characterization

We applied three methods to focus specifically on *de novo* STRs/VNTRs—two of which have been previously described (Dolzhenko et al. [35], Sulovari et al. [173]). We also developed a custom k-mer based pipeline that defines a library of uniquely mappable 30 bp k-mers (i.e., 30-mers) from the set of 21,000 phased tandem repeats of the human genome defined in Sulovari et al. [173]. We used seqtk to create the reverse complement of each haplotype-resolved STR and VNTR sequence in our library of tandem repeat sequences, followed by

generating all possible 30-mers using jellyfish [116]. All 30-mers were aligned using mrsFAST v3.3.8 and each was deemed uniquely mappable if and only if it mapped unambiguously to at most one specific genomic location of the unmasked GRCh38.p12 reference after allowing a hamming distance of two (i.e., command-line option -e 2) [63]. The vast majority of 21,000 polymorphic tandem repeats had at least one uniquely mappable k-mer associated with them. Next, we count each of the uniquely mappable k-mers in the Illumina short-read BAM files of each sample using VariantBam [185] to pull down reads matching the k-mer sequence and Kanalyze [6] for counting the number of specific k-mers (and their reverse complements using –reverse=canonical option) across the short reads, irrespective of read mapping information and including both mapped and unmapped reads in our search space. Importantly, the information between each k-mer sequence and the GRCh38 coordinates of the STR/VNTR contigs that they originated from were stored throughout the process. After normalizing the k-mer counts by sequencing depth and GC bias coefficient (as described by Sudmant et al. [172]), the adjusted k-mer counts become a proxy for repeat length. The sites that appeared to have a significantly higher number of k-mer counts in either child relative to both parents were subsequently investigated as putative *de novo* sites. The CLR reads and long-range phasing information from 10x were used to carry out a targeted phased assembly for each locus with a putative *de novo* STR/VNTR [173]. The loci where the targeted assembly results supported the existence of putative DNM underwent PCR validation.

Cell line artifacts

In the process of identifying *de novo* SVs, we found evidence for several cell-line artifacts. We discovered these events in early exploratory phases of the project using merged callsets from PacBio CLR (Phased-SV, SMRT-SV, PBSV), 10X (LongRanger v2.2.2), and Bionano (assembly-based calls from Solve). Although SV callsets were not generated from Strand-seq, it was used to find orthogonal support for variant calls. We applied subseq [43] to find support for variants in aligned CLR reads for all family members to validate both the original variant call and inheritance status. Briefly, subseq expands a window around each variant,

finds all reads in the region, and determines the length of the read spanning the region, which will be longer than the reference for insertions and shorter for deletions. This allows us to sensitively identify support for SVs down to two or more reads. We then selected *de novo* SVs with concordance from more than one technology and manually inspected them for supporting evidence across callsets, subseq, Strand-seq. We considered genomic location, such as VNTRs and high-identity segmental duplications, which may have led to false variant calls, and we looked for clusters of SVs commonly associated with poor mapping, false calls, and poor reproducibility. Most *de novo* SVs could be explained by a missing parental allele or poor SV quality. However, we found several SVs that were strongly supported using sequence data originating from cell-line-derived sources (10X, Strand-seq, Bionano), but clearly lacked support in blood-derived sources, such as CLR (by SV discovery and subseq) and Illumina (by WSSD).

Detection of meiotic recombination breakpoints

In order to detect meiotic recombination, we realigned demultiplexed Strand-seq reads to the human reference assembly CHM13v1.0 using default parameters of BWA-MEM (version 0.7.17-r1188). Aligned BAM files were sorted by genomic position using SAMtools (version 1.10) and duplicated reads marked using sambamba (version 1.0). Next, we phased Strand-seq reads using StrandPhaseR using default parameters used for Illumina paired-end reads [143]. We proceeded with integrative phasing by merging long-range Strand-seq haplotypes with local PacBio phasing, embedded in each long-read, using WhatsHap (versions 0.18) [143]. Having chromosome length and dense haplotype for all family members, we set to detect all recombination breakpoints as positions where a child's haplotype switches from matching H1 to H2 of a given parent or vice versa. In order to detect these positions, we first established what homolog in a child was inherited from either parent by calculating the level of agreement between child's alleles and homozygous variants in each parent. Next, we compared each child's homolog to both homologs of the corresponding parent and encoded them as 0 or 1 if they match H1 or H2, respectively. We applied a circular binary segmenta-

tion algorithm on such binary vectors using R function 'fastseg' implemented in R package fastseg (version 1.36.0) with parameter 'minSeg' set to 50 and 1000 for high-sensitivity and high-specificity breakpoint detection, respectively. Detected regions with segmentation mean $\leq$0.25 have been assigned H1 while regions with segmentation mean $\geq$0.75 have been assigned H2. Regions with segmentation mean in between these values were deemed ambiguous and were excluded. In addition, we filtered out regions shorter than 500 kbp and merged consecutive regions assigned the same haplotype.

Chapter 3

LONG-READ SEQUENCING REVEALS INCREASED GERMLINE AND POSTZYGOTIC MUTATION RATES IN REPETITIVE DNA

This work in preparation was performed by:

Noyes, MD, Harvey, WT, Sui, Y, Munson, KM, Hoekzema, K, Kordosky, J, Garcia, GH, Knuth, J, Lewis, AP, Zody, MC, Eichler, EE

3.1 Abstract

While most *de novo* mutations (DNMs) arise in the parental germline, there are several postzygotic mutations (PZMs) that contribute to the overall mutational burden of an individual. These PZMs are difficult to distinguish from germline mutations with short-read sequencing data, but long reads can provide the haplotype information necessary to discern between the two. We used three sequencing technologies—Illumina, Oxford Nanopore Technologies, and Pacific Biosciences—to discover DNMs in 42 samples across 24 trios. We found a total of 88.2 *de novo* single-nucleotide variants (SNVs) and 7.8 indels per sample, including an average of 13.1 postzygotic SNVs per sample. Approximately 5% of all DNMs were identified on the sex chromosomes, including five SNVs and one indel on the Y chromosome. This constitutes an approximately 40% increase in DNM discovery over previous Illumina-based studies of the same samples, as well as the first characterization of postzygotic variation in these samples. In the 91% of the genome we were able to assess, we calculated a germline mutation rate of 1.31×10^{-8} and a PZM rate of 0.23×10^{-8} substitutions per base pair per generation. In segmental duplications, we observed enrichments of both germline (38% increase) and postzygotic (threefold increase) SNVs, averaging for an overall 66% increase in *de novo*

variation in duplicated regions. We show that long-read data not only increase sensitivity for germline variant discovery but also help to discern postzygotic variation, both in unique and repetitive sequence.

3.2 Introduction

De novo mutations (DNMs) are variants unique to a child that are absent from the parents. We typically think of DNMs as arising from mutational processes in the parental germline, such as double-stranded break repair or errors in DNA replication [61, 81, 85, 91, 162]. In addition, a small number of *de novo* variants arise in the rounds of cell division just after fertilization, early enough in development that they can still be detected without tissue-specific sequencing. The most common type of DNMs are single base-pair substitutions and small (<50 bp) insertions and deletions; previous short-read DNM discovery efforts report an average mutation rate of at least 70 DNMs per individual per generation, over three quarters of which arise in the paternal germline [61, 91, 162]. Postzygotic mutations (PZMs) are challenging to identify and validate, and the PZM rate has been estimated to be up to 10% of the germline rate [1, 36, 107, 162, 191].

Most studies of *de novo* variation have used short-read Illumina data to both identify variants and assign them to parental haplotypes, which can then be used to infer if a variant is germline or postzygotic in origin. In a trio, parental haplotype assignment is performed using informative single-nucleotide polymorphisms (SNPs), or alleles that can be uniquely traced to one parent or the other. We expect there to be one such site approximately every 1,000 bp in the genome, meaning that 150 bp Illumina reads can only be used to phase fewer than 20% of DNMs [61]. Long-read sequencing can capture many of these informative SNPs on a single read and can, therefore, be used to phase more than 95% of DNMs and distinguish more postzygotic from germline variants [72, 94]. Another limitation of short-read sequencing is that the reads cannot align unambiguously in repetitive regions, restricting the search space for DNMs to as little as 84% of the genome [174]. Some DNM studies exclude these regions altogether, neglecting variation in sequences such as segmental duplications (SDs), which are

thought to be more mutable than unique sequences [41, 131]. Repeat hypermutability is in part due to interlocus gene conversion, a process by which breaks in DNA are repaired using paralogous sequence as a template, thereby creating a new allele in the original sequence that matches the allele in its paralog [39]. Other unique mutational processes are likely at play in duplicated sequences too, as their mutational profile significantly differs from unique sequence, including a depletion of CpG>TpG substitutions, the most common of all mutation types [180].

In this study, we set out to comprehensively identify human germline and postzygotic *de novo* variation across 42 children from 24 families. We leveraged long-read Pacific Biosciences (PacBio) high-fidelity (HiFi) sequencing data derived from blood for variant discovery, and both long-read Oxford Nanopore Technologies (ONT) and short-read Illumina data for validation. These families are part of the Simons Simplex Collection [47], and each has a proband affected by simplex autism. They have been examined for *de novo* variation before using Illumina whole exome sequencing and whole genome sequencing data [4, 51, 176, 189] but were selected for long-read sequencing because no genetic cause had been confirmed. Here, we quantify both germline and postzygotic DNMs in each child. Using long-read sequencing, we expect to recover more PZMs than previous Illumina-based studies, increasing our estimate of the PZM rate, perhaps at a cost to the germline mutation rate. Further, we can evaluate the germline mutation and PZM rates across different genomic regions, quantifying the enrichment of DNMs in hypermutable repetitive sequences.

3.3 Results

3.3.1 De novo and postzygotic single-nucleotide variants (SNVs) on the autosomes

We examined HiFi data derived from blood and cell lines for 90 total samples from 24 families affected with simplex autism (18 quads and 6 trios, n=42 transmissions) for *de novo* variant discovery (Figure 3.1). With T2T-CHM13v2.0 as a reference genome, we used two variant callers to identify SNVs, selecting those unique to a child for validation with two

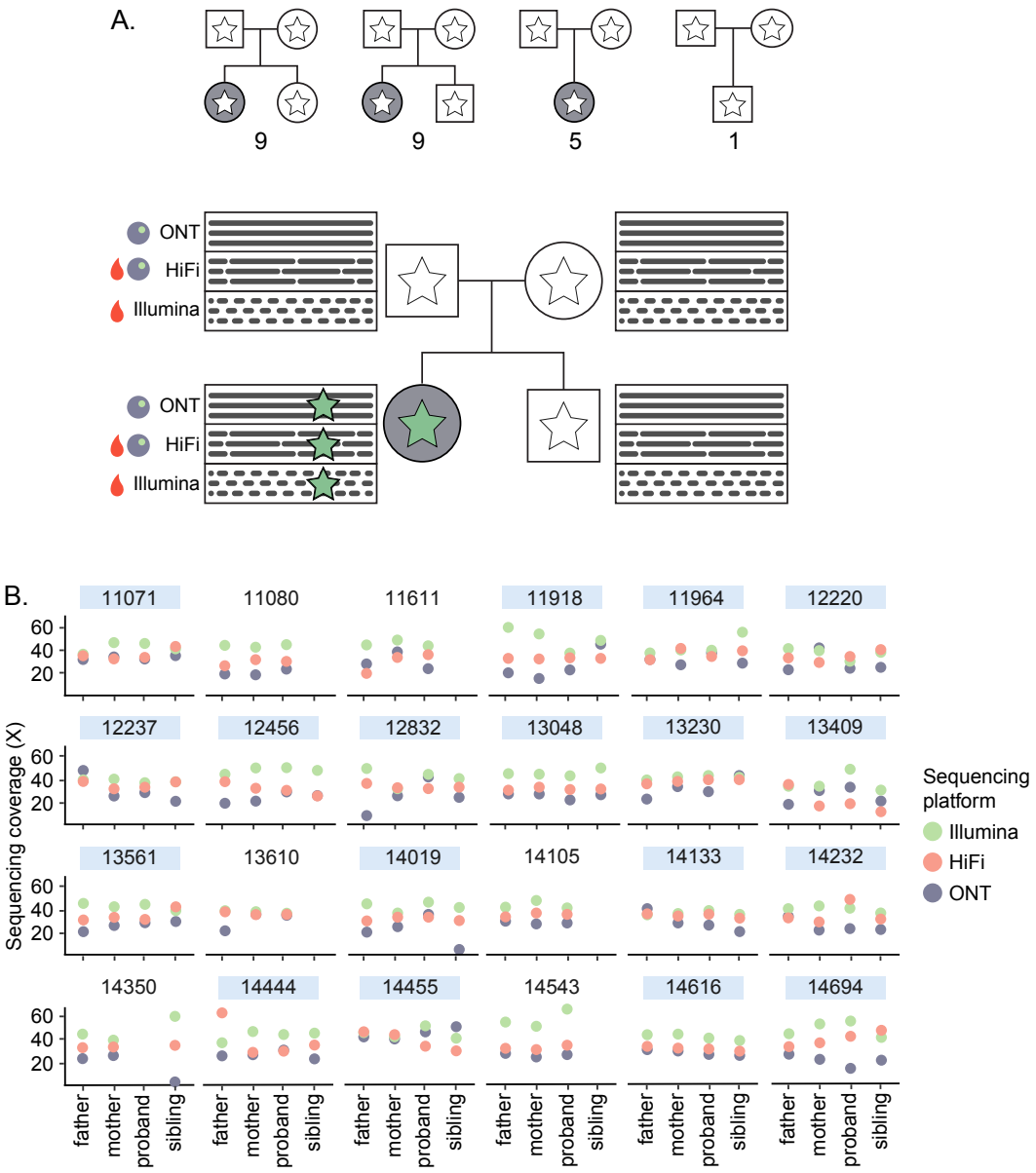


Figure 3.1: **Overview of the dataset.** A. This dataset is composed of 24 quads and trios, and all but one have a proband affected with simplex autism. Each family is sequenced with ONT (cell line derived DNA), PacBio HiFi (blood DNA topped off with cell line), and Illumina (blood DNA). B. Sequencing depth of coverage across all three platforms for every sample. Quads are highlighted in blue, and trios are not.

orthogonal sequencing technologies: ONT and Illumina. For a variant to pass validation, we require that it be observed in a child's blood-derived HiFi reads in addition to their ONT or Illumina reads, and that it be absent from both parents across all three types of read data. Additionally, every validated *de novo* variant is unique to the sample in which it was called; any variant observed in the HiFi reads of unrelated individuals was assumed to be either a segregating allele that went unsequenced in a parent or a recurrent sequencing error. We found 3,505 *de novo* SNVs that passed our validation criteria (Methods), for an average of 83.5 SNVs/child.

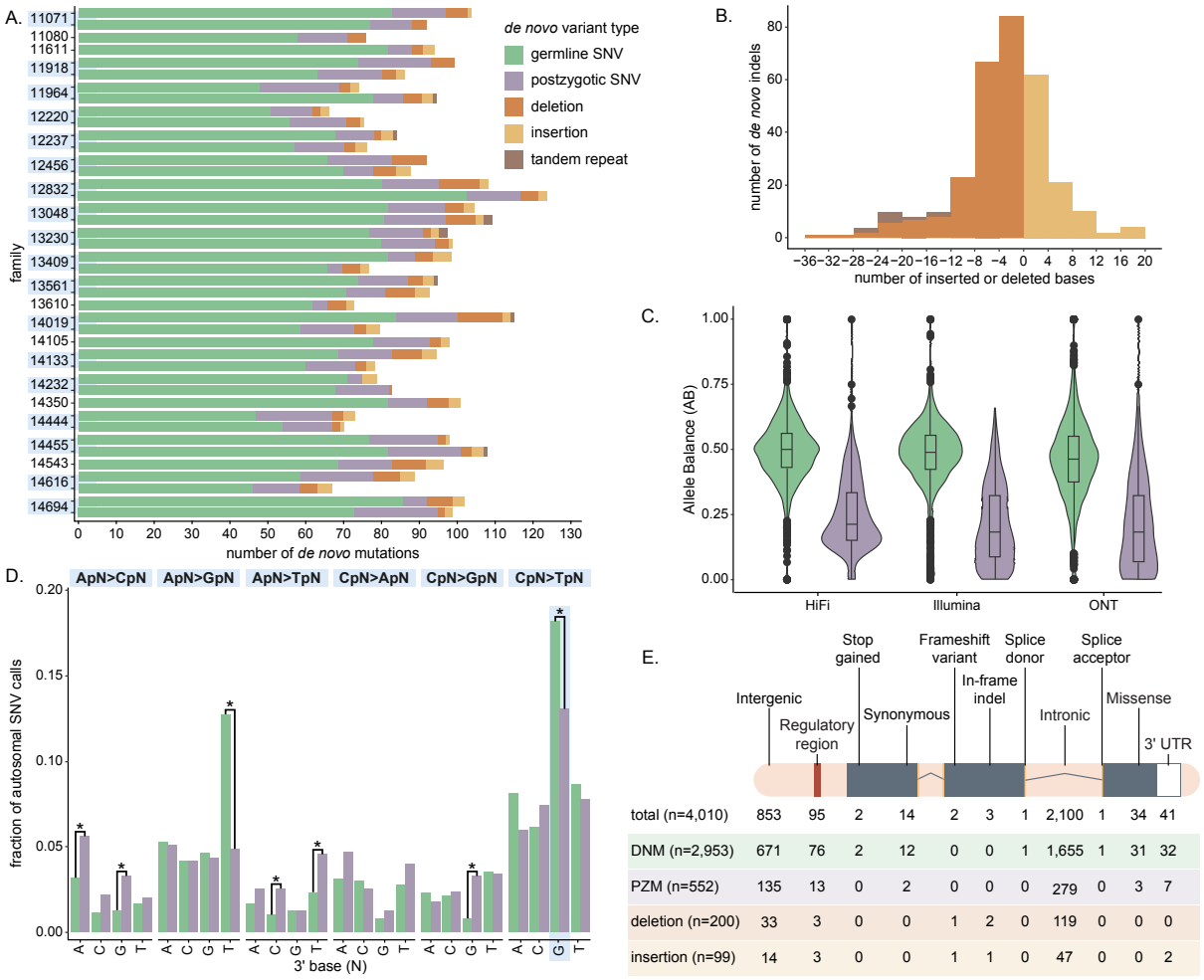
We determined whether these SNVs were germline or postzygotic in origin first by haplotype assignment. A variant was considered to be germline if it was present in all the HiFi and ONT reads from its parental haplotype of origin, and cases of conflict were resolved by examining the allele balance (AB), or the fraction of reads with the *de novo* allele, across all three sequencing platforms (Methods, Figure B.1). We determined that 552 SNVs were postzygotic in origin, because they were only present on a fraction of reads from a given parental haplotype. The remaining 2,953 SNVs likely arose in the parental germline (Figure 3.2A). The average germline mutation has AB=0.48 across the sequencing data generated by all three technologies, while the average PZM has AB=0.22 across the sequencing platforms (Figure 3.2C, Figure B.2). Although HiFi and Illumina data were generated from blood and ONT data from cell lines, 79.2% of PZMs and 87.7% of DNMs do not have significantly different AB across platforms (chi-squared test, $p>0.05$). A small number (n=7) of germline DNMs have AB=1 across HiFi and at least one other sequencing platform, and all but one of these events fall in repetitive regions (5 in retrotransposable elements, 1 in satellite DNA).

To confirm that our germline mutations and PZMs were the result of different mutational processes, we first examined the parental haplotypes on which they arose (Figure 3.3A), using the same haplotype assignments as we used to determine the variants' origins. We were able to phase 97.9% and 96.9% of germline and postzygotic SNVs, respectively. As expected, we observe a significant enrichment of germline mutations on paternal haplotypes (Wilcoxon signed-rank test, $p=1.67\times10^{-8}$), with a 3.85:1 paternal:maternal ratio, while PZMs are evenly

distributed across parental haplotypes (Wilcoxon signed-rank test, p=0.489), with a 1.07:1 paternal:maternal ratio. Further, germline mutations increase in number with parental age, by approximately 1.64 and 0.45 additional DNMs per year of paternal and maternal age, respectively (Figure 3.3B, Figure B.3), while PZM counts are not correlated with either parent's age (Figure 3.3D). The absence of parent-of-origin effects on PZMs supports our hypothesis that these variants arose after fertilization. To further characterize the differences between germline and postzygotic SNVs, we compared their dinucleotide mutational spectra (Figure 3.2D). We find there are 29% fewer postzygotic CpG>TpG mutations relative to germline SNVs. In addition, PZMs are enriched for A>C and A>T substitutions (chi-squared test, $p=3.76\times10^{-5}$ and 2.88×10^{-4}). Notably, we observe an expected transition to transversion ratio (Ti/Tv) of 2.13 for germline DNMs, but PZMs are enriched for transversion mutations, with a Ti/Tv of 1.11. We also compared the germline mutational spectra between parents (Figure 3.3C), and found that A>C mutations are significantly enriched in the paternal germline (chi-squared test, p=0.00389). While other differences between parental spectra do not rise to significance, we observe previously reported signatures, such as an increased rate of C>T substitutions in the maternal germline [81].

3.3.2 Small de novo insertions and deletions

Using the same HiFi variant calls that we examined for single-nucleotide variation, we selected insertions and deletions that were unique to a child. We validated candidate *de novo* indels using ONT and Illumina data with similar logic that we applied to the SNVs, ensuring that each variant is present in a sample and absent from its parents across all three sequencing platforms. Notably, we divided indel calls into two categories: those that were expansions or contractions of short tandem repeats (STRs), and those that were not. For the candidate tandem repeat mutations, we used a validation strategy adapted from Goldberg et al. [60], counting the number of STR subunits that were present on every HiFi and Illumina read for a child and their parents. This analysis yielded nine mutations, a likely underestimate of the true mutation rate at STRs.



total (n=4,010)	853	95	2	14	2	3	1	2,100	1	34	41
DNM (n=2,953)	671	76	2	12	0	0	1	1,655	1	31	32
PZM (n=552)	135	13	0	2	0	0	0	279	0	3	7
deletion (n=200)	33	3	0	0	1	2	0	119	0	0	0
insertion (n=99)	14	3	0	0	1	1	0	47	0	0	2

Figure 3.2: **Autosomal SNVs and indels.** A. The number of autosomal *de novo* germline and postzygotic SNVs, insertions, and deletions <50 bp, and tandem repeat mutations observed in each sample. Sibling pairs are grouped by family and highlighted in blue, with the proband on top of the sibling. B. Distribution of the size of autosomal insertions, deletions, and tandem repeat mutations. C. Allele balance (AB) distribution for autosomal germline and postzygotic SNVs across PacBio HiFi, Illumina, and ONT read data. D. Autosomal germline and postzygotic dinucleotide mutation spectrum. E. The most severe functional annotations of each class of autosomal mutation, as assigned by VEP after lifting variants from T2T-CHM13 to GRCh38.

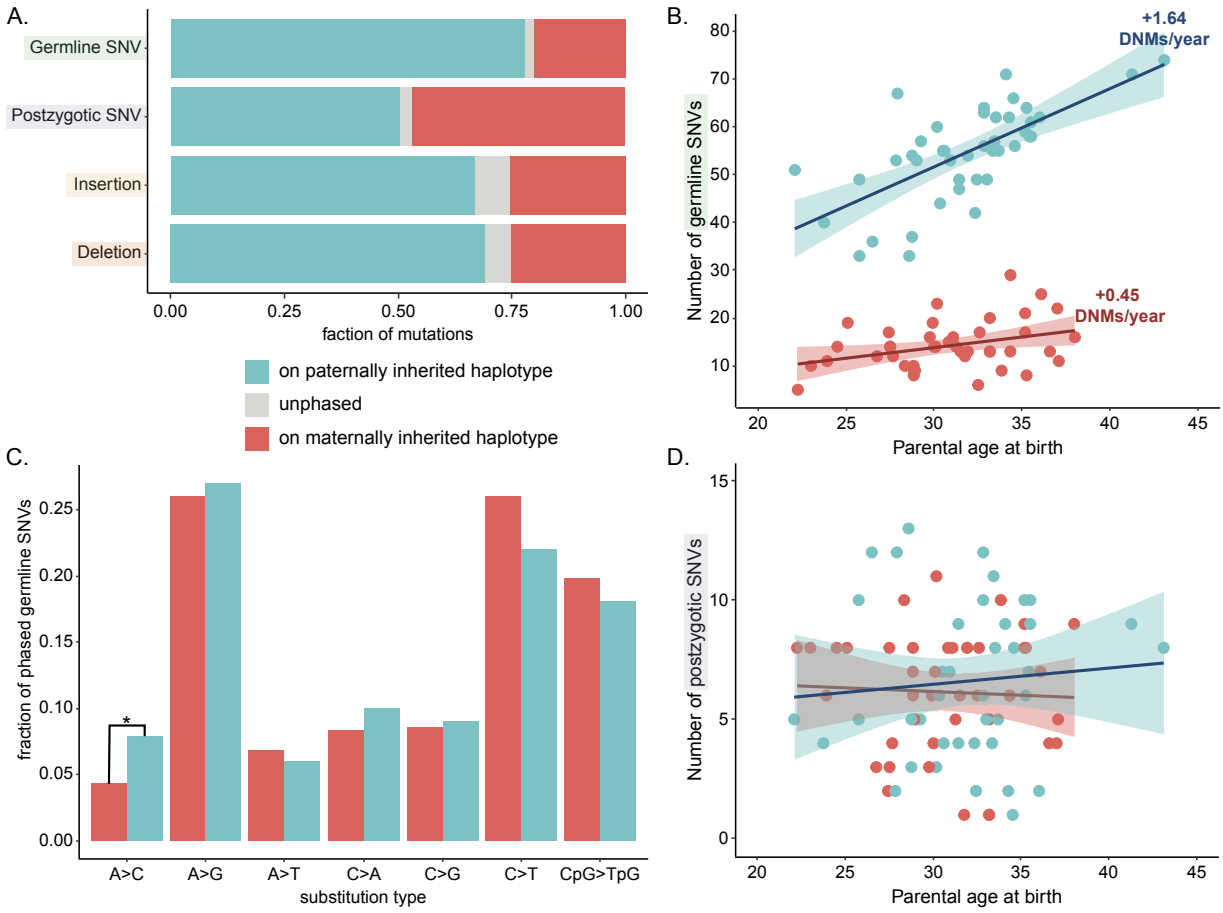


Figure 3.3: **Parent-of-origin effects on autosomal SNVs.** A. Fraction of autosomal mutations assigned to paternal or maternal haplotypes. B. Number of germline SNVs assigned to each parental haplotype as a function of parental age at birth. The slope was defined by linear regression. C. Proportion of each type of substitution in maternal and paternal germline SNVs. A>C mutations are significantly enriched in the paternal germline, p-value from chi-squared = 0.00389. D. Same analysis as B, repeated with postzygotic SNVs.

Outside of STRs, we identified 99 and 200 *de novo* insertions and deletions, respectively, for an average of 7.1 indels/child on the autosomes, with a median insertion size and deletion size of ±4 bp (Figure 3.2B). This is a 78.0% increase over the four indels/child identified by Wilfert et al. [189], a 50.2% increase over the combined Talkowski callsets [4], and an 18.9% decrease relative to Turner et al. [176]. We find that 57.3% (n=177) of our indel calls are unique to our HiFi callset. Of the 142 calls observed in Illumina studies that were absent from our final callset, only one appeared to be truly *de novo* across long-read data, while 63% appeared inherited and the remainder were absent from long-read data.

We used the same haplotype assignment method to determine whether indels were germline or postzygotic in origin. Only 11 indels appeared to emerge after fertilization, while 268 were germline, and the remaining 20 did not have enough tagging information to be unambiguously assigned. Notably, all postzygotic indels arose on paternal haplotypes, while germline indels had a 2.57:1 paternal:maternal ratio, which is significantly different from what we observed in germline SNVs (chi-squared, p=0.0061). There is not a significant relationship between parental age and the size of indel mutations, but we do observe an additional 0.16 deletions per year of paternal age (Figure B.4), which is approximately 10% of the paternal age effect for SNVs.

3.3.3 DNM rate

To estimate the DNM rate, we first had to determine the amount of the genome in which we could discover variation. For a given sample, we determined a site was callable if both their parents had homozygous reference genotypes, and if the sample and both parents had at least one blood-derived HiFi read aligned to the site with high mapping quality. We evaluated every base in the genome for every sample, calculating the callable genome for each individual, and on average, we could identify variation in 91.6% of the autosomes (2.66 Gbp/2.90 Gbp, standard deviation = 24.9 Mbp) (Figure 3.4A). Because high mapping quality is used to determine whether a site is callable with our method, we were not able to call in regions of the genome with the highest sequence identity. For example, we can call in more

than 94% of sequence space in SDs with sequence identity less than 98%, whereas we can only assess 35% of sequence space in duplications with over 99% identity. Our calling method performs even worse in centromeres, where we can only assess approximately 5% of higher order repeats and 8% of human satellite sequence. Although we cannot fully examine the variation in these repetitive regions, we are still better equipped to study them than previous Illumina-based *de novo* studies, which typically exclude them completely [81, 85, 176, 189]. Using the same Illumina criteria for all our samples, we estimate that 89.3% of the genome is callable (2.59 Gbp, standard deviation = 36.0 Mbp).

Excluding postzygotic mutations, we calculate an autosomal germline mutation rate of 1.31×10^{-8} substitutions per base pair per generation (95% C.I. 1.24 - 1.39×10^{-8}) (Figure 3.4B) and a PZM rate of 2.40×10^{-9} substitutions per base pair per generation (95% C.I. 2.13 - 2.67×10^{-9}) (Figure 3.4D). When we restrict our analysis to GENCODEv35 exonic regions of the genome (144 Mbp of sequence, on average 96.8% callable), we find that neither the germline nor PZM rate is significantly different from the autosome-wide rate. We also examined Alus, where we found a significant enrichment of germline variation but not postzygotic (two-sided t-test, p=0.0025 and p=0.12).

In SDs, the germline mutation rate is 1.80×10^{-8} substitutions per base pair per generation (95% C.I. 1.49 - 2.11×10^{-8}), a significant 38% increase over the rate across the autosomes (two-sided t-test, p=0.010). Of note, this signal is entirely driven by SDs with greater than 99% sequence identity, where the mutation rate is more than double that of the lowest identity duplications (Figure 3.4C, Figure B.5). The PZM rate is more sensitive to sequence identity and is enriched threefold in both SDs and centromeres (two-sided t-test, p=0.000039 and p=0.0031). In the case of PZMs, the strongest enrichment is in the highest identity duplications, where we observe an average of one substitution per sample (n=42). We observe the highest PZM rates in centromeric higher order repeats and human satellite DNA, where we found a total of 12 substitutions but could only call variation in 3.6 Mbp and 7.3 Mbp, respectively.

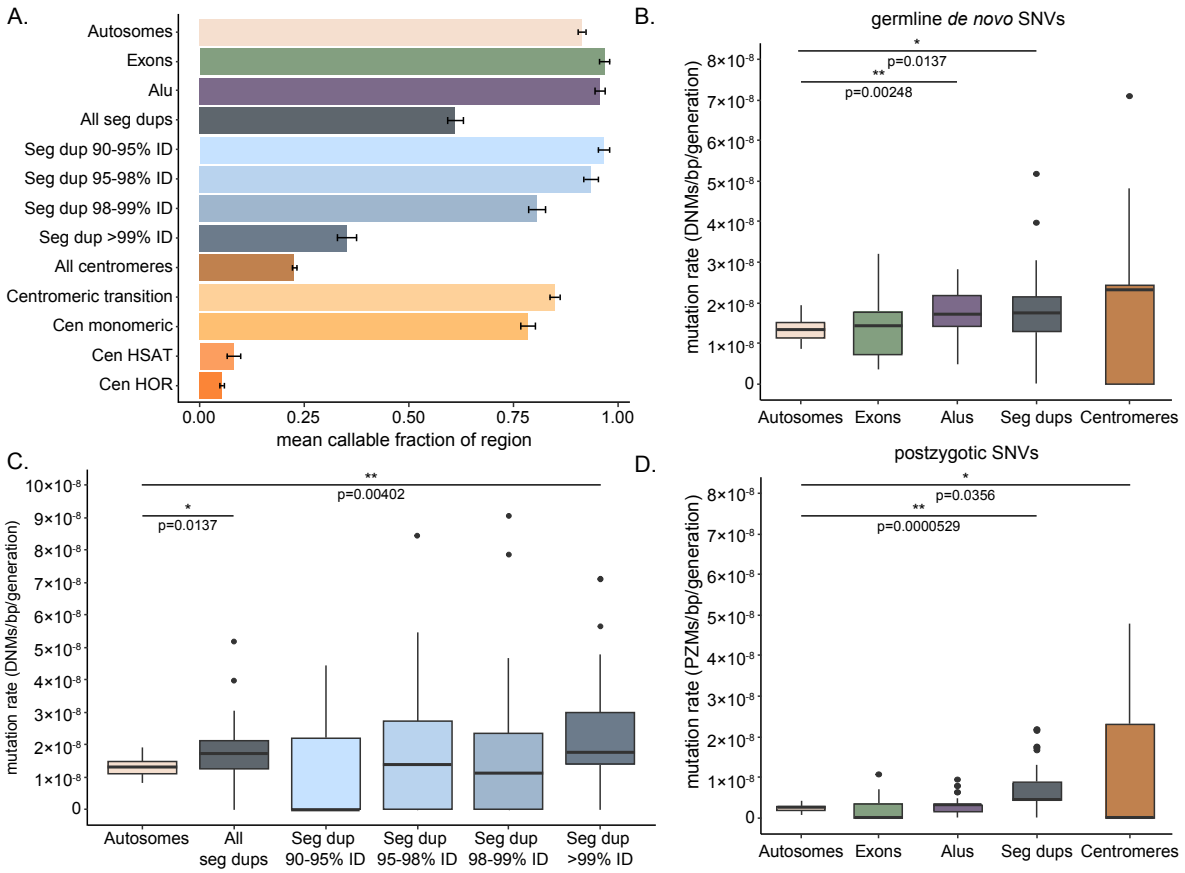


Figure 3.4: **Autosomal SNV mutation rates.** A. Mean fraction of callable space in different autosomal regions across all samples in the dataset, with error bars representing 1 standard deviation. B. Distributions of autosomal germline DNM rate for each sample across different genomic regions reveal an enrichment i n Alus and segmental duplications (seg dups). P-values calculated with t-test and adjusted for multiple testing by Benjamini-Hochberg. C. Autosomal germline DNMs across SDs stratified by percent identity (% ID). D. The same analysis as B, repeated for autosomal PZMs, finds enrichment in SDs and centromeres.

3.3.4 X and Y chromosome variation

We used a modified version of our *de novo* SNV and indel discovery strategies to call variation on the sex chromosomes, treating the females (n=32) and males (n=10) separately. On male X chromosomes, we found a total of 10 *de novo* SNVs (1 per X chromosome) and 2 indels, double the number of DNMs observed on Y chromosomes (5 SNVs and 1 indel) (Figure 3.5B). In female samples, we found a total of 183 SNVs (2.86 per X chromosome) and 15 indels. We determined the origin of female mutations using the same haplotype strategy that we used for the autosomes, and we found that 15.3% (n=28) of female chrX mutations were postzygotic in origin. We were able to assign 107 and 22 germline mutations to paternal and maternal haplotypes, respectively. Combined with the 10 SNVs observed on chrX in males that must have been inherited from their mothers, we see a 3.34:1 paternal:maternal ratio on the X chromosome, which is not significantly different from the ratio we observe on the autosomes (chi-squared, p=0.5643).

According to our HiFi filters, we were able to discover variation in 96.0% of the female X chromosome on average (standard deviation=8.9%) (Figure 3.5A). In males, we modified our filters such that we only require high-quality HiFi reads from a sample and their mother to be able to call on the X chromosome, and conversely, only from a sample and their father on the Y chromosome. We were able to call on 94.5% of the male X chromosome (standard deviation=1.1%). The Y chromosome is highly repetitive, as 30 Mbp of the 62 Mbp chromosome is comprised of satellite arrays, and another 3 Mbp of sequence is in pseudoautosomal regions, regions of homology between the X and Y chromosome [154]. Because read data cannot align unambiguously in these regions, our ability to call DNMs was limited to 29.3% of the Y chromosome (standard deviation=5.6%).

We estimate the X chromosome mutation rate to be 0.53×10^{-8} substitutions per base pair per generation in the maternal germline and 2.30×10^{-8} substitutions per base pair per generation in the paternal germline (Figure 3.5C). This enrichment of mutations in the paternal germline is even more stark on the Y chromosome, where we see a mutation rate of

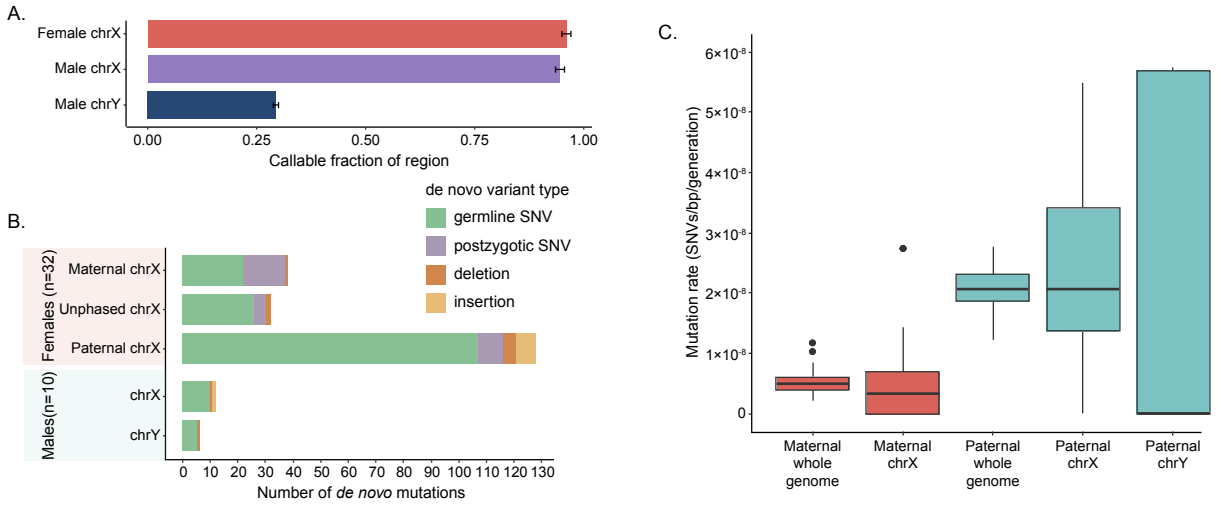


Figure 3.5: **Sex chromosome DNMs.** A. Mean fraction of callable sex chromosomes in males and females, with error bars representing 1 standard deviation. B. Number of *de novo* germline and postzygotic SNVs, as well as insertions and deletions, observed on the sex chromosomes. Mutations were phased if possible. C. Distribution of whole-genome and sex chromosome mutation rates calculated using phased germline mutations.

2.85×10^{-8} substitutions per base pair per generation. Combined with the germline mutations we observed on the autosomes, we calculate the whole-genome mutation rate in the maternal and paternal germline to be 0.53×10^{-8} and 2.07×10^{-8} substitutions per base pair per generation, respectively. The average whole-genome rate is 1.31×10^{-8} substitutions per base pair per generation, which is not significantly different from the autosomal mutation rate.

3.4 Discussion

We used long-read sequencing data aligned to T2T-CHM13v2.0 to discover and validate small *de novo* variants in 42 samples from 24 families. Studies to characterize this class of variation typically use short-read sequence data, which limits their ability to examine repetitive regions of the genome due to their inability to map uniquely. We have previously shown that by combining long reads with a more complete reference genome, we can increase

our capacity to discover *de novo* variation by at least 25% [131]. Across our samples, we identified a total of 3,813 *de novo* variants (3,505 SNVs and 308 indels) on the autosomes. We found an additional 218 variants on the sex chromosomes (198 SNVs and 18 indels), including 6 variants on the Y chromosome, which is often excluded from *de novo* studies due to its repetitive sequence content. In total, we identified approximately 95.9 small *de novo* variants per child.

Focusing on autosomal SNVs, we see a 48.7% increase over the 56.1 SNVs/child discovered in the same Illumina data by Wilfert et al. [189] (Figure B.6A). When we compare to a combined callset of An et al. [51] and Stephan Sanders [unpublished, personal correspondence], we see an increase of 40.5% over their SNV callset, whereas we see a 44.9% decrease compared to Turner et al. [176]. When comparing to the Turner study, it is important to note that their variant discovery method focused on sensitivity over specificity, resulting in an inflated callset enriched for false positives. Across all three other studies, there were a total of 664 autosomal SNV calls that were not represented in our final callset that we were able to lift over to T2T-CHM13v2.0 (Figure B.6B). We ran these calls through our variant-filtering pipeline, finding that 58.8% appeared to be inherited and 29.5% were not present in long-read data, leaving 93 SNV calls that passed our first round of read-based filtering. Only 41 of these 93 variants passed our full suite of filters, with the majority (73% of failed variants) failing because they were observed in unrelated samples. While it is certainly possible for the same mutation to occur independently in two unrelated samples, the probability is vanishingly low, and it is more conservative to assume a recurrent error in sequencing or basecalling has disguised a segregating variant as *de novo*.

In addition to the short-read studies, two samples from this dataset have previously been examined with long-read data, revealing 195 autosomal SNVs on the autosomes that had been validated by HiFi and ONT data [131]. In comparison, we identified 196 autosomal SNVs for the same samples, 14.8% (n=29) of which are unique to this study. Of those 29 SNVs, 21 are postzygotic in origin. Conversely, there were 28 SNVs reported in the previous study that are absent from this one. More than a third (n=10) were originally discovered

only in GRCh38-aligned reads and were consequently not identified by our variant callers. Of the remaining 18 variants, 6 were part of mutational clusters and excluded from our SNV calls, 5 were observed in multiple samples and excluded as errors, 4 failed validation by T2T-CHM13v2.0 aligned-reads, and the remaining 3 SNVs had conflicting parental haplotype assignments. While the estimated number of DNMs for this family remained nearly the same, this comparison reveals that we missed approximately 16 events across these two samples, giving a false negative rate of approximately 7.5%.

Besides increasing our ability to discover *de novo* variation, long-read data have a significant advantage when it comes to phasing variants or assigning them to parental haplotypes. Without pedigree information, short-read data can typically be used to phase up to 20% of DNM calls, but we were able to phase 97.7% and 99.3% of autosomal SNVs and indels, respectively. We found that 79.4% of germline SNVs arose in the parental germline, falling within the 3.1-4.0:1 paternal to maternal ratio observed in previous studies [81, 91, 162]. We observed parental age effects of an additional 1.64 and 0.45 DNMs per year of paternal and maternal age, respectively, both of which are slightly higher but within the range of previous studies [61, 81, 91, 162]. There was a significant enrichment of A>C mutations in the paternal germline compared to the maternal germline, but while we observed other previously characterized differences [81], none of them rose to significance, likely due to our smaller callset size. Notably, the paternal mutation spectrum is significantly correlated with paternal age, but we do not observe any such correlation in the maternal spectrum.

Another perk to our ability to phase *de novo* variants is that we can use phasing information to infer whether a variant arose in the parental germline or postzygotically, in early rounds of cell division after fertilization. We determined variant origin by checking whether a *de novo* variant was observed on every read from a parental haplotype in HiFi and ONT data, using AB across both platforms and Illumina to resolve any cases of disagreement. Based on this method, we found that 84.3% of autosomal SNVs arose in the parental germline, and the remaining 15.7% were postzygotic in origin, an average of 13.1 PZMs per child. This is higher than even the most sensitive previous studies of postzygotic variation, which

estimated that up to 10% of all *de novo* variants are actually postzygotic in origin [162]. One caveat is that we cannot say with certainty that these PZMs are not somatic mutations that arose in blood, but the same analysis applied to a multi-generational family with 6-8 children per generation reveals that 64.5% of identified PZMs were transmitted to the next generation, confirming their presence in a sample's germline in addition to blood [141]. This does not, however, indicate that the remaining 35% of PZMs are exclusive to blood. Despite the large number of offspring in this family, the average AB of transmitted PZMs was 0.18; we would expect 20-30% of PZMs to go untransmitted if they were present in only 18% of all cells. Without sampling other tissues, we can confidently assert that at least two-thirds of our PZMs are truly early embryonic events, but the true number is likely much higher. Because PZMs do not arise in parental gametes, we do not observe an enrichment on paternal haplotypes or a parental age effect like we do for germline events. In addition, we identified novel differences in the PZM spectrum [162], like an 80% enrichment of A>C and 29% depletion of CpG>TpG substitutions (Figure 3.2D), reflecting different mutational mechanisms compared to germline mutations. Both germline and postzygotic variants are, however, enriched in repetitive sequence, like SDs, indicating that hypermutability is conferred by high sequence identity.

To estimate the mutation rate, we first determined that we were able to identify variation in 91.0% of the genome on average (2.79 Gbp). Across the genome, we found a germline mutation rate of 1.31×10^{-8} substitutions per base pair per generation, which is very close to estimates of $1.1\text{-}1.29\times10^{-8}$ made by Illumina-based studies [81, 91, 150, 162]. We find that the autosomal PZM rate is 0.23×10^{-8} substitutions per base pair per generation, approximately 18% of the germline rate. When comparing mutation rates to previous studies, it is important to note that every study treats PZMs differently: most studies use a filter to exclude mutations with AB<0.25 or AB<0.30 [81, 91], some do not account for PZMs [150], and some explicitly break them out as their own mutational class [162]. We expect studies that do not account for PZMs to overestimate the germline mutation rate by approximately 10-15%, and given that 57.4% of PZMs in our callset have AB<0.25, studies that implement

an AB filter should overestimate the germline rate by approximately 4-6%. Adjusting for PZMs, the germline mutation rate estimates from previous studies would be expected to fall into the range of 1.1-1.21$\times 10^{-8}$. Not only does our study accurately report a higher germline mutation rate, but the combined DNM and PZM rate is 1.54×10^{-8}, a marked increase over previous reports [81, 91, 150, 162].

Both germline mutations and PZMs are enriched in SDs, specifically in those with greater than 99% sequence identity, where the germline mutation rate is 79% higher and the postzygotic rate is sixfold higher. Across all SDs, we see a more modest increase in germline mutation rate (38%), which is lower than we might expect given previous estimates of a 60% increase of SNVs in SDs [180]. However, if we pool germline mutation and PZM rates, we see a 66% higher rate in SDs, consistent with that previous estimate. In fact, the increased postzygotic signal is in line with the types of mutations observed in SDs, as both groups are depleted in CpG substitutions and have an excess of transversions (PZM Ti/Tv = 1.11:1).

While we were able to assess small variants in approximately 91% of the genome, there is still the potential to discover additional *de novo* variation in these families. For example, we were not able to align HiFi read data to more than half of the highest identity SDs and over three-quarters of the Y chromosome, both of which we know to be highly mutable [141]. We also did not assess large-scale structural variation, which alignment-based methods are not suited to discover. Instead, analysis of the most complex regions and variants will have to be performed with assembly-based discovery techniques.

3.5 Methods

Illumina sequencing

We used previously published Illumina data generated by the NYGC for the Simons Simplex Collection [131, 176, 189]. Briefly, each sample was sequenced on the Illumina X Ten platform using 1 μg of blood-derived DNA with an Illumina PCR-free library protocol.

HiFi sequencing

We used the same method as described in Noyes et al. [131] to generate HiFi data from blood and cell-line derived DNA.

ONT sequencing

ONT data were generated from DNA extracted from lymphoblastoid cell lines using a modified Gentra Puregene protocol. Libraries were constructed using the Ligation Sequencing Kit (ONT, SQK-LSK110) with modifications to the manufacturer's protocol. The libraries were loaded onto a primed FLO-PRO002 R9.4.1 flow cell for sequencing on the PromethION, with two nuclease washes and reloads after 24 and 48 hours of sequencing.

Alignment to the reference genome

We selected T2T-CHM13v2.0 as our reference genome, as it allows us to align reads to repetitive regions that were not represented in the previous reference. For all females, we masked the Y chromosome from the reference before alignment, and for all males we masked the pseudoautosomal regions of the Y, following the guidance from Rhie et al. [154]. We aligned both HiFi and ONT data using minimap2 v2.17 [101], with the help of the pbmm2 v1.13.1 (https://github.com/PacificBiosciences/pbmm2) wrapper for handling the HiFi data. Illumina data were aligned using BWA-MEM v0.7.17 [99].

de novo SNV discovery and validation on the autosomes

Variant calling was performed using aligned HiFi data and two variant callers, GATK HaplotypeCaller v4.3.0.0 [140] and DeepVariant v1.4.0 [139], following the same filtering strategy outlined in Noyes et al. [131]. For each caller, we naively identified candidate *de novo* events by selecting any variant where both parents were homozygous for the reference allele and the child had at least one alternate allele. We took the union of both candidate *de novo* callsets, retaining only variant calls where the child's genotype quality was at least 20, resulting in an

initial callset of 1,200,313 SNVs across all 42 samples. To eliminate runs of candidate *de novo* events that were actually the result of a dropped haplotype in one parent, we eliminated any regions where three or more SNVs were found in a sliding 1 kbp window, removing a total of 1,011,837 candidate events from our callset.

Next, we examined the HiFi, ONT, and Illumina reads that spanned each candidate variant in a child and both parents. For a child's HiFi read to be considered, it had to be derived from blood data, but both parental blood and cell line reads were retained (all ONT data were derived from cell lines, and all Illumina data from blood). Long reads with mapping quality <59 were excluded (we did not filter short reads on the basis of mapping quality). We partitioned reads into three categories based on the base quality (probability that a base was correctly called) at the site of the variant: reads with base quality >20 (high quality), reads with base quality between 10 and 20 (low quality), and reads with quality <10, which were discarded. For each sequencing platform, we counted the number of reads that supported the reference and alternate alleles in both parents and the child and used them to determine whether a variant was truly *de novo* or inherited. For HiFi and Illumina data, we required that each parent have fewer than one high-quality or two low-quality reads with the *de novo* allele, and the child have at least one read with the *de novo* allele. Since ONT is slightly less accurate, we required fewer than two high-quality or three low-quality reads with the *de novo* allele. Once each variant was examined in each platform, we combined the validations, determining that a variant was inherited if it looked inherited in at least one platform, and truly *de novo* if it was supported in at least two platforms. Across all samples, 8,559 candidate mutations passed this initial round of filtering.

We returned to the aligned HiFi reads for every sample in the dataset (parents and children), checking every candidate *de novo* allele to see if it was represented in more than one sample, removing 4,237 variants that we determined to be recurrent errors. If a variant was not present in a tandem repeat (TR), we required that it be unique to the child it was identified in, and if the variant was in a TR, we allowed it to be observed in one unrelated sample. In addition, variants in TRs had to have an average AB greater than 0.05

across all platforms. We removed another 754 low-quality variants, either because they had dubious support across two or more platforms (typically noisy parental data with alternate alleles different from the *de novo* allele). We excluded 41 variants in regions flagged by RepeatMasker that failed AB filters (0.1 if it was also in a TR, 0.08 if not) [165]. Lastly, we removed 22 variants in or adjacent to homopolymers that involved the homopolymer subunit (i.e., an A to T substitution on the edge of an A homopolymer), as those variants are typically sequencing artifacts, and difficult to validate across all sequencing platforms. We validated a total of 3,575 variants that we then assigned to haplotypes. A final 70 variants could not be uniquely assigned to one parent so were excluded, resulting in a final callset of 3,505 autosomal SNVs.

de novo indel discovery and validation on the autosomes

We generated candidate indel callsets using the same combined GATK and DeepVariant callset, naively selecting for any insertions or deletions present in a child and absent from the parents. We divided these calls into two categories: TR mutations, where one or more subunits were added or subtracted (n=1,294) and indels that either did not involve a perfect TR motif or did not overlap with a TR (n=384,395). We applied a similar read-based validation strategy for indels as we did for SNVs. We required that HiFi, ONT, and Illumina reads have mapping quality of 60 (the highest possible), and that they fully span the variant site, with at least 10 bp of flanking sequence before and after, to ensure that we captured the full allele.

For variants in TRs, we counted the number of subunits present on every read and determined whether the child had a unique number of subunits across both HiFi and Illumina data (we excluded ONT, which was too noisy to determine the true number of subunits present in a read). If a variant was supported in the child and absent from the parents across both platforms, we considered it to be truly *de novo*. If they conflicted about the size of the *de novo* allele, we examined every unique allele in a child to determine whether it could be verified across both platforms. In total, only nine TR events passed our validation

filters.

To validate indels outside of TRs, we examined child and parental data across all three sequencing platforms, counting the number of *de novo* alleles. If a child had a sibling in our dataset, we also examined the sibling's read data. We considered a variant to be inherited if one parent or the sibling had a read supporting the *de novo* allele in any platform. We deemed a variant to be truly *de novo* if we observed the *de novo* allele in the child in both HiFi and Illumina data, resulting in a total callset of 318 *de novo* indels.

Sex chromosome de novo discovery

To identify variation, we used ploidy-aware GATK HaplotypeCaller v4.3.0.0 [140], treating the female chromosome X as diploid, and the male sex chromosomes as haploid. Females and males from each family were jointly genotyped separately, and variant calls were filtered using the same parameters as autosomal variants.

For female children, we naively identified *de novo* variation by selecting sites that were homozygous reference in the mother and hemizygous reference in the father, and the child had an alternate allele, identifying 20,282 SNVs and 13,923 indels. We excluded 15,346 SNVs that were in clusters of three or more within 1 kbp, then used most of the SNV filtering strategy that we applied to autosomal variants, examining HiFi, ONT, and Illumina reads to ensure each variant was unique to the child in which it was identified. We used the same filtering parameters for sites in TRs but did not filter based on RepeatMasker or homopolymer annotations [165], as few sites were in such regions. In total, 187 SNVs on female X chromosomes passed our validations. After assigning these variants to parental haplotypes, four had conflicting parental information and were excluded from our final callset of 183 SNVs. We used the same autosomal indel validation pipeline for non-TR variants, resulting in a final callset of 15 indels on female X chromosomes.

For male children, we treated X and Y chromosome variation separately, excluding the pseudoautosomal regions on both. We naively identified variants on the X chromosome by comparing a child to the mother, selecting any sites where the child had a different allele (not

required to be reference or alternate). Conversely, we selected sites on the Y chromosome where a child had a different allele than the father. We identified 20,509 SNVs and 34,929 indels on male X chromosomes, and 45,816 SNVs and 4,453 indels on male Y chromosomes. To validate SNV calls on the male X chromosome, we applied the same filtering strategy as the female X, first evaluating HiFi, ONT, and Illumina sequencing data to ensure that a variant was unique to a child, and then filtering sites in TRs, resulting in a total of 10 SNVs on male X chromosomes. For SNV calls on the Y chromosome, we simply checked a child and father's sequencing data across all three platforms to ensure that the variant was not present in the father. No further filtering steps were performed, resulting in a final callset of five SNVs on male Y chromosomes. We used the same indel filtering strategy for male sex chromosome variants as for female, except that we only examined maternal sequencing data for X and paternal data for Y chromosome variants. In total, we found two indels on male X and one indel on male Y chromosomes.

Phasing

For every DNM, we identified informative SNPs within an 80 kbp window centered at the mutation site based on variant calls from our GATK4 run using HiFi data. An informative SNP is defined as any SNP whose origin can unambiguously be assigned to one parent: for example, a site where one parent is 0/0, the other parent is 0/1 or 1/1, and their child is 0/1. We then examined the HiFi read data for the child, examining every read derived from blood DNA that passed our read filters (mapping quality $\geq$59 and base quality $\geq$20 at the DNM site). We assigned each read to a maternal or paternal haplotype by calculating an inheritance score based on the presence of tagging SNPs. We gave tagging SNPs a value of $\pm$1 depending on whether they were inherited from the mother or father and then took the average of these values, inversely weighted by each SNP's distance from the DNM site. A negative inheritance score indicated a paternally inherited read, and a positive score indicated a maternally inherited read. If all the reads with the *de novo* allele could be assigned to one parent, we assigned them as the parent of origin. If the *de novo* allele was present on both

paternal and maternal haplotypes, it was left unphased. Using this method, we were able to phase 91.1% of all SNVs (n=3,221), while 158 had no tagging SNPs and 155 had ambiguous parental data.

DNM origin was also evaluated using ONT reads. We applied the same read-filtering steps, with the caveat that all ONT data were derived from cell lines, and calculated inheritance scores based on informative SNPs. For each SNP, we compared the HiFi and ONT haplotype assignments, preferentially selecting the HiFi assignment in cases of disagreement (n=4) between the two sequencing platforms. With the ONT data, we were able to phase an additional 204 DNMs, leaving just 2.3% of our DNMs (n=80) unphased. In cases where tagging SNPs were present, but neither ONT nor HiFi data could unambiguously phase a *de novo* variant (n=29), we assumed the variant was due to sequencing error and excluded it from the final DNM callset.

PZM assessment

Using the filtered and parentally assigned HiFi and ONT reads from our phasing pipeline, we counted the number of reference and alternate alleles derived from each parent and determined that a DNM was postzygotic in origin if we detected at least two reads with the reference and one read with the alternate allele assigned to one parent's haplotype.

We also predicted whether a mutation originated in the parental germline or the early embryo based on the new variant's AB across HiFi, ONT, and Illumina sequencing data. First, we filtered reads according to the same parameters used in the phasing pipeline, restricting HiFi data to only blood-derived reads, and counted the number of reads with the reference and alternate alleles. We used these counts to calculate AB for each platform and then used a chi-squared test (using chi2_contingency from the python package scipy.stats) to determine whether the three AB values were concordant. In cases where the AB was not concordant, we could not confidently predict whether the variant was postzygotic, so it was supposed to be germline. If the AB was concordant across platforms, we pooled the reference and alternate allele counts and tested whether the total AB was significantly less than 0.5

using a binomial test (binomtest from scipy.stats). Any variants with significantly low AB were predicted to be postzygotic in origin.

To make the final determination of mutation origin, we combined results from the HiFi and ONT haplotypes and AB-based predictions. In cases where HiFi and ONT haplotypes disagreed (n=228), we used the origin assignment that matched our AB prediction; in cases where HiFi and ONT haplotypes were ambiguous (n=109), we used the AB prediction. In total, we determined that 552 *de novo* SNVs were likely postzygotic in origin, and 2,953 arose in the parental germline.

Callable genome and mutation rate calculation

To determine where we were able to identify *de novo* variation in the genome, we assessed HiFi data for every trio. We first used GATK HaplotypeCaller v4.3.0.0 with the option "ERC BP_RESOLUTION" to generate a genotype call at every site in the genome [140]. Only sites where both parents were genotyped as homozygous reference (0/0) were considered callable, as sites with a parental alternate allele were excluded from our *de novo* discovery pipeline. We then examined the HiFi reads from a sample and its parents, restricting to only primary alignments with mapping quality of at least 59. For children, we only considered HiFi reads derived from blood, but we considered blood and cell line data for parents. We counted the number of reads with a base quality score of at least 20 at every site in the genome, and then combined this information with our variant calls. A site was deemed callable if both parents and the child each had at least one high-quality read with a high-quality base call. We observed an average of 2.66 Gbp/2.90 Gbp (standard deviation = 24.9 Mbp) such sites across the autosomes. For female children, the callable chromosome X was determined the same way, whereas for male children, we only considered the mother's HiFi data when examining the X chromosome and the father's HiFi data when examining the Y chromosome. In addition, male sex chromosomes were not restricted to sites where both parents were genotyped as reference—each parent was allowed to carry an alternate allele.

We calculated the germline autosomal mutation rate for every sample by dividing the

number of germline autosomal DNMs by twice the number of base pairs we determined to be callable. For PZMs, we used the same denominator. In females, the amount of callable sex chromosomes was defined as twice the number of callable bases on the X chromosome, and in males it was defined as the sum of the callable bases on the X and Y chromosomes. For each feature-specific mutation rate (such as SDs), we intersected both a sample's *de novo* SNVs and the sample's callable regions with coordinates of the relevant feature. We then calculated the mutation rate by dividing the number of SNVs in the region by the amount of callable space in the region.

Chapter 4

A FAMILIAL, TELOMERE-TO-TELOMERE REFERENCE FOR HUMAN *DE NOVO* MUTATION AND RECOMBINATION FROM A FOUR-GENERATION PEDIGREE

This chapter is adapted with modification from:

> Porubsky, D, Dashnow, H*, Sasani, TA*, Logsdon, GA*, Hallast, P*, **Noyes, MD***, Kronenberg, ZN*, Mokveld, T*, Koundinya, N, Nolan, C, Steely, CJ, Guarracino, A, Dolzhenko, E, Harvey, WT, Rowell, WJ, Grigorev, K, Nicholas, TJ, Oshima, KK, Lin, J, Ebert, P, Watkins, WS, Leung, TY, Hanlon, VCT, McGee, S, Pedersen, BS, Goldberg, ME, Happ, HC, Jeong, H, Munson, KM, Hoekzema, K, Chan, DD, Wang, Y, Knuth, J, Garcia, GH, Fanslow, C, Lambert, C, Lee, C, Smith, JD, Levy, S, Mason, CE, Garrison, E, Lansdorp, PM, Neklason, DW, Jorde, LB, Quinlan, AR, Eberle, MA, Eichler, EE. A familial, telomere-to-telomere reference for human *de novo* mutation and recombination from a four-generation pedigree. 2024.08.05.606142
>
> Preprint at https://doi.org/10.1101/2024.08.05.606142 (2024).

*These authors contributed equally to this manuscript

4.1 Author contributions

M.D.N., D.P., P.H., G.A.L. and Z.N.K. Analysis of *de novo* mutations.

N.K., W.T.H. and D.P. Generation of *de novo* assemblies and validation.

P.H., P.E., and C.Li Chromosome Y analysis.

D.P., K.K.O. and G.A.L. Centromere analysis.

4.2 Abstract

Using five complementary short- and long-read sequencing technologies, we phased and assembled >95% of each diploid human genome in a four-generation, 28-member family (CEPH 1463) allowing us to systematically assess *de novo* mutations (DNMs) and recombination. From this family, we estimate a range of 98-206 DNMs per transmission, including 74.5 *de novo* single-nucleotide variants (SNVs), 7.4 non-tandem repeat indels, and 4.4 centromeric *de novo* SVs and SNVs. Among males, we find 12.4 *de novo* Y chromosome events per generation. We accurately assemble 288 centromeres and six Y chromosomes across the generations, documenting *de novo* SVs, and demonstrate that the DNM rate varies by an order of magnitude depending on repeat content, length, and sequence identity. We show a strong paternal bias (75-81%) for all forms of germline DNM, yet we estimate that 14.6% of *de novo* SNVs are postzygotic in origin with no paternal bias. The use of multiple orthogonal technologies, near-telomere-to-telomere phased genome assemblies, and a multi-generation family to assess transmission has created the most comprehensive, publicly available "truth set" of all classes of genomic variants. The resource can be used to test and benchmark new algorithms and technologies to understand the most fundamental processes underlying human genetic variation.

4.3 Introduction

The complete sequencing of a human genome was an important milestone in understanding some of the most complex regions of our genome [132]. Its completion added an estimated 8% of the most repeat-rich DNA, including regions typically excluded from studies of human genetic variation and recombination analyses, such as centromeres [3], segmental duplications (SDs) [182], and acrocentric regions [62, 132]. Long-read sequencing has also driven assembly-based approaches to understand human genetic variation, revealing new insights into mutational mechanisms and access to regions previously considered intractable [119, 142, 180]. The ability to construct a phased genome assembly where the paternal and

maternal complements are nearly fully resolved from telomere-to-telomere (T2T) opens up, in principle, the discovery of all forms of variation irrespective of class or complexity or the regions where they occur, placing them into the haplotypic context in which they immediately arose [43, 106]. Direct comparison of parental genomes to their offspring increases the power to discover *de novo* mutation (DNM) as opposed to mapping reads to an intermediate reference, such as GRCh38 or T2T-CHM13 [131].

The goal of this study was to construct a high-quality T2T human pedigree resource where chromosomes were fully assembled and phased and their transmission studied intergenerationally to serve as a reference for understanding both recombination and DNM processes in the human species. We sought to eliminate three ascertainment biases with respect to discovery, including biases to specific genomic regions, classes of genetic variation, and reference genome effects. In addition to read-based approaches, we directly compare parent and child genomes to increase specificity and sensitivity of discovery in difficult regions of the genome, such as centromeres or chromosome Y. To achieve this, we focused on a four-generation, 28-member family, CEPH 1463, which has been intensively studied over the last three decades [32], and sequenced members with five sequencing technologies having distinct and complementary error modalities. This particular pedigree has served as a benchmark for early linkage mapping studies [32, 59] and optimization of short-read sequencing data by Illumina [42] and continues to serve as reference for understanding human variation, including patterns of mosaicism [7, 162].

Different from previous investigations, we focused our discovery on the sequencing and analysis of DNA obtained from primary tissue (i.e., peripheral blood leukocytes) as opposed to cell lines. We reconsented living family members (generations 2-3) and extended the sample collection to the fourth generation providing the opportunity to assess the transmission of DNMs. While all sequencing data and assemblies are available in dbGaP, 17 family members consented for their data to be publicly accessible similar to the 1000 Genomes Project samples. Just as the initial T2T genome1 served as a reference for understanding all regions of the genome, our objective was to create a reference truth set for both inherited and *de*

novo variation. Our integration of multiple long- and short-read sequencing technologies across four generations allows us to understand the factors that affect the pattern and rates of DNMs in regions that were previously inaccessible.

4.4 Results

4.4.1 Analysis of de novo SNVs and small indels

To discover small variants, we examined HiFi reads aligned to T2T-CHM13 for variant discovery, then leveraged orthogonal ONT and Illumina data to confirm that a variant is in fact present in a sample and absent from both its parents (Methods). This strategy reduces bias introduced by any one sequencing platform, but it restricts DNM discovery to G2 and G3 individuals, as we did not have corresponding G4 cell lines for ONT data generation. Our *de novo* callset included 755 SNVs and 73 indels across the autosomes of 10 individuals (n=2 G2; n=8 G3 individuals, Figure 4.1A), as well as 27 SNVs and 1 indel on Chromosome X.

To further characterize autosomal DNMs, we used flanking SNVs from long-read data to construct haplotypes, phase variants, and trace a mutation back either to a parental gamete or the early embryo. We determined that a mutation occurred somatically, and likely early in embryonic development, if it met one of two criteria: it was incompletely linked to a parental haplotype (n=112), or, if it could not be phased, it had an allele balance significantly less than 0.5 across all three sequencing platforms (n=7). We further validated each postzygotic mutation (PZM) by tracing its haplotype backward across generations and forward for the four individuals with sequenced offspring (Appendix C, Supplementary Note 1). Of the 62 PZMs in these four samples, 64.5% (n=40) are transmitted to the next generation, compared to 97.1% of germline SNVs (n=242/249) and 100% of indels (Figure 4.2A-D). We found that 10 PZMs failed these haplotype-based validations, resulting in a final callset of 109 PZMs, accounting for 14.6% of total autosomal SNVs. Previous Illumina-based analysis of this family [162] identified 605 *de novo* SNVs of either germline (G2 and G3) or postzygotic (only

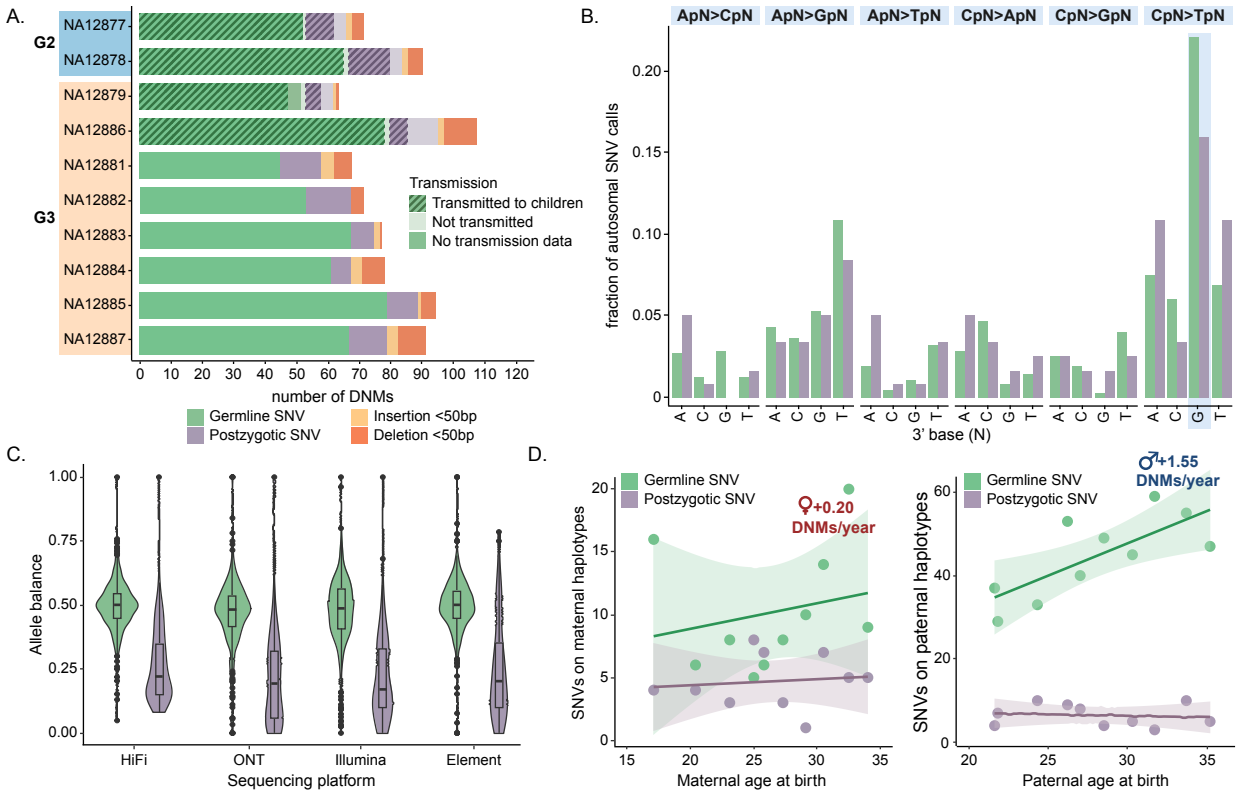


Figure 4.1: **Summary of DNMs.** A. The number of *de novo* germline/postzygotic mutations (PZMs) and indels (<50 bp) for the parents (G2) and 8 children in CEPH 1463. Crosshatch bars are the number of SNVs confirmed as transmitting to the next generation. B. The dinucleotide mutation spectrum of DNMs and PZMs does not reveal any signficant differences. C. Germline SNVs have a mean allele balance near 0.50 across sequencing platforms, while the mean postzygotic allele balance is less than 0.25. D. A strong paternal age effect is observed for germline *de novo* SNVs but not for PZMs.

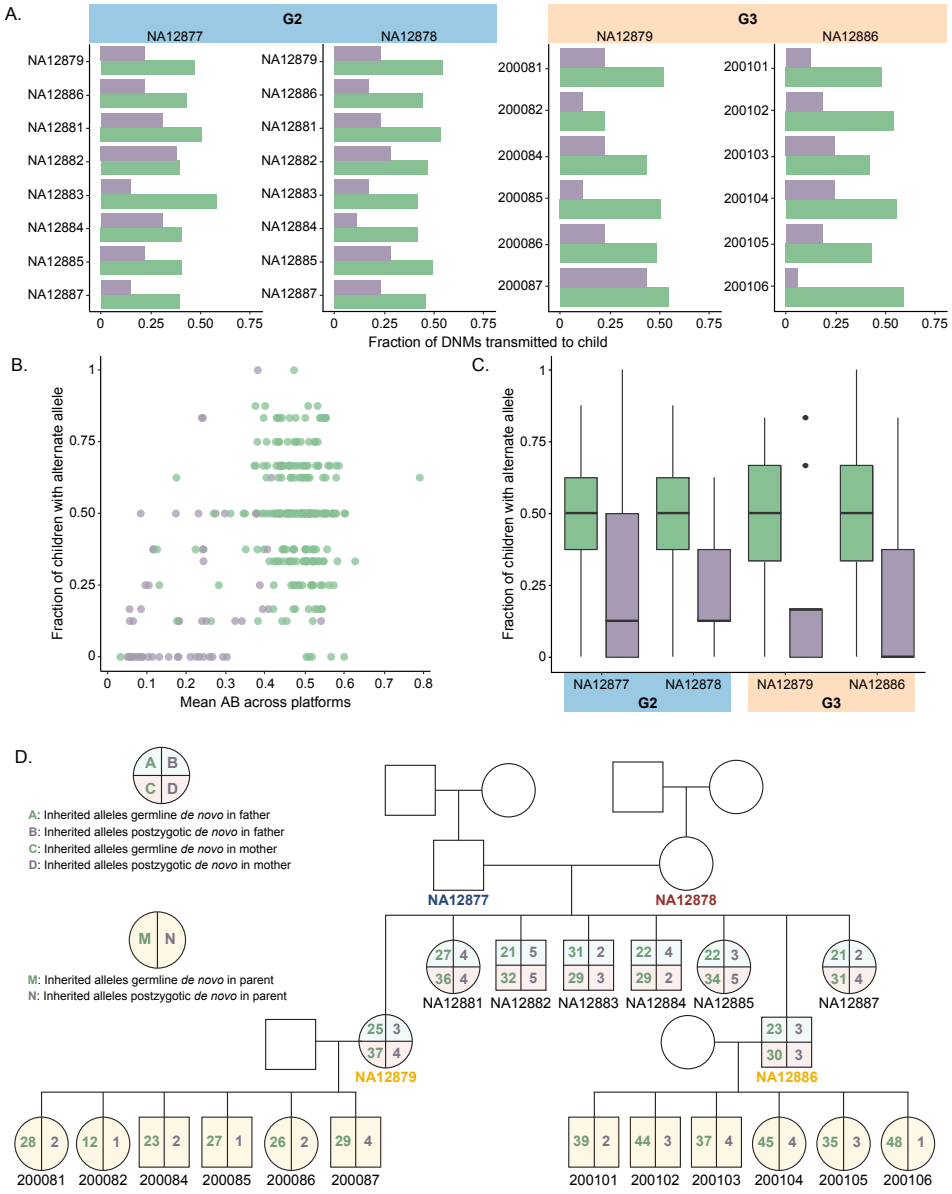


Figure 4.2: **Number of germline and postzygotic SNVs transmitted to children.** A. The fraction of a parent's germline SNVs (green, DNMs) and postzygotic SNVs (purple, PZMs) transferred to each child. B. The mean allele balance (AB) of DNMs and PZMs across HiFi, Illumina, ONT, and Element data plotted against the fraction of children who inherited a variant reveals that about half the PZMs with AB <0.25 get transmitted to at least one child. C. On average, DNMs are transmitted to 50% of children, while PZMs are transmitted to less than 25% of children. D. Number of DNMs and PZMs transmitted to each child in the pedigree.

G2) origin, 92.4% (n=559) of which were represented in our final callset, while all but four of the absent variants failed validation with long-read data. Not only were we able to identify an additional 72 PZMs in G3 for the first time, but we also identified a total of 186 novel DNMs, a 6.1% and 2% increase in germline SNV and indel discovery, respectively.

We find that 81.4% of germline small DNMs originate on paternal haplotypes (4.38:1 paternal:maternal ratio, Wilcoxon signed-rank test, $p<2\times10^{-16}$), with a significant parental age effect of 1.55 germline DNMs per additional year of paternal age when fitting with linear regression (two-sided t-test, p=0.013). In contrast, PZMs show no significant difference with respect to parental origin (1.38:1 paternal:maternal ratio, Wilcoxon signed-rank test, p=0.09) and no parental age effects (Figure 4.1D). While our small sample size does not provide sufficient power to detect significant differences between the *de novo* and postzygotic mutational spectra (Figure 4.3B), we do observe a novel depletion of CpG>TpG PZMs (chi-squared test, p=0.17) and an enrichment of postzygotic T>A substitutions (chi-squared test, p=0.268) that has been previously observed [162].

Using this approach, we successfully assay 91.9% of the autosomal genome (2.66 Gbp) (4.3A). Excluding all variants classified as postzygotic, we find that the parental germline contributes 1.17×10^{-8} SNVs/bp/generation (95% CI: 1.08 – 1.27×10^{-8}). *De novo* SNVs are significantly enriched in repetitive sequences, as much as 2.8-fold in centromeres (95% CI: 1.79 – 5.51×10^{-8} SNVs/bp/generation, two-sided t-test, p=0.017) and 1.9-fold in SDs (95% CI: 1.64 – 2.88×10^{-8} SNVs/bp/generation, two-sided t-test, p=0.0066) (Figure 4.3B). We observe a lower PZM rate of 2.04×10^{-8} SNVs/bp/generation (95% CI: 1.68 – 2.47×10^{-9}) across the autosomes, yet we see 3.9-fold enrichment of PZMs in SDs (95% CI: 4.84×10^{-9} – 1.25×10^{-8} SNVs/bp/generation, two-sided t-test, p=0.049).

4.4.2 Centromere familial transmission and de novo SNVs

Among the 288 completely sequenced and assembled centromeres, we were able to assess 150 transmissions (33 from G1 to G2 and another 117 transmissions from G2 to G3) Comparing these assembled centromeres between parent and child, we identify 18 (12%) *de novo* SVs

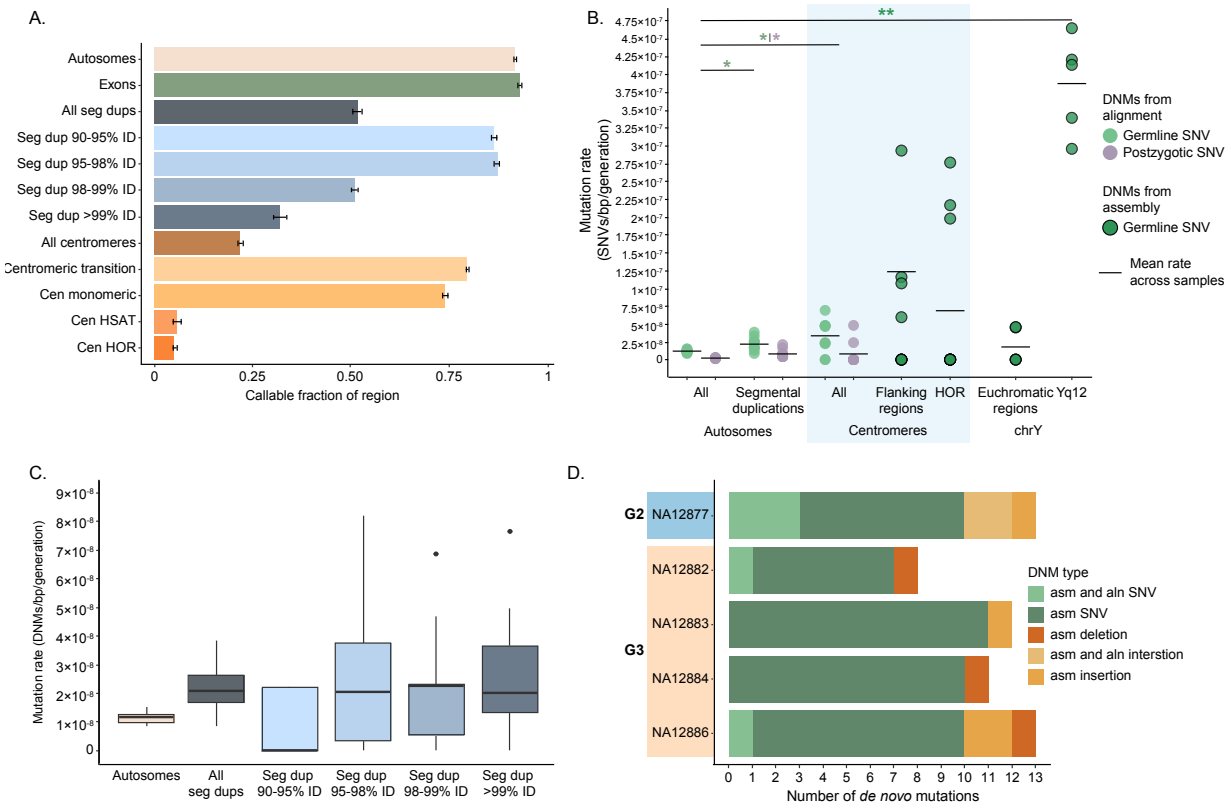


Figure 4.3: **DNM rates.** A. Mean fraction of callable space in different autosomal regions across all samples in the dataset, with error bars representing 1 standard deviation. B. Estimated SNV DNM rate by region of the genome shows a significant excess of DNM for large repeat regions, including segmental duplications (SDs). Assembly-based DNM calls on the centromeres and Y chromosome show an excess of DNM in the satellite DNA. C. The highest identity SDs drive the observed enrichment of DNMs. D. Comparison between alignment-based and assembly-based DNM calls on the Y chromosome. All DNMs discovered in alignments were also discovered in assemblies.

validated by both ONT and HiFi data with roughly equivalent number of insertions and deletions (Methods). All *de novo* SVs (n=8) that had a child sequenced as part of this study confirmed transmission of the SV to the next generation. We also identify 16 SNV DNMs in centromeres, including five within the α-satellite HOR arrays, revealing a DNM rate of 1.01×10^{-7} mutations/bp/generation (95% CI: 5.75×10^{-8} – 1.63×10^{-7}). This rate is comparable to the rate from our read-based mapping approach, which identified 14 centromeric SNVs, albeit over more than 10 times the amount of sequence, resulting in a DNM rate of 3.27×10^{-8} mutations/bp/generation (95% CI: 1.79×10^{-8} – 5.51×10^{-7}) (Figure 4.3B). Combined, we estimate a significantly higher SNV DNM rate for centromeres of 4.94×10^{-8} (two-sided t-test, p=0.017). While discovery of these DNMs still remains challenging, we believe this a conservative estimate because we required validation of all events by ONT and HiFi sequencing platforms.

4.4.3 Y chromosome mutations

Here, we focus on the ∼59.7 Mbp male-specific Y-chromosomal region (MSY, i.e., excluding pseudoautosomal regions) considering both read-based as well as assembly-based approaches to discover DNMs (Methods). There are nine male members who carry the R1b1a-Z302 Y haplogroup across the four generations (Figure 4.4A,B) and we use the great-grandfather (G1-NA12889) chromosome Y assembly as a reference for DNM detection across the 48.8 Mbp MSY. The *de novo* assembly-based approach increases by >2-fold the number of accessible base pairs when compared to HiFi read-based calling but increases by >7-fold the discovery of *de novo* SNVs (Methods, Figure 4.3D). In total, we identify 48 *de novo* SNVs in the MSY across the five G2-G3 males, ranging from 7-11 SNVs per Y transmission (mean 9.6, median 10). Only two SNVs map to the Y euchromatic regions, and one to the pericentromeric, with the remaining 45/48 to the Yq12 heterochromatic satellite regions (Figure 4.4C). We thus estimate the *de novo* SNV rate of 1.99×10^{-7} (95% CI: 1.59 - 2.39×10^{-7}) for the MSY combining both read- and assembly-based approaches. It is important to note that 13/45 (29%) of the DNMs had 100% identical matches elsewhere in the Yq12 region (but not

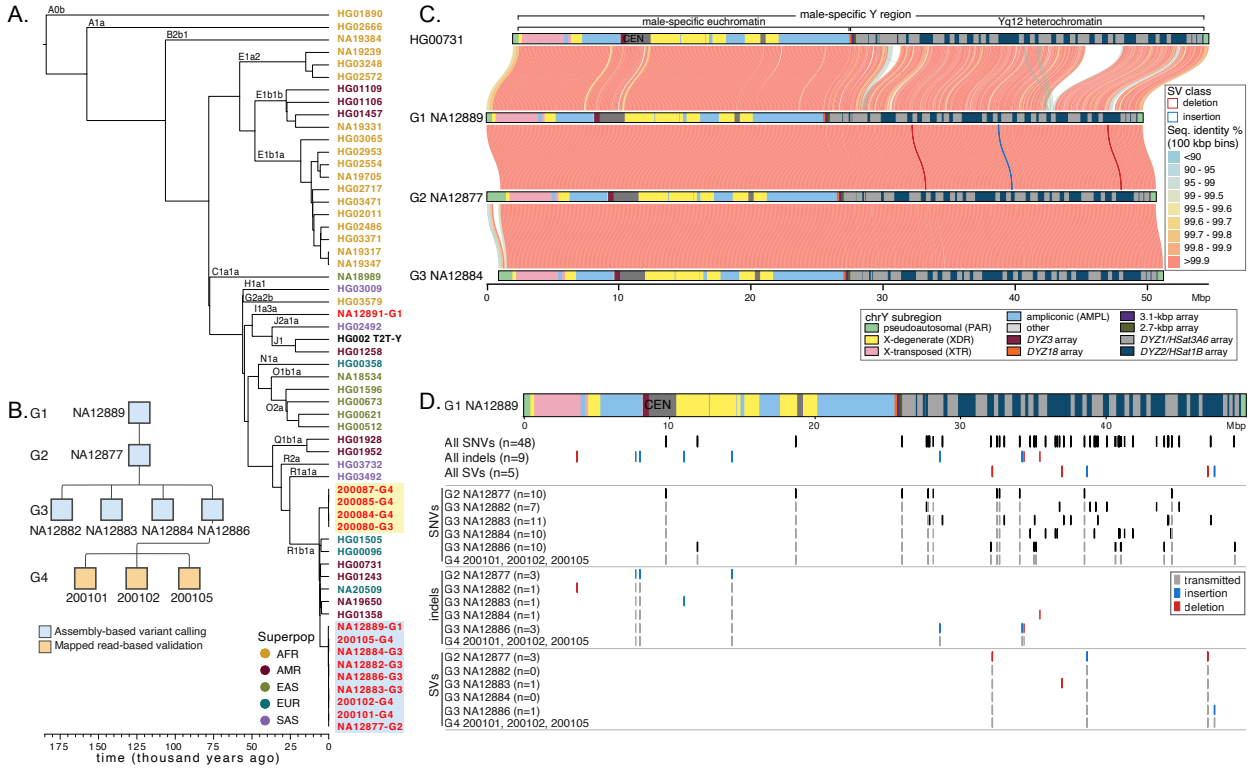


Figure 4.4: **Chromosome Y.** A. Y phylogeny incorporating 14 pedigree males and 44 males with previously published long-read Y assemblies [66]. Red text - pedigree males, other colors indicate 1000 Genomes Project superpopulations. Yellow background shows the Y chromosomes of G3 spouse (200080) and his three male offspring. Blue background indicates the 9 males with R1b1a-Z302 Y chromosomes analyzed in detail here (see B). B. Pedigree of the 9 males carrying the R1b1a-Z302 Y chromosomes. C. Pairwise comparison of Y assemblies: closely related Y from HG00731 (R1b1a1b1a1a2a5-Z225) and the most contiguous R1b1a-Z302 Y assemblies from three generations. Y-chromosomal sequence classes are shown with pairwise sequence identity between samples in 100 kbp bins, with QC-passed SVs identified in the pedigree males shown. D. Summary of chrY *de novo* mutations. Top - Y structure of G1 NA12889. Below the Y structure - all identified *de novo* mutations across G1-G3 Y assemblies. Bottom - breakout by mutation class and by sample. In light gray are shown *de novo* mutations that show evidence of transmission from G2 to G3-G4, and from G3 NA12886 to his male descendants in G4.

at orthologous positions) and could, therefore, result from interlocus gene conversion events (Methods) consistent with the DYZ1/HSat3A6 and DYZ2/HSat1B organization of the region [154]. We also identify a total of nine *de novo* indels (<50 bp, homopolymers excluded) ranging from 1-3 indels/sample (mean 1.8 events/Y transmission) and five *de novo* SVs (≥50 bp) (Figure 4.4D). The latter range from 2,416 to 4,839 bp in size, each affecting an entire DYZ2 repeat unit(s), with an average of one SV per Y transmission. Variants detected in the G3 parents of G4 are confirmed by both transmission and read data, supporting the high quality of the variant calls. Overall, 83% (52/63) of the DNMs identified on chrY (42/48 SNVs, 4/9 of indels and 5/5 SVs) are located in regions where short reads cannot be reliably mapped (mapping quality = 0).

4.5 Discussion

The origin, rate, and distribution of DNMs is arguably one of the most important aspects of human genetics and key to our understanding of genetic disease, phenotypic variation, and the evolution of our species [178]. Most studies [50, 61, 68, 91, 81] that establish DNM rates utilize short reads amongst large groups of trios and generally agree on 60-70 DNMs per generation; however, this largely excludes highly mutable regions of the genome, e.g., long TRs, SDs, and satellite sequences [180]. Our approach differs in that we generated a comprehensive multi-generational assembly-based resource with five orthogonal short- and long-read sequencing technologies with the aim to catalog transmitting and *de novo* variation of all classes—establishing a truth set for human genetic variation and all subsequent sequencing technologies. The multiplatform and multigenerational, assembly-based approach provides access to some of the most difficult regions of the genome, such as centromeres and heterochromatic regions on the Y chromosome. The use of parental references in addition to the standard GRCh38 and T2T-CHM13 references and the ability to confirm transmissions across subsequent generations improves both sensitivity and specificity. In this multigenerational pedigree, we estimate a range of 98-206 DNMs per transmission (average of 152 per generation), including 74.5 *de novo* SNVs and 7.4 non-TR indels, 4.4 centromeric *de novo*

SVs and SNVs. Among males, we find 12.4 *de novo* Y chromosome events per generation. We observe a strong paternal *de novo* bias (70-80%) and an increase with advancing paternal age not only for SNVs but also for indels and SVs.

We find that the rate of *de novo* SNVs varies by more than an order of magnitude depending on the genomic context. In particular, we observe elevated rates of *de novo* SNVs in repetitive regions both for germline and postzygotic events, consistent with recent human population-based analyses [110, 180] and theoretical predictions [175]. SD regions show an 88% increase (2.2 vs. 1.17×10^{-8}), which is driven by SDs with >95% identity. Although the number of validated SNV DNMs is still rather modest, we currently estimate that satellite DNA in the Yq12 heterochromatic region is at least 30 times more mutable (3.68×10^{-7} - 7.41×10^{-7} mutations per bp per generation) than autosomal euchromatin. The Yq12 region in particular has never been studied at base-pair resolution as it is largely missing from the GRCh38 reference and its complete assembly has only recently become possible [66, 154]. It is composed of thousands of short satellite DNA repeats (DYZ1/Hsat3A6 and DYZ2/Hsat1B) organized into blocks that are >98% identical, in total tens of megabase pairs in size [66, 154]. This, along with the fact that 29% of mutational changes match to non-orthologous sites in Yq12, is consistent with "interlocus gene conversion" driving this excess potentially as a result of increased sister chromatid exchange events [154]. While our DNM SNV rate estimate for Y euchromatic regions is comparable to previous pedigree-based work (∼22 Mbp, 1.81×10^{-8} mutations per bp per generation in this study compared to 2.87×10^{-8} mutations per bp per generation from Helgason et al. [68]), the SNV estimate for Yq12 is >20x higher.

Previous studies have predicted that 6-10% of DNMs are not germline in origin, but instead arise sometime after fertilization, giving rise to a mosaic variant [1, 162]. These studies distinguished between *de novo* postzygotic and germline SNVs using allele balance thresholds [1] or by identifying incomplete linkage to nearby SNVs across three generations [162], but long-read data can substantially increase the sensitivity of these approaches. We assign nearly every *de novo* SNV to a parental haplotype, defining a PZM by its incomplete linkage to that haplotype. We classify 14.6% of *de novo* SNVs as postzygotic in origin

(n=109 PZMs/745 *de novo* SNVs). Because all sequencing data in this study are derived from blood, we cannot demonstrate that every PZM is present in multiple tissues, but we can use transmission to the next generation as a proxy, as it reveals the mutation is also present in germ cells. In the four samples with sequenced children, we find that PZMs account for 12% of all SNVs transmitted to the next generation (n=33 PZMs/275 transmitted SNVs), an increase over previous estimates. Early cell divisions of human embryos are frequently error prone [80, 135] with an accelerated rate of cell division and these properties may contribute to the large fraction of PZMs with high (>25%) allele balance (43% are estimated to have high allele balance and 83% of these (n=20/24) are transmitted to the next generation). Such events would previously have been classified as germline but, consistent with PZM expectations, we find no paternal bias associated with these *de novo* variants (Figure 4.1D).

There are several limitations to this study. First, we were unable to successfully characterize *de novo* variation in the acrocentric regions due to the repetitive nature of the regions and rampant heterologous recombination. An important next step will be to assign acrocentric contigs to their respective chromosomes and assess patterns of mutation and ectopic recombination in regions predicted to be the most dynamic [62]. Second, we examined only one multigenerational family and familial variation is expected depending on the genetic background [150, 162, 170]. Many more families will be required to establish a reliable estimate of the mutation rate especially for complex regions of the genome. In that regard, it is perhaps noteworthy that efforts are underway to characterize an additional 10 CEPH pedigrees. Notwithstanding, this study highlights the fact that a single sequencing technology and a single human genome reference are insufficient to comprehensively estimate mutation rates. This is especially problematic in complex regions such as centromeres and heterochromatic regions of chromosome Y where assembly-based parent-to-offspring comparisons are required to catalog DNMs. More variation, including *de novo* variation, remains to be discovered—we were conservative in our callset requiring multiple technologies supporting the discovery of DNM, assessing transmission, where possible, to the next generation for all variants, and being careful to consider DNA from primary tissue as opposed to cell lines. It is noteworthy

that several studies with long-read sequencing technologies have claimed higher DNM rates [130, 131]. The multigenerational resource we generated will further refine these estimates and serve as a useful benchmark for new algorithms and new sequencing technologies similar to Genome in a Bottle [198].

4.6 Methods

Sample and DNA preparation

Family members from G2 and G3 were re-engaged for the purpose of updating informed consent and health history, and for enrolling their children (G4) and the marry- in parent (G3). Archived DNA from G2 and G3 was extracted from whole blood. Newly enrolled family members underwent informed consent and blood was obtained for DNA and cell lines. DNA was extracted from whole blood using the Flexigene system (Qiagen 51206). All samples are broadly consented for scientific purposes, which makes this dataset ideal for future tool development and benchmarking studies.

Sequence data generation

Sequencing data from orthogonal short- and long-read platforms were generated as follows:

Illumina data generation

Illumina WGS on G1-G3 was generated as previously described [162]. Illumina WGS on G4 and marry-in spouses for G3 were generated by the Northwest Genomics Center using the TruSeq library prep kit and sequenced to approximately 30x on the NovaSeq 6000 with paired end 150 bp reads.

PacBio HiFi sequencing

PacBio HiFi data were generated per manufacturer's recommendations. Briefly, DNA was extracted from blood samples as described or cultured lymphoblasts using the Monarch

HMW DNA Extraction Kit for Cells & Blood (New England Biolabs, T3050L). At all steps, quantification was performed with Qubit dsDNA HS (ThermoFisher, Q32854) measured on DS-11 FX (Denovix) and size distribution checked using FEMTO Pulse (Agilent, M5330AA & FP-1002-0275) HMW DNA was sheared with Megaruptor 3 (Diagenode, B06010003 & E07010003) using settings 28/30, 28/31, or 27/29 based on initial quality check to target a peak size of 22 kbp. After shearing, the DNA was used to generate PacBio HiFi libraries via the SMRTbell prep kit 3.0 (PacBio, 102-182-700). Size selection was performed either with diluted AMPure PB beads per the protocol, or with Pippin HT using a high-pass cutoff between 10-17 kbp based on shear size (Sage Science, HTP0001 & HPE7510). Libraries were sequenced either on the Sequel II platform on SMRT Cells 8M (PacBio, 101-389-001) using Sequel II sequencing chemistry 3.2 (PacBio,102-333-300) with 2-hour pre-extension and 30-hour movies, or on the Revio platform on Revio SMRT Cells (PacBio, 102-202-200) and Revio polymerase kit v1 (PacBio, 102-817-600) with 2-hour pre-extension and 24-hour movies.

ONT sequencing

To generate UL sequencing reads >100 kbp, we used ONT sequencing. Ultra-high molecular weight gDNA was extracted from the lymphoblastoid cell lines according to a previously published protocol [109]. Briefly, 3-5 $\times 10^7$ cells were lysed in a buffer containing 10 mM Tris-Cl (pH 8.0), 0.1 M EDTA (pH 8.0), 0.5% w/v SDS, and 20mg/mL RNase A for 1 hour at 37°C. 200 μg/mL Proteinase K was added, and the solution was incubated at 50°C for 2 hours. DNA was purified via two rounds of 25:24:1 phenol- chloroform-isoamyl alcohol extraction followed by ethanol precipitation. Precipitated DNA was solubilized in 10 mM Tris (pH 8.0) containing 0.02% Triton X-100 at 4°C for two days.

Libraries were constructed using the Ultra-Long DNA Sequencing Kit (ONT, SQK-ULK001) with modifications to the manufacturer's protocol: $\sim$40 μg of DNA was mixed with FRA enzyme and FDB buffer as described in the protocol and incubated for 5 minutes at RT, followed by a 5-minute heat-inactivation at 75°C. RAP enzyme was mixed with the

DNA solution and incubated at RT for 1 hour before the clean-up step. Clean-up was performed using the Nanobind UL Library Prep Kit (Circulomics, NB-900- 601-01) and eluted in 450 μL EB. 75 μL of library was loaded onto a primed FLO- PRO002 R9.4.1 flow cell for sequencing on the PromethION, with two nuclease washes and reloads after 24 and 48 hours of sequencing. All G1-G3 ONT base calling was done with guppy (v6.3.7).

Element (AVITI) sequencing

Element WGS data was generated per manufacturer's recommendations. Briefly, DNA was extracted from whole blood as described above. PCR-free libraries were prepared using mechanical shearing, yielding 350 bp fragments, and the Element Elevate library preparation kit (Element Biosciences, 830-00008). Linear libraries were quantified by qPCR and sequenced on AVITI 2 x 150 bp flow cells (Element Biosciences, not yet commercially available). Bases2Fastq Software (Element Biosciences) was used to generate demultiplexed FASTQ files.

Generation of phased genome assemblies

Phased genome assemblies were generated using two different algorithms, namely Verkko (v1.3.1 and v1.4.1) [151] and hifiasm (UL) with ONT support (v0.19.5) [20]. Due to active development of the Verkko and hifiasm algorithms, assemblies were generated with two different versions. Phased assemblies for G2-G3 were generated using a combination of HiFi and ONT reads using parental Illumina k-mers for phasing. To generate phased genome assemblies of G1, we still used a combination of HiFi and ONT reads with the Verkko pipeline and used Strand-seq to phase assembly graphs [69]. Lastly, G4 samples were assembled using HiFi reads only with hifiasm (v0.19.5) [106].

NOTE: Trio-based phasing with Verkko assigns maternal to haplotype 1 and paternal to haplotype2. In contrast, for hifiasm assemblies we report switched haplotype labeling such that haplotype 1 is paternal and haplotype 2 is maternal in order to match HPRC standard for hifiasm assemblies.

Evaluation of phased genome assemblies

To evaluate the base pair and structural accuracy of each phased assembly, we employed a multitude of assembly evaluation tools as well as orthogonal datasets such as PacBio HiFi, ONT, Strand-seq, Illumina, and Element data. We note that we fixed four haplotype switch errors in our assembly-based variant callsets to avoid biases in subsequent analysis.

Strand-seq validation

We used Strand-seq data to evaluate directional and structural accuracy of each phased assembly. First, we aligned selected Strand-seq libraries for each sample to the phased *de novo* assembly using BWA [103] (v0.7.17-r1188). Then we ran breakpointR [145] (v1.15.1) using aligned BAM files as input. Next, we created directional composite files using breakpointR function 'createCompositeFiles' followed by running breakpointR on such composite files using 'runBreakpointR' function. This provided us, for any given sample, with regions where strand-state changes across all single-cell Strand-seq libraries. Many such regions point to real heterozygous inversions. However, regions where Strand-seq reads mapped in opposite orientation with respect to surrounding regions are likely caused by misorientation. Also positions where the strand state of Strand-seq reads changes repeatedly in multiple libraries might be a sign of an assembly misjoin and such regions were investigated more closely to rule out any such large structural assembly inconsistencies.

Read to assembly alignment

To evaluate *de novo* assembly accuracy, we aligned sample-specific PacBio HiFi reads to their corresponding phased genome assemblies using Winnowmap (v2.03) with the following parameters: -I 10G -Y -ax map-pb –MD –cs -L –eqx

Flagger validation

Flagger [106] was used to detect misassemblies using HiFi read alignments to the assemblies and the assemblies aligned to the reference genome. Regions were flagged based on read alignment divergence and specific reference-biased regions. A reference-specific BED file (chm13v2.0.sd.bed) was used setting a maximum read divergence of 2% and specifying reference-biased blocks. These flagged regions were analyzed to identify collapses, false duplications, erroneous regions, and correctly assembled haploid blocks with the expected read coverage.

We used Flagger v0.3.3 (https://github.com/mobinasri/flagger) to run the "flagger_end_to_end" WDL. Required inputs include following:

1. Read-to-contig alignments - Winnowmap alignments of all HiFi reads to the assembly (hap1, hap2 and unassigned.fasta)

2. A combined assembly fasta file with hap1, hap2 and unassigned contigs

3. BAM alignments of assembly to the CHM13v2.0 reference hap1, hap2 and unassigned fasta files of the assembly were aligned to CHM13v2.0 using this pipeline:

https://github.com/mrvollger/asm-to-reference-alignment.

NucFreq validation

NucFreq [181] was used to calculate nucleotide frequencies for HiFi reads aligned using Winnowmap [78]. This was used to identify regions of collapses: where the second-highest nucleotide count exceeded [119], and misassembly: where all nucleotide counts were zero. NucFreq analysis pipeline is available at GitHub: https://github.com/mrvollger/NucFreq

Assembly base-pair quality

To evaluate the accuracy of the genome assembly, we employed a pipeline that uses Meryl (v1.0) to count the k-mers of length 21 from Illumina reads using the following command:

meryl k=21 count {input.fastq} output {output.meryl}

We then used Merqury (v1.1) [155], which compares the k-mers from the sequencing reads against those in the assembled genome and flags discrepancies where k-mers are uniquely found only in the assembly. These unique k-mers indicate potential base-pair errors. Merqury then calculates the quality value based on the k-mer survival rate, estimated from Meryl's k-mer counts, providing a quantitative measure to assess the completeness and correctness of the genome assembly.

Detection of small de novo variants

Following the parameters outlined in Noyes et al. [131], we called variants in HiFi data aligned to T2T-CHM13 using GATK HaplotypeCaller (v4.3.0.0) and DeepVariant (v1.4.0) and naively identified variants unique to each G2 and G3 sample [139]. We separated out SNV and indel calls and applied basic quality filters, such as removing clusters of three or more SNVs in a 1 kbp window. We combined this set of variant calls generated by a secondary calling method, (https://github.com/Platinum-Pedigree-Consortium/Platinum-Pedigree-Inheritance/blob/main/analyses/Denovo.md) and subjected all calls to the following validation process.

We validated both SNVs and indels by examining them in HiFi, ONT, and Illumina read data, excluding reads that failed to reach mapping quality (59 for long reads, 0 for short reads) thresholds. Reads with high base quality (>20) and low base quality (<20) at the variant site were counted separately. We retained variants that were present in at least two types of sequencing data for the child, and absent from high base quality parental reads. For SNV calls, we next examined HiFi data for every sample in the pedigree. We determined an SNV was truly *de novo* if it was absent from every family member that was not a direct descendant of the *de novo* sample. Finally, we examined the allele balance of every variant, determined which variants were in TRs, and reevaluated parental read data across all sequencing platforms, removing variants with noisy sequencing data or more than two low-quality parental reads supporting the alternate allele.

DNM phasing and postzygotic assignment

To determine the parent-of-origin for the *de novo* SNVs, we reexamined the long reads containing the *de novo* allele. First, we used our initial GATK variant calls to identify informative sites in an 80 kbp window around the DNM, selecting any SNPs where one allele could be uniquely assigned to one parent (for example, a site that is homozygous reference in a father and heterozygous in a mother). For every DNM, we evaluated every ONT and HiFi read that aligned to the site of the *de novo* allele and assigned it to either a paternal or maternal haplotype (if informative SNPs were available) by calculating an inheritance score as outlined in Noyes et al. [131]. DNMs that were exclusively assigned to maternal or paternal haplotypes were successfully phased, whereas DNMs on conflicting haplotypes were excluded from our final callset. Unphased variants were determined to be postzygotic in origin (n=7) if their allele balance was not significantly different across platforms (by a chi-squared test) and if their combined allele balance was significantly different from 0.5.

Once we assigned every read to a parental haplotype, we counted the number of maternal and paternal reads that had either the reference or alternate allele. We determined that a DNM was germline in origin if it was present on every read from a given parent's haplotype. Conversely, if a DNM was present on only a fraction of reads from a parental haplotype, we determined that it was postzygotic in origin.

Sex chromosome DNM calling and validation

To identify DNMs on the X chromosome, we applied the same strategy as autosomal variants, with one exception: we only used variant calls generated by GATK. For males, we reran GATK in haploid mode, such that it would only identify one genotype on the X chromosome.

To identify DNMs on the Y chromosome, we aligned male HiFi, ONT, and Illumina data to the G1-NA12889 chrY assembly and then called variants using GATK in haploid mode on the aligned HiFi data. We directly compared each male to his father, selecting variants unique to the son. We validated SNVs and indels by examining the father's HiFi, ONT, and

Illumina data and excluded any variants that were present in the parental reads, applying the same logic that we used for autosomal variants.

Callable genome and mutation rate calculations

We determined the size of the callable genome for each individual based on their HiFi data, using two criteria. First, we reran GATK HaplotypeCaller with the option "ERC BP_RESOLUTION" for every *de novo* sample and their parents to generate a genotype at every site in the genome. We excluded any site where both parents were not homozygous for the reference allele. For male sex chromosomes, we only considered the mother's genotype in the case of the X, and the father's genotype in the case of the Y. Second, we examined the HiFi data for each sample and their parents and excluded any site where all three members of the trio did not have at least one HiFi read that passed our mapping and base quality thresholds. Any sites that were not excluded were considered to be "callable" with our DNM pipeline. We intersected these sites with annotations to calculate the amount of callable space in a region such as SDs. To calculate the mutation rate on the autosomes in each sample, we divided the number of DNMs in a given region by twice the number of bases deemed to be callable.

Analysis of centromeric regions

To identify completely and accurately assembled centromeres from each genome assembly, we first aligned the genome assemblies generated via Verkko [151] or hifiasm (UL) [20] to the T2T-CHM13 reference genome1 using minimap2 [101] and the following parameters: -a –eqx -x asm20 -s 5000 -I 10G -t {threads}. Then, we filtered the whole-genome alignments to only those contigs that aligned to the centromeres in the T2T-CHM13 reference genome. We checked if these centromeric contigs spanned the centromeres by checking to see if they contained sequence from the p- and the q-arms in the regions directly adjacent to the centromere. Then, we validated the assembly of the centromeric regions by aligning native PacBio HiFi data from the same source genome to each whole-genome assembly using pbmm2 (v1.1.0;

https://github.com/PacificBiosciences/pbmm2) and the following command: align –log-level DEBUG –preset SUBREAD –min-length 5000 -j {threads}, and next assessed the assemblies for uniform read depth across the centromeric regions via NucFreq18. We also aligned native ONT data >30 kbp in length from the same source genome to each whole-genome assembly using minimap2 [101] (v2.28) and assessed the assemblies for uniform read depth across the centromeric regions via IGV browser [158].

To identify *de novo* SVs and SNVs within each centromeric region, we first aligned each child's genome assembly to the relevant parent's genome assembly using minimap2 [101] and the following parameters: -a –eqx -x asm20 -s 5000 -I 10G -t {threads}. Then, we used the resulting PAF file to identify *de novo* SVs and SNVs using SVbyEye [144], filtering our results to only those centromeres that were completely and accurately assembled. We checked each SV and SNV call with NucFreq [181], Flagger [106], and native ONT data to ensure that the underlying data supported each call.

Construction and dating of Y phylogeny

The construction and dating of Y-chromosomal phylogeny for 58 total samples, combining the 14 pedigree males from the current study with 44 individuals, for which long-read-based Y assemblies have previously been published, was done as described previously in detail [66]. In short, all sites were called from the Illumina high-coverage data [162] of the 14 pedigree males using the approximately 10.4 Mbp of Y-chromosomal sequence previously defined as accessible to short-read sequencing [148]. BCFtools [31, 98] (v1.16) was used with minimum base quality 20, mapping quality 20, and ploidy 1. SNVs within 5 bp of an indel call (SnpGap) and all indels were removed, followed by filtering all calls for a minimum read depth of 3 and a requirement of $\geq$85% of reads covering the position to support the called genotype. The VCF was merged with a similarly filtered VCF from Hallast et al. [66] for the 44 individuals using BCFtools, followed by removal of sites with $\geq$5% of missing calls, that is, missing in more than 3 out of 58 samples, were removed using VCFtools [30] (v0.1.16). After filtering, a total of 10,404,104 sites remained, including 13,443 variant sites.

The Y haplogroups of each sample were predicted as previously described [65] and correspond to the International Society of Genetic Genealogy nomenclature (ISOGG, https://isogg.org, v15.73). A coalescence-based method implemented in BEAST [38] (v1.10.4) was used to estimate the ages of internal nodes. RAxML [168] (v8.2.10) with the GTRGAMMA substitution model was used to construct a starting maximum-likelihood phylogenetic tree for BEAST. Markov chain Monte Carlo samples were based on 200 million iterations, logging every 1,000 iterations, with the first 10% of iterations discarded as burn-in. A constant-sized coalescent tree prior, the GTR substitution model, accounting for site heterogeneity (gamma), and a strict clock with a substitution rate of 0.76×10^{-9} (95% CI = 0.67×10^{-9} – 0.86×10^{-9}) single-nucleotide mutations per bp per year was used [52]. A prior with a normal distribution based on the 95% CI of the substitution rate was applied. A summary tree was produced using Tree-Annotator (v1.10.4) and visualized using the FigTree software (v1.4.4).

Identification of sex-chromosome contigs

Detailed analysis of Y-chromosomal DNMs focused on seven males (R1b1a-Z302 Y haplogroup, G1-NA12889, G2-NA12877, G3-NA12882, G3-NA12883, G3-NA12884 and G3-NA12886) for which phased Verkko assemblies were generated. Contigs containing X- and Y-chromosomal sequences were identified and extracted from the whole- genome assemblies as previously described [66]. In addition, the pseudoautosomal regions from the G1 grandmother NA12890 and G2 mother NA12878 genome assemblies were identified by aligning the respective sequences from the T2T-CHM13 reference genome to these assemblies using minimap2 [101] (v2.26).

Annotation of Y-chromosomal subregions

The annotation of Y-chromosomal subregions of the Verkko assemblies was performed using both the GRCh38 and T2T-CHM13 Y reference sequences as previously described [66]. The centromeric α-satellite repeats for the purpose of Y subregion annotation were identified

using RepeatMasker (v4.1.2-p1) with default parameters. The Yq12 repeat annotations were generated using HMMER [121] (v3.3.2dev) with published DYZ1, DYZ2, DYZ18, 2k7bp and 3k1bp sequences [66], followed by manual checking of repeat unit orientation and distance from each other. Dot plots to compare Y-chromosomal sequences were generated using Gepard [93] (v2.0).

Detection and validation of DNMs

Human Y chromosomes vary extensively in the size and composition of repetitive regions [66], including the T2T-CHM13 Y (haplogroup J1a-L816) and the R1b1a-Z302 haplogroup Y chromosomes carried by the seven pedigree males analyzed in detail here (Figure 4.1A, Figure C.2). For this reason, the Y assembly of the G1 grandfather NA12889 was used as a reference for DNM detection (Figure C.3). The DNMs were called from the Y assemblies of five G2 (NA12877) and G3 (NA12882, NA12883, NA12884, NA12886) males using Dipcall [102] (v0.3) with the default parameters recommended for male samples. Variants were identified from the male-specific Y regions only, i.e., the pseudoautosomal regions were excluded from this analysis. All identified variants were filtered as follows: any variant calls overlapping with regions flagged by Flagger or NucFreq in either reference or query assembly were filtered out. For SNVs, the final filtered calls were supported by 100% of HiFi reads (i.e., no reads supported the reference allele in offspring or alternative allele in the father) and ONT reads mapped to both the reference and each individual assembly were checked for support.

For indels ($\leq$50 bp), homopolymer tracts were excluded from the analysis, while the rest of the calls were validated using the read data (HiFi, ONT, Illumina) as follows. Individual reads mapped to the reference (G1 NA12889 Y assembly) and covering the indel call plus 150 bp of flanking sequence were extracted from all samples using subseq (https://github.com/EichlerLab/subseq), followed by alignment using MAFFT (v7.508) with default parameters [83, 84]. All alignments were manually checked and any calls where the HiFi data had two or more reads supporting a reference allele and one or more reads supporting an alternate allele were removed. All final SNV and indel calls were additionally

supported (if unique mapping to the region was possible) by both Illumina and Element read data mapped to the reference.

For all SV calls, HiFi read depth for reference and alternative alleles were visualized and SVs in regions showing high levels of read depth variation coinciding with clusters of SNVs with >10% of reads supporting an alternative allele removed. HiFi and ONT reads mapped to both the reference and individual assemblies were checked for support. For all variants, concordance with the expected transmission through generations was confirmed. Additionally, the HiFi data available for three G4 males (200101, 200102 and 200105) were checked for support of the identified variants.

Y-chromosomal DNM rate calculation

The assembly-based DNM rates were calculated for each of the five males based on the accessible regions of each individual Y assembly (i.e., any regions flagged by Flagger and/or NucFreq were removed).

Chapter 5
DISCUSSION

When I started working independently on my first project, I was given a framework to understand research: imagine humanity is on an island in the middle of the ocean, and every rock, pebble, and grain of sand on that island is a bit of knowledge that we have discovered. The stone you're sitting on? It's the Human Genome Project. That boulder jutting out into the sea? There's Turing machines. Our job as researchers is to wade into the ocean, gather a pail of sand, and contribute it to the shore. Here, I'll sort through the contents of my pail, and tell you how I think we can go about filling the next.

5.1 My sandcastle

Through my incredible good fortune and the hard work of my labmates and collaborators, I have had the opportunity to work with a wealth of cutting-edge sequencing data generated by some incredibly talented scientists. While we were not the first to use long-read sequencing to identify *de novo* variation [117, 136, 152], I hope we demonstrated the power that long reads bring to the study of small variants.

In Chapter 2, I developed an approach to use sequencing data generated by three disparate technologies to discover and validate *de novo* mutations (DNMs). The object of this project was not only to characterize DNMs in a family but also to try out a variety of new sequencing and data analysis techniques to see what would stick. By using long-read HiFi and ONT data, I was able to search for DNMs in an additional 5% of the genome that was mostly composed of repetitive DNA. With the gain in search space came a concomitant gain in discovery: I found 20% more DNMs using long-read data than with short-read data, including a handful of variants in segmental duplications (SDs; n=4) and on the edges of centromeres (n=3).

All told, we found a total of 195 DNMs across two children, for a DNM rate of 1.41×10^{-8} DNMs/bp/gen. Of course, this initial study had a few limitations. First, and perhaps most obviously, we still could not access the whole genome, leaving over 10% untouched—I'll revisit this problem a little later. Next, we examined just two samples, meaning we did not have enough statistical power to say anything about the kinds of new DNMs we were discovering. Finally, we treated postzygotic mutations (PZMs) as something of an afterthought, sifting them out of discarded variant calls that did not initially meet our quality thresholds. Luckily, I had an opportunity to address these issues in Chapter 3.

We expanded the scope of the project from one family to dozens of families, creating an opportunity to see how well my results could replicate. I rewrote my code, devised new variant filters, and, most importantly, developed an approach to discern germline from postzygotic variants, working with 42 samples from 24 families. I had expected that this new project would further increase our estimate of the mutation rate, and while I found an average of 96 DNMs in each sample, a comparable amount to my previous study, my estimate of the germline mutation rate actually dropped to 1.31×10^{-8} DNMs/bp/gen. Using long reads to classify PZMs, I found that 15% of *de novo* single-nucleotide variants (SNVs) were postzygotic in origin, a higher proportion than had been observed in previous studies [1, 162]. To be fair, I had an advantage. Many *de novo* studies use a minimum allele balance (AB) threshold, which eliminates any variants that are observed in less than 25% of all sequencing reads [81, 91, 176]. I did not need to apply that filter, because I had the power of three orthogonal sequencing platforms to validate my PZM calls, allowing me to trust low AB variants that were observed in multiple data types. I found an average of 13.1 PZMs per sample, for a PZM rate of 0.23×10^{-8} PZMs/bp/gen. Had I not distinguished between germline and postzygotic variants, this study would yield a mutation rate of 1.54×10^{-8} DNMs/bp/gen, living up to my expectations; however, finding the PZMs is, I think, much more satisfying and biologically interesting.

With variants from 42 samples, I had enough statistical power to unpack a few differences between germline DNMs and PZMs. Some differences, like a postzygotic enrichment for A>T

mutations, had been previously reported [162], while others, like the enrichment of A>C mutations and depletion of CpG>TpG substitutions, were novel. In addition to the discovery of PZMs, I was also able to evaluate the hypermutability of repetitive regions. I found a 38% enrichment of germline SNVs in SDs and an astonishing threefold increase of PZMs in the same regions. When combined, DNMs and PZMs are enriched 66% in duplications, which matches estimates made by using population genetics methods to examine segregating variation [108, 180]. In addition, the novel mutational signatures I found in PZMs at least partly match the observed mutational spectrum in SDs.

With this in mind, I propose the hypothesis that much of the variation we observe in SDs is driven by interlocus gene conversion (IGC) in the first few cell divisions after fertilization. Looking back in time to shortly before fertilization, we know that the tightly packed DNA in differentiated sperm are inaccessible for DNA repair, despite accumulating damage, and that spermatozoa themselves do not contain the full arsenal of DNA repair genes [73, 79, 124, 129, 166]. Once fertilization occurs, there is no transcription until the 4-8-cell stage, meaning that new embryos mostly rely on whatever proteins were present in the oocyte to repair any DNA breaks or transcriptional mistakes that occur in the first few cell divisions [14, 15]. These early divisions in embryogenesis are more error-prone than later divisions, creating more mistakes for the limited amount of repair proteins to fix [28]. Perhaps these repair enzymes become saturated in the early embryo, making homology-driven repair highly preferred as a means to fix errors in duplicated sequences [2, 22, 34, 86]. This preference could lead to a rise in IGC, a means of homology-driven repair that uses paralogous sequence as a template [127]. An increased rate of IGC could result in our observed increase of PZMs in SDs. Furthermore, we expect mutations that occur as a result of IGC to have a different mutational profile than mutations that arise from normal replication and repair errors in unique DNA.

Obviously, I cannot confidently say if the increased rate of PZMs in SDs contributes to the segregating variation in those regions if those postzygotic variants are not transmitted to the next generation. I don't expect every PZM to make it into the germline of the

sample in which it arose, since with every round of cell division, cells get put onto different developmental pathways. In Chapter 4, I was able to see both DNM and PZM transmission in a four-generation pedigree using a modified version of my variant discovery pipeline from Chapter 3 adapted to use the structure of the pedigree itself to weed out inherited variation masquerading as *de novo*. In total, I found 249 germline DNMs and 55 PZMs in samples with children in the dataset; 97% of DNMs and 60% of PZMs were transmitted to the next generation. To me, this is one of the most exciting results of this dissertation. Not only does it confirm that my DNM and PZM calls are indeed real variants, but it also demonstrates that these PZMs contribute to segregating variation in the population.

Another interesting observation from the pedigree study is that compared to the autism quads, I observed a similar PZM rate of 0.20×10^{-8} PZMs/bp/gen, yet a much lower germline mutation rate of 1.17×10^{-8} DNMs/bp/gen. While we know that mutation rates vary between families, this difference can be explained by parental age [61, 81, 91, 162]. In Chapters 3 and 4, I found parental age has no effect on PZMs though does contribute to an increase in germline variants. The average age of both fathers and mothers at the time of their children's birth was about 3.5 years older for the autism quads compared to the samples in the pedigree. We observe an additional 1.5 DNMs per year of additional paternal age and 0.5 DNMs per year of maternal age, so we would expect on average seven fewer mutations per sample in the pedigree. If we do a little back of the envelope math and add those mutations into the pedigree mutation rate, it comes out to 1.3×10^{-8} DNMs/bp/gen—nearly the same as I observed in the quads!

If we want to accurately compare mutation rate calculations across different sample sets, we need to examine either a yearly mutation rate or a generational rate scaled to the same average parental age. While many *de novo* studies provide the ingredients to calculate these rates, not every study explicitly spells them out [91, 150, 162]. To compare my results to previous studies, I estimated the per-year mutation rates by dividing each study's per-generation mutation rate by the average age of parents in their dataset, following the precedent of Jonsson et al. (Table 5.1) [81]. After adjusting for parental age, I find that my DNM rate falls

perfectly within the 1.2-1.3 DNMs/bp/gen range previously reported. However, my germline mutation rate does not contain any PZMs. With the exception of Sasani et al. [162], the studies I compare against did not make efforts to discern between germline mutations and PZMs, so I suspect their mutation rates are somewhat inflated. If I include PZMs in my mutation rate and adjust it for 30-year-old parents, I find a total mutation rate of 1.48-1.52$\times10^{-8}$ SNVs/bp/gen in Chapters 3 and 4, an approximately 25% increase over previous studies.

In summary, over the last five years, I have shown that long-read sequencing data can improve our power to identify DNMs and discern whether they originated in a parental germline or in early embryogenesis, increasing the estimated contribution of PZMs from 10% to 15% of all DNMs [1, 162]. I calculated an age-adjusted germline mutation rate of 1.2-1.3$\times10^{-8}$ DNMs/bp/gen, and an age-agnostic PZM rate of 0.23$\times10^{-8}$ PZMs/bp/gen. Lastly, I confirmed my hypothesis that duplicated regions are in fact hypermutable, and I found evidence that some of the segregating variation in SDs is postzygotic in origin. However, I left approximately 9% of the genome untouched, and given the repetitive nature of those regions, I suspect there are still many DNMs waiting to be discovered.

The pioneers who find these new mutations have their work cut out for them, and the generation of an iron-clad truth set may be a fool's errand. If there's one thing I learned in my time DNM hunting, it's that a DNM can never be validated enough. At the end of the day, someone must decide what level of uncertainty is acceptable for a variant call, especially one in a repetitive region, and we all have different thresholds that we can tolerate. Studying DNMs has made me a more careful scientist and instilled in me a healthy skepticism of the variation we report as "true." I'm sure one day a new sequencing technology will be developed, new assembly methods will be written, and my own DNM calls will come under careful scrutiny. I've done my best to ensure they withstand the test of time, though I know better than to expect perfection.

study	per generation rate	average parental age	per year rate	rate for 30 year old parents
Chapter 4	1.17×10^{-8}	27.3	4.33×10^{-10}	1.30×10^{-8}
Chapter 3	1.31×10^{-8}	31.3	4.20×10^{-10}	1.26×10^{-8}
Chapter 2	1.41×10^{-8}	35.4	3.99×10^{-10}	1.20×10^{-8}
Sasani et al. 2019	1.10×10^{-8}	27.6	3.99×10^{-10}	1.20×10^{-8}
Jonsson et al. 2017	1.29×10^{-8}	30.1	4.24×10^{-10}	1.27×10^{-8}
Rahbari et al. 2015	1.28×10^{-8}	29.8	4.30×10^{-10}	1.29×10^{-8}
Kong et al. 2012	1.20×10^{-8}	29.7	4.04×10^{-10}	1.21×10^{-8}

Table 5.1: **Age-adjusted mutation rates.** Estimates of mutation rates from Illumina-based DNM studies (in colors) and my thesis chapters [81, 91, 150, 162, 176]

.

5.2 Beyond the pail

My alignment-based method cannot handle the most complicated regions of the genome, where the mutation rate can be up to 30-fold higher. While it is perhaps naive to assume that one day we will be able to identify every DNM in an individual with 100% certainty, it shouldn't stop us from trying. So, what steps can we take to get us closer?

As is the case with most challenges we tackle in the Eichler lab, the path forward lies with genome assembly. Apart from a few unsuccessful tests, all of my efforts to identify *de novo* variants relied on aligning sequencing reads to a reference genome. Even with a complete reference, like T2T-CHM13, this strategy will never be able to capture the variation present in centromeres or the Y chromosome. These regions are characterized by heterochromatic satellite DNA, or large arrays of tandem repeats, spanning up to megabases of sequence [24, 25]. If aligning reads is challenging in small repeats, imagine how they look in hundreds of thousands of base pairs of repeated repeats. It's not pretty [111]. What's more, these regions are marked by extensive structural variation, with a *de novo* structural variant observed once every three transmissions [141]. So, even if reads could perfectly align to a reference sequence, the size of these regions differs between individuals, making them impossible to accurately represent without an assembly.

In Chapter 4, we demonstrated the power of assembling the Y chromosome to identify *de novo* variants. Compared to my read-based method, the assembly method identified eight times as many DNMs in about 2.6 times as much genomic space. It is worth mentioning that even by comparing a son's Y chromosome to his father's, only 80% of the Y chromosome was correctly assembled across both samples and accessible for variant calling. If the mutation rate in the newly accessible Yq12 region was 32 times what I observed in the autosomes, who knows what mutations are waiting for us in the remaining inaccessible regions! The Y chromosome was a great proof-of-concept for an assembly-based approach, as it is essentially a haploid case. Males only have one copy of the chromosome, and there is no question from which parent it was inherited. There is a notable exception, the pseudoautosomal regions (PARs), where crossovers with the X chromosome occur. PARs are essentially diploid, so read data cannot be uniquely attributed to the X or Y chromosome, making assembly much more challenging [154]. Shrewd readers will notice that PARs were excluded from our *de novo* analysis, and our tools will need to be further developed for DNM calling in a diploid case.

My guess is that future generations of *de novo* studies will be performed by assembling a child and both their parents and directly comparing the sequences of their chromosomes. There are some existing tools for this, such as PAV [43], which compares an assembled genome to a reference, and Dipcall [102], which compares the haplotypes from a phased assembly to a reference. In both cases, a parental assembly can be used as the reference, but to optimize for *de novo* calling, one would have to determine which set of chromosomes each parent transmitted to their child (and perhaps their recombination breakpoints), a trivial task with the Y chromosome that would require careful bookkeeping on the autosomes. Another complication is that assembly-based variant callers tend to perform better with larger variants and can generate many false positive SNV calls, especially in sequences with multiple copies. SNVs in these repetitive regions can be difficult to validate, as long-read data cannot map uniquely and traditional DNM validation methods, like Sanger sequencing, cannot be applied in repeats. In Chapter 4, we showed that some alignment-based validation

can be performed if mapping quality filters are not applied to read data, but on a whole-genome scale and across many samples, an assembly-based variant caller with more precise SNV calling will be necessary for *de novo* discovery.

One improvement I can foresee is an analog for joint genotyping, a method that is ubiquitous in alignment-based calling pipelines [140, 196]. Joint genotypers simultaneously call variants across multiple samples, ensuring consistent representation of sequence differences and uniform variant quality calculations, reducing the number of false positive variant calls across a dataset. An assembly-based equivalent might be to use a pangenome strategy to directly compare multiple assemblies [44, 55, 57, 104]. In addition to calling precision, validation methods will need to be improved as well. Perhaps the solution to validation lies in a concept that underlies genome assembly: the humble k-mer, defined as a sequence of length k. A *de novo* variant should create a k-mer unique to a child and absent from the parents, so by comparing all the k-mers in a trio, one can determine which variants are indeed *de novo* [12, 54]. However, an approach like this would fail to validate IGC events, in which the resulting k-mer would look identical to one from the homologous sequence that mediated the mutation. Using a larger k-mer size could help to solve this problem by capturing uniquely identifying sequence, but this comes with the trade-off of additional computational expense. So, while k-mers might provide some help validating variants in regions where reads cannot map uniquely, they will also struggle with one of the most common types of mutations in these regions.

Luckily, even though identifying and validating DNMs from assemblies will remain a challenge in the near future, distinguishing PZMs is fairly straightforward and follows from current methods development for the identification of somatic mutations. Using a child's assembled genome as a reference, one can align sequencing data and use traditional variant-calling methods to discover alternate alleles [187, 193]. By the same logic that I used to discern PZMs, sites that are multi-allelic should be somatic in origin. The trick here is distinguishing between a mutation that occurred early in embryogenesis and a somatic mutation that occurred later in life. The only way to truly trace the origin of a PZM is to sample

DNA across multiple tissues and identify variants that are common across most tissue types. So, for comprehensive PZM discovery at scale, we will probably need long-read sequencing to get a little cheaper.

As sequencing becomes increasingly accessible, I suspect DNM discovery will move further into the clinical space. Clinicians already search for DNMs to make diagnoses, even using long-read sequencing to compare the genomes of a child and their parents to find clues about the cause of a child's phenotype [71, 120]. However, these tests are rare and expensive, not to mention further complicated by our imperfect understanding of variant effects [29]. Our work in variant discovery is complemented by research determining how any given mutation might affect the function of cells or cause disease in an individual. If a mutation does contribute to a developmental disease, like many of the DNMs we study, identifying it in a patient may help guide treatment but cannot reverse the disease, at least with the technology we have today.

A new frontier for DNMs is *in vitro* fertilization, where genetic screening is common but *de novo* discovery is not. Although the use of assistive reproductive technologies (ART) in itself does not create DNMs [167], we might expect that some of the factors that increase parental infertility, such as age, could lead to a higher mutation rate in families using ART [186]. Further, the change in environment for the newly fertilized embryo may lead to an increased rate of postzygotic mutations [23]. Outside of the cost of sequencing, one challenge with preimplantation genetic testing is the small amount of DNA that can be recovered when biopsying just a few cells from an embryo. In order to get enough DNA for analysis, whole-genome amplification must be performed, which can introduce variants that appear *de novo* but are actually amplification artifacts [123, 179, 183]. With careful filtering, these artifacts can be removed from DNM callsets, leaving behind truly *de novo* events [123]. A few studies have already shown on a small scale that preimplantation screening for DNMs can identify disease-causing variation [123, 137, 179], and as we refine DNM discovery methods and further characterize variant effects, we may be able to vastly reduce the rate of genetic disease in babies born with ART. Achieving this goal will require not only rigorous scientific inquiry,

but also ethical inquiry. I would be remiss not to mention the thorny ethical space that preimplantation genetic screening occupies, as it opens the door to discriminating against embryos with "benign polymorphisms" [190]. It is up to us as a society to thread the needle, finding the correct path to reducing the harm caused by genetic disease without veering towards eugenics.

BIBLIOGRAPHY

[1] Rocio Acuna-Hidalgo, Tan Bo, Michael P. Kwint, Maartje van de Vorst, Michele Pinelli, Joris A. Veltman, Alexander Hoischen, Lisenka E. L. M. Vissers, and Christian Gilissen. Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *The American Journal of Human Genetics*, 97(1):67–74, July 2015.

[2] Robert John Aitken. Role of sperm DNA damage in creating de-novo mutations in human offspring: the 'post-meiotic oocyte collusion' hypothesis. *Reproductive BioMedicine Online*, 45(1):109–124, July 2022. Publisher: Elsevier.

[3] Nicolas Altemose, Glennis A. Logsdon, Andrey V. Bzikadze, Pragya Sidhwani, Sasha A. Langley, Gina V. Caldas, Savannah J. Hoyt, Lev Uralsky, Fedor D. Ryabov, Colin J. Shew, Michael E. G. Sauria, Matthew Borchers, Ariel Gershman, Alla Mikheenko, Valery A. Shepelev, Tatiana Dvorkina, Olga Kunyavskaya, Mitchell R. Vollger, Arang Rhie, Ann M. McCartney, Mobin Asri, Ryan Lorig-Roach, Kishwar Shafin, Julian K. Lucas, Sergey Aganezov, Daniel Olson, Leonardo Gomes de Lima, Tamara Potapova, Gabrielle A. Hartley, Marina Haukness, Peter Kerpedjiev, Fedor Gusev, Kristof Tigyi, Shelise Brooks, Alice Young, Sergey Nurk, Sergey Koren, Sofie R. Salama, Benedict Paten, Evgeny I. Rogaev, Aaron Streets, Gary H. Karpen, Abby F. Dernburg, Beth A. Sullivan, Aaron F. Straight, Travis J. Wheeler, Jennifer L. Gerton, Evan E. Eichler, Adam M. Phillippy, Winston Timp, Megan Y. Dennis, Rachel J. O'Neill, Justin M. Zook, Michael C. Schatz, Pavel A. Pevzner, Mark Diekhans, Charles H. Langley, Ivan A. Alexandrov, and Karen H. Miga. Complete genomic and epigenetic maps of human centromeres. *Science*, 376(6588):eabl4178, April 2022. Publisher: American Association for the Advancement of Science.

[4] Joon-Yong An, Kevin Lin, Lingxue Zhu, Donna M. Werling, Shan Dong, Harrison Brand, Harold Z. Wang, Xuefang Zhao, Grace B. Schwartz, Ryan L. Collins, Benjamin B. Currall, Claudia Dastmalchi, Jeanselle Dea, Clif Duhn, Michael C. Gilson, Lambertus Klei, Lindsay Liang, Eirene Markenscoff-Papadimitriou, Sirisha Pochareddy, Nadav Ahituv, Joseph D. Buxbaum, Hilary Coon, Mark J. Daly, Young Shin Kim, Gabor T. Marth, Benjamin M. Neale, Aaron R. Quinlan, John L. Rubenstein, Nenad Sestan, Matthew W. State, A. Jeremy Willsey, Michael E. Talkowski, Bernie Devlin, Kathryn Roeder, and Stephan J. Sanders. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science*, 362(6420):eaat6576, December 2018. Publisher: American Association for the Advancement of Science.

[5] Barbara Arbeithuber, Andrea J. Betancourt, Thomas Ebner, and Irene Tiemann-Boege. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc Natl Acad Sci U S A*, 112(7):2109–2114, February 2015.

[6] Peter Audano and Fredrik Vannberg. KAnalyze: a fast versatile pipelined K-mer toolkit. *Bioinformatics*, 30(14):2070–2072, July 2014.

[7] Jonathan R. Belyeu, Harrison Brand, Harold Wang, Xuefang Zhao, Brent S. Pedersen, Julie Feusier, Meenal Gupta, Thomas J. Nicholas, Joseph Brown, Lisa Baird, Bernie Devlin, Stephan J. Sanders, Lynn B. Jorde, Michael E. Talkowski, and Aaron R. Quinlan. *De novo* structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *The American Journal of Human Genetics*, 108(4):597–607, April 2021.

[8] Lucie A Bergeron, Søren Besenbacher, Jaco Bakker, Jiao Zheng, Panyi Li, George Pacheco, Mikkel-Holger S Sinding, Maria Kamilari, M Thomas P Gilbert, Mikkel H Schierup, and Guojie Zhang. The germline mutational process in rhesus macaque and its implications for phylogenetic dating. *GigaScience*, 10(5):giab029, 05 2021.

[9] Lucie A Bergeron, Søren Besenbacher, Tychele Turner, Cyril J Versoza, Richard J Wang, Alivia Lee Price, Ellie Armstrong, Meritxell Riera, Jedidiah Carlson, Hwei-yen Chen, Matthew W Hahn, Kelley Harris, April Snøfrid Kleppe, Elora H López-Nandam, Priya Moorjani, Susanne P Pfeifer, George P Tiley, Anne D Yoder, Guojie Zhang, and Mikkel H Schierup. The mutationathon highlights the importance of reaching standardization in estimates of pedigree-based germline mutation rates. *eLife*, 11:e73577, jan 2022.

[10] Søren Besenbacher, Christina Hvilsom, Tomas Marques-Bonet, Thomas Mailund, and Mikkel Heide Schierup. Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nat Ecol Evol*, 3(2):286–292, February 2019. Publisher: Nature Publishing Group.

[11] Søren Besenbacher, Siyang Liu, José M. G. Izarzugaza, Jakob Grove, Kirstine Belling, Jette Bork-Jensen, Shujia Huang, Thomas D. Als, Shengting Li, Rachita Yadav, Arcadio Rubio-García, Francesco Lescai, Ditte Demontis, Junhua Rao, Weijian Ye, Thomas Mailund, Rune M. Friborg, Christian N. S. Pedersen, Ruiqi Xu, Jihua Sun, Hao Liu, Ou Wang, Xiaofang Cheng, David Flores, Emil Rydza, Kristoffer Rapacki, John Damm Sørensen, Piotr Chmura, David Westergaard, Piotr Dworzynski, Thorkild I. A. Sørensen, Ole Lund, Torben Hansen, Xun Xu, Ning Li, Lars Bolund, Oluf Pedersen, Hans Eiberg, Anders Krogh, Anders D. Børglum, Søren Brunak, Karsten Kristiansen, Mikkel H. Schierup, Jun Wang, Ramneek Gupta, Palle Villesen, and Simon Rasmussen.

Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun*, 6(1):5969, January 2015. Publisher: Nature Publishing Group.

[12] Jörn Bethune, April Kleppe, and Søren Besenbacher. A method to build extended sequence context models of point mutations and indels. *Nat Commun*, 13(1):7884, December 2022. Publisher: Nature Publishing Group.

[13] Adrian P. Bird. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, 8(7):1499–1504, April 1980.

[14] Susanna E. Brantley and Stefano Di Talia. Cell cycle control during early embryogenesis. *Development*, 148(13):dev193128, June 2021.

[15] Peter Braude, Virginia Bolton, and Stephen Moore. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature*, 332(6163):459–461, March 1988. Publisher: Nature Publishing Group.

[16] G R Bunin, A T Meadows, B S Emanuel, J D Buckley, W G Woods, and G D Hammond. Pre- and postconception factors associated with sporadic heritable and nonheritable retinoblastoma. *Cancer Res*, 49(20):5730–5735, October 1989.

[17] Malgorzata Bzymek and Susan T. Lovett. Instability of repetitive DNA sequences: The role of replication in multiple mechanisms. *Proceedings of the National Academy of Sciences*, 98(15):8319–8325, July 2001. Publisher: Proceedings of the National Academy of Sciences.

[18] Mark J. P. Chaisson, Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, Oscar L. Rodriguez, Li Guo, Ryan L. Collins, Xian Fan, Jia Wen, Robert E. Handsaker, Susan Fairley, Zev N. Kronenberg, Xiangmeng Kong, Fereydoun Hormozdiari, Dillon Lee, Aaron M. Wenger, Alex R. Hastie, Danny Antaki, Thomas Anantharaman, Peter A. Audano, Harrison Brand, Stuart Cantsilieris, Han Cao, Eliza Cerveira, Chong Chen, Xintong Chen, Chen-Shan Chin, Zechen Chong, Nelson T. Chuang, Christine C. Lambert, Deanna M. Church, Laura Clarke, Andrew Farrell, Joey Flores, Timur Galeev, David U. Gorkin, Madhusudan Gujral, Victor Guryev, William Haynes Heaton, Jonas Korlach, Sushant Kumar, Jee Young Kwon, Ernest T. Lam, Jong Eun Lee, Joyce Lee, Wan-Ping Lee, Sau Peng Lee, Shantao Li, Patrick Marks, Karine Viaud-Martinez, Sascha Meiers, Katherine M. Munson, Fabio C. P. Navarro, Bradley J. Nelson, Conor Nodzak, Amina Noor, Sofia Kyriazopoulou-Panagiotopoulou, Andy W. C. Pang, Yunjiang Qiu, Gabriel Rosanio, Mallory Ryan, Adrian Stütz, Diana C. J. Spierings, Alistair Ward, AnneMarie E. Welch, Ming Xiao, Wei Xu, Chengsheng Zhang, Qihui Zhu, Xiangqun Zheng-Bradley, Ernesto Lowy, Sergei Yakneen, Steven McCarroll, Goo Jun, Li Ding, Chong Lek Koh,

Bing Ren, Paul Flicek, Ken Chen, Mark B. Gerstein, Pui-Yan Kwok, Peter M. Lansdorp, Gabor T. Marth, Jonathan Sebat, Xinghua Shi, Ali Bashir, Kai Ye, Scott E. Devine, Michael E. Talkowski, Ryan E. Mills, Tobias Marschall, Jan O. Korbel, Evan E. Eichler, and Charles Lee. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun*, 10(1):1–16, April 2019. Publisher: Nature Publishing Group.

[19] Jian-Min Chen, Claude Férec, and David N. Cooper. Gene Conversion in Human Genetic Disease. *Genes (Basel)*, 1(3):550–563, December 2010.

[20] Haoyu Cheng, Mobin Asri, Julian Lucas, Sergey Koren, and Heng Li. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. *Nat Methods*, 21(6):967–970, June 2024. Publisher: Nature Publishing Group.

[21] Haoyu Cheng, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*, 18(2):170–175, February 2021. Publisher: Nature Publishing Group.

[22] Eui-Hwan Choi, Seobin Yoon, Young Eun Koh, Young-Jin Seo, and Keun Pil Kim. Maintenance of genome integrity and active homologous recombination in embryonic stem cells. *Exp Mol Med*, 52(8):1220–1229, August 2020. Publisher: Nature Publishing Group.

[23] Wouter Coppieters, Carole Charlier, Michel Georges, Chad Harland, and Erik Mullaart. *Rate of de novo mutation in dairy cattle and potential impact of reproductive technologies*. February 2018.

[24] Gianmarco Corneo, Enrico Ginelli, and Elio Polli. A satellite DNA isolated from human tissues. *Journal of Molecular Biology*, 23(3):619–622, February 1967.

[25] Gianmarco Corneo, Enrico Ginelli, and Elio Polli. Repeated sequences in human DNA. *Journal of Molecular Biology*, 48(2):319–327, March 1970.

[26] Christine Coulondre, Jeffrey H. Miller, Philip J. Farabaugh, and Walter Gilbert. Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, 274(5673):775–780, August 1978. Publisher: Nature Publishing Group.

[27] James F. Crow. The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet*, 1(1):40–47, October 2000. Publisher: Nature Publishing Group.

[28] Cerys E. Currie, Emma Ford, Lucy Benham Whyte, Deborah M. Taylor, Bettina P. Mihalas, Muriel Erent, Adele L. Marston, Geraldine M. Hartshorne, and Andrew D.

McAinsh. The first mitotic division of human embryos is highly error prone. *Nat Commun*, 13(1):6755, November 2022. Publisher: Nature Publishing Group.

[29] Nikhita Damaraju, Angela L Miller, and Danny E Miller. Long-Read DNA and RNA Sequencing to Streamline Clinical Genetic Testing and Reduce Barriers to Comprehensive Genetic Testing. *The Journal of Applied Laboratory Medicine*, 9(1):138–150, January 2024.

[30] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, August 2011.

[31] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, and Heng Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, February 2021.

[32] Jean Dausset, Howard Cann, Daniel Cohen, Mark Lathrop, Jean-Marc Lalouel, and Ray White. Centre d'Etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. *Genomics*, 6(3):575–577, March 1990.

[33] Geoffry N. De Iuliis, Laura K. Thomson, Lisa A. Mitchell, Jane M. Finnie, Adam J. Koppers, Andrew Hedges, Brett Nixon, and R. John Aitken. DNA Damage in Human Spermatozoa Is Highly Correlated with the Efficiency of Chromatin Remodeling and the Formation of 8-Hydroxy-2'-Deoxyguanosine, a Marker of Oxidative Stress1. *Biology of Reproduction*, 81(3):517–524, September 2009.

[34] Alwin Derijck, Godfried van der Heijden, Maud Giele, Marielle Philippens, and Peter de Boer. DNA double-strand break repair in parental chromatin of mouse zygotes, the first cell cycle as an origin of de novo mutation. *Human Molecular Genetics*, 17(13):1922–1937, July 2008.

[35] Egor Dolzhenko, Mark F. Bennett, Phillip A. Richmond, Brett Trost, Sai Chen, Joke J. F. A. van Vugt, Charlotte Nguyen, Giuseppe Narzisi, Vladimir G. Gainullin, Andrew M. Gross, Bryan R. Lajoie, Ryan J. Taft, Wyeth W. Wasserman, Stephen W. Scherer, Jan H. Veldink, David R. Bentley, Ryan K. C. Yuen, Melanie Bahlo, and Michael A. Eberle. ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol*, 21(1):1–14, December 2020. Number: 1 Publisher: BioMed Central.

[36] Yanmei Dou, Xiaoxu Yang, Ziyi Li, Sheng Wang, Zheng Zhang, Adam Yongxin Ye, Linlin Yan, Changhong Yang, Qixi Wu, Jiarui Li, Boxun Zhao, August Yue Huang, and Liping Wei. Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. *Human Mutation*, 38(8):1002–1013, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.23255.

[37] Joni B. Drost and William R. Lee. Biological basis of germline mutation: Comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. *Environmental and Molecular Mutagenesis*, 25(S2):48–64, 1995. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/em.2850250609.

[38] Alexei J. Drummond and Andrew Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 7(1):214, November 2007.

[39] Beth L. Dumont. Interlocus gene conversion explains at least 2.7 % of single nucleotide variants in human segmental duplications. *BMC Genomics*, 16(1):456, June 2015.

[40] Laurent Duret and Nicolas Galtier. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annual Review of Genomics and Human Genetics*, 10(Volume 10, 2009):285–311, September 2009. Publisher: Annual Reviews.

[41] Mark T. W. Ebbert, Tanner D. Jensen, Karen Jansen-West, Jonathon P. Sens, Joseph S. Reddy, Perry G. Ridge, John S. K. Kauwe, Veronique Belzil, Luc Pregent, Minerva M. Carrasquillo, Dirk Keene, Eric Larson, Paul Crane, Yan W. Asmann, Nilufer Ertekin-Taner, Steven G. Younkin, Owen A. Ross, Rosa Rademakers, Leonard Petrucelli, and John D. Fryer. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biology*, 20(1):97, May 2019.

[42] Michael A. Eberle, Epameinondas Fritzilas, Peter Krusche, Morten Källberg, Benjamin L. Moore, Mitchell A. Bekritsky, Zamin Iqbal, Han-Yu Chuang, Sean J. Humphray, Aaron L. Halpern, Semyon Kruglyak, Elliott H. Margulies, Gil McVean, and David R. Bentley. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.*, 27(1):157–164, January 2017. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[43] Peter Ebert, Peter A. Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari,

Feyza Yilmaz, Xuefang Zhao, PingHsun Hsieh, Joyce Lee, Sushant Kumar, Jiadong Lin, Tobias Rausch, Yu Chen, Jingwen Ren, Martin Santamarina, Wolfram Höps, Hufsah Ashraf, Nelson T. Chuang, Xiaofei Yang, Katherine M. Munson, Alexandra P. Lewis, Susan Fairley, Luke J. Tallon, Wayne E. Clarke, Anna O. Basile, Marta Byrska-Bishop, André Corvelo, Uday S. Evani, Tsung-Yu Lu, Mark J. P. Chaisson, Junjie Chen, Chong Li, Harrison Brand, Aaron M. Wenger, Maryam Ghareghani, William T. Harvey, Benjamin Raeder, Patrick Hasenfeld, Allison A. Regier, Haley J. Abel, Ira M. Hall, Paul Flicek, Oliver Stegle, Mark B. Gerstein, Jose M. C. Tubio, Zepeng Mu, Yang I. Li, Xinghua Shi, Alex R. Hastie, Kai Ye, Zechen Chong, Ashley D. Sanders, Michael C. Zody, Michael E. Talkowski, Ryan E. Mills, Scott E. Devine, Charles Lee, Jan O. Korbel, Tobias Marschall, and Evan E. Eichler. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537):eabf7117, April 2021. Publisher: American Association for the Advancement of Science.

[44] Jana Ebler, Peter Ebert, Wayne E. Clarke, Tobias Rausch, Peter A. Audano, Torsten Houwaart, Yafei Mao, Jan O. Korbel, Evan E. Eichler, Michael C. Zody, Alexander T. Dilthey, and Tobias Marschall. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet*, 54(4):518–525, April 2022. Publisher: Nature Publishing Group.

[45] Harrison Echols and Myron F. Goodman. FIDELITY MECHANISMS IN DNA REPLICATION. *Annual Review of Biochemistry*, 60(Volume 60, 1991):477–511, July 1991. Publisher: Annual Reviews.

[46] Ester Falconer, Mark Hills, Ulrike Naumann, Steven S. S. Poon, Elizabeth A. Chavez, Ashley D. Sanders, Yongjun Zhao, Martin Hirst, and Peter M. Lansdorp. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat Methods*, 9(11):1107–1112, November 2012. Publisher: Nature Publishing Group.

[47] Gerald D. Fischbach and Catherine Lord. The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. *Neuron*, 68(2):192–195, October 2010.

[48] Walter M. Fitch. Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *Journal of Molecular Biology*, 26(3):499–507, June 1967.

[49] Susan Folstein and Michael Rutter. Genetic influences and infantile autism. *Nature*, 265(5596):726–728, February 1977. Publisher: Nature Publishing Group.

[50] Laurent C. Francioli, Paz P. Polak, Amnon Koren, Androniki Menelaou, Sung Chun, Ivo Renkens, Cornelia M. van Duijn, Morris Swertz, Cisca Wijmenga, Gertjan van Ommen, P. Eline Slagboom, Dorret I. Boomsma, Kai Ye, Victor Guryev, Peter F. Arndt, Wigard P. Kloosterman, Paul I. W. de Bakker, and Shamil R. Sunyaev. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet*, 47(7):822–826, July 2015. Publisher: Nature Publishing Group.

[51] Jack M. Fu, F. Kyle Satterstrom, Minshi Peng, Harrison Brand, Ryan L. Collins, Shan Dong, Brie Wamsley, Lambertus Klei, Lily Wang, Stephanie P. Hao, Christine R. Stevens, Caroline Cusick, Mehrtash Babadi, Eric Banks, Brett Collins, Sheila Dodge, Stacey B. Gabriel, Laura Gauthier, Samuel K. Lee, Lindsay Liang, Alicia Ljungdahl, Behrang Mahjani, Laura Sloofman, Andrey N. Smirnov, Mafalda Barbosa, Catalina Betancur, Alfredo Brusco, Brian H. Y. Chung, Edwin H. Cook, Michael L. Cuccaro, Enrico Domenici, Giovanni Battista Ferrero, J. Jay Gargus, Gail E. Herman, Irva Hertz-Picciotto, Patricia Maciel, Dara S. Manoach, Maria Rita Passos-Bueno, Antonio M. Persico, Alessandra Renieri, James S. Sutcliffe, Flora Tassone, Elisabetta Trabetti, Gabriele Campos, Simona Cardaropoli, Diana Carli, Marcus C. Y. Chan, Chiara Fallerini, Elisa Giorgio, Ana Cristina Girardi, Emily Hansen-Kiss, So Lun Lee, Carla Lintas, Yunin Ludena, Rachel Nguyen, Lisa Pavinato, Margaret Pericak-Vance, Isaac N. Pessah, Rebecca J. Schmidt, Moyra Smith, Claudia I. S. Costa, Slavica Trajkova, Jaqueline Y. T. Wang, Mullin H. C. Yu, David J. Cutler, Silvia De Rubeis, Joseph D. Buxbaum, Mark J. Daly, Bernie Devlin, Kathryn Roeder, Stephan J. Sanders, and Michael E. Talkowski. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat Genet*, 54(9):1320–1331, September 2022. Publisher: Nature Publishing Group.

[52] Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M. Slepchenko, Aleksei A. Bondarev, Philip L. F. Johnson, Ayinuer Aximu-Petri, Kay Prüfer, Cesare de Filippo, Matthias Meyer, Nicolas Zwyns, Domingo C. Salazar-García, Yaroslav V. Kuzmin, Susan G. Keates, Pavel A. Kosintsev, Dmitry I. Razhev, Michael P. Richards, Nikolai V. Peristov, Michael Lachmann, Katerina Douka, Thomas F. G. Higham, Montgomery Slatkin, Jean-Jacques Hublin, David Reich, Janet Kelso, T. Bence Viola, and Svante Pääbo. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 514(7523):445–449, October 2014. Publisher: Nature Publishing Group.

[53] Ziyue Gao, Priya Moorjani, Thomas A. Sasani, Brent S. Pedersen, Aaron R. Quinlan, Lynn B. Jorde, Guy Amster, and Molly Przeworski. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proceedings of the National Academy of Sciences*, 116(19):9491–9500, May 2019. Publisher: Proceedings of the National Academy of Sciences.

[54] Kiran V. Garimella, Zamin Iqbal, Michael A. Krause, Susana Campino, Mihir Kekre,

Eleanor Drury, Dominic Kwiatkowski, Juliana M. Sá, Thomas E. Wellems, and Gil McVean. Detection of simple and complex de novo mutations with multiple reference sequences. *Genome Res.*, 30(8):1154–1169, August 2020. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[55] Erik Garrison, Andrea Guarracino, Simon Heumos, Flavia Villani, Zhigui Bao, Lorenzo Tattini, Jörg Hagmann, Sebastian Vorbrugg, Santiago Marco-Sola, Christian Kubica, David G. Ashbrook, Kaisa Thorell, Rachel L. Rusholme-Pilcher, Gianni Liti, Emilio Rudbeck, Sven Nahnsen, Zuyu Yang, Mwaniki N. Moses, Franklin L. Nobrega, Yi Wu, Hao Chen, Joep de Ligt, Peter H. Sudmant, Nicole Soranzo, Vincenza Colonna, Robert W. Williams, and Pjotr Prins. Building pangenome graphs, April 2023. Pages: 2023.04.05.535718 Section: New Results.

[56] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv*, July 2012. arXiv:1207.3907 [q-bio].

[57] Erik Garrison, Jouni Sirén, Adam M. Novak, Glenn Hickey, Jordan M. Eizenga, Eric T. Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F. Lin, Benedict Paten, and Richard Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol*, 36(9):875–879, October 2018. Publisher: Nature Publishing Group.

[58] Ann Genovese and Merlin G. Butler. Clinical Assessment, Genetics, and Treatment Approaches in Autism Spectrum Disorder (ASD). *International Journal of Molecular Sciences*, 21(13):4726, January 2020. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

[59] Richard A. Gibbs, John W. Belmont, Paul Hardenbol, Thomas D. Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, Paul Kwong-Hang Tam, Lap-Chee Tsui, Mary Miu Yee Waye, Jeffrey Tze-Fei Wong, Changqing Zeng, Qingrun Zhang, Mark S. Chee, Luana M. Galver, Semyon Kruglyak, Sarah S. Murray, Arnold R. Oliphant, Alexandre Montpetit, Thomas J. Hudson, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Michael S. Phillips, Andrei Verner, Pui-Yan Kwok, Shenghui Duan, Denise L. Lind, Raymond D. Miller, John P. Rice, Nancy L. Saccone, Patricia Taillon-Miller, Ming Xiao, Yusuke Nakamura, Akihiro Sekine, Koki Sorimachi, Toshihiro Tanaka, Yoichi Tanaka, Tatsuhiko Tsunoda, Eiji Yoshino, David R. Bentley, Panos Deloukas, Sarah Hunt, Don Powell, David Altshuler, Stacey B. Gabriel, Houcan Zhang, Changqing Zeng, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R. Macer, Eiko Suda, Charles N. Rotimi, Clement A. Adebamowo, Toyin Aniagwu, Patricia A. Marshall, Olayemi Matthew, Chibuzor Nkwodimmah, Charmaine D. M.

Royal, Mark F. Leppert, Missy Dixon, Lincoln D. Stein, Fiona Cunningham, Ardavan Kanani, Gudmundur A. Thorisson, Aravinda Chakravarti, Peter E. Chen, David J. Cutler, Carl S. Kashuk, Peter Donnelly, Jonathan Marchini, Gilean A. T. McVean, Simon R. Myers, Lon R. Cardon, Gonçalo R. Abecasis, Andrew Morris, Bruce S. Weir, James C. Mullikin, Stephen T. Sherry, Michael Feolo, David Altshuler, Mark J. Daly, Stephen F. Schaffner, Renzong Qiu, Alastair Kent, Georgia M. Dunston, Kazuto Kato, Norio Niikawa, Bartha M. Knoppers, Morris W. Foster, Ellen Wright Clayton, Vivian Ota Wang, Jessica Watkin, Richard A. Gibbs, John W. Belmont, Erica Sodergren, George M. Weinstock, Richard K. Wilson, Lucinda L. Fulton, Jane Rogers, Bruce W. Birren, Hua Han, Hongguang Wang, Martin Godbout, John C. Wallenburg, Paul L'Archevêque, Guy Bellemare, Kazuo Todani, Takashi Fujita, Satoshi Tanaka, Arthur L. Holden, Eric H. Lai, Francis S. Collins, Lisa D. Brooks, Jean E. McEwen, Mark S. Guyer, Elke Jordan, Jane L. Peterson, Jack Spiegel, Lawrence M. Sung, Lynn F. Zacharia, Karen Kennedy, Michael G. Dunn, Richard Seabrook, Mark Shillito, Barbara Skene, John G. Stewart, David L. Valle (chair), Ellen Wright Clayton (co chair), Lynn B. Jorde (co chair), John W. Belmont, Aravinda Chakravarti, Mildred K. Cho, Troy Duster, Morris W. Foster, Marla Jasperse, Bartha M. Knoppers, Pui-Yan Kwok, Julio Licinio, Jeffrey C. Long, Patricia A. Marshall, Pilar N. Ossorio, Vivian Ota Wang, Charles N. Rotimi, Charmaine D. M. Royal, Patricia Spallone, Sharon F. Terry, Eric S. Lander (chair), Eric H. Lai (co chair), Deborah A. Nickerson (co chair), Gonçalo R. Abecasis, David Altshuler, David R. Bentley, Michael Boehnke, Lon R. Cardon, Mark J. Daly, Panos Deloukas, Julie A. Douglas, Stacey B. Gabriel, Richard R. Hudson, Thomas J. Hudson, Leonid Kruglyak, Pui-Yan Kwok, Yusuke Nakamura, Robert L. Nussbaum, Charmaine D. M. Royal, Stephen F. Schaffner, Stephen T. Sherry, Lincoln D. Stein, Toshihiro Tanaka, †The International HapMap Consortium, Genotyping centres: Baylor College of Medicine and ParAllele BioScience, Chinese HapMap Consortium, Illumina, McGill University and Génome Québec Innovation Centre, University of California at San Francisco and Washington University, University of Tokyo and RIKEN, Wellcome Trust Sanger Institute, Whitehead Institute/MIT Center for Genome Research, Community engagement/public consultation and sample-collection groups: Beijing Normal University and Beijing Genomics Institute, Eubios Ethics Institute and Shinshu University Health Sciences University of Hokkaido, Howard University and University of Ibadan, University of Utah, Analysis Groups: Cold Spring Harbor Laboratory, Johns Hopkins University School of Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics University of Oxford, US National Institutes of Health, Legal and Social Issues: Chinese Academy of Social Sciences Ethical, Genetic Interest Group, Howard University, Kyoto University, Nagasaki University, University of Montréal, University of Oklahoma, Vanderbilt University, Wellcome Trust, SNP Discovery: Baylor College of Medicine, Washington University, Scientific Management: Chinese Academy of Sciences, Chinese Ministry of Science and Technology, Genome Canada, Génome Québec, Science and

Technology Culture Japanese Ministry of Education, Sports, The SNP Consortium, Legal and Social Issues Group Initial Planning Groups: Populations and Ethical, and Methods Group. The International HapMap Project. *Nature*, 426(6968):789–796, December 2003. Publisher: Nature Publishing Group.

[60] Michael E Goldberg, Michelle D Noyes, Evan E Eichler, Aaron R Quinlan, and Kelley Harris. Effects of parental age and polymer composition on short tandem repeat de novo mutation rates. *Genetics*, 226(4):iyae013, April 2024.

[61] Jakob M. Goldmann, Wendy S. W. Wong, Michele Pinelli, Terry Farrah, Dale Bodian, Anna B. Stittrich, Gustavo Glusman, Lisenka E. L. M. Vissers, Alexander Hoischen, Jared C. Roach, Joseph G. Vockley, Joris A. Veltman, Benjamin D. Solomon, Christian Gilissen, and John E. Niederhuber. Parent-of-origin-specific signatures of de novo mutations. *Nat Genet*, 48(8):935–939, August 2016. Publisher: Nature Publishing Group.

[62] Andrea Guarracino, Silvia Buonaiuto, Leonardo Gomes de Lima, Tamara Potapova, Arang Rhie, Sergey Koren, Boris Rubinstein, Christian Fischer, Jennifer L. Gerton, Adam M. Phillippy, Vincenza Colonna, and Erik Garrison. Recombination between heterologous human acrocentric chromosomes. *Nature*, 617(7960):335–343, May 2023. Publisher: Nature Publishing Group.

[63] Faraz Hach, Fereydoun Hormozdiari, Can Alkan, Farhad Hormozdiari, Inanc Birol, Evan E. Eichler, and S. Cenk Sahinalp. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods*, 7(8):576–577, August 2010. Publisher: Nature Publishing Group.

[64] J. B. S. Haldane. The rate of spontaneous mutation of a human gene. *Journ. of Genetics*, 31(3):317–326, October 1935.

[65] Pille Hallast, Anastasia Agdzhoyan, Oleg Balanovsky, Yali Xue, and Chris Tyler-Smith. A Southeast Asian origin for present-day non-African human Y chromosomes. *Hum Genet*, 140(2):299–307, February 2021.

[66] Pille Hallast, Peter Ebert, Mark Loftus, Feyza Yilmaz, Peter A. Audano, Glennis A. Logsdon, Marc Jan Bonder, Weichen Zhou, Wolfram Höps, Kwondo Kim, Chong Li, Savannah J. Hoyt, Philip C. Dishuck, David Porubsky, Fotios Tsetsos, Jee Young Kwon, Qihui Zhu, Katherine M. Munson, Patrick Hasenfeld, William T. Harvey, Alexandra P. Lewis, Jennifer Kordosky, Kendra Hoekzema, Rachel J. O'Neill, Jan O. Korbel, Chris Tyler-Smith, Evan E. Eichler, Xinghua Shi, Christine R. Beck, Tobias Marschall, Miriam K. Konkel, and Charles Lee. Assembly of 43 human Y chromosomes reveals

extensive complexity and variation. *Nature*, 621(7978):355–364, September 2023. Publisher: Nature Publishing Group.

[67] Bjarni V. Halldorsson, Gunnar Palsson, Olafur A. Stefansson, Hakon Jonsson, Marteinn T. Hardarson, Hannes P. Eggertsson, Bjarni Gunnarsson, Asmundur Oddsson, Gisli H. Halldorsson, Florian Zink, Sigurjon A. Gudjonsson, Michael L. Frigge, Gudmar Thorleifsson, Asgeir Sigurdsson, Simon N. Stacey, Patrick Sulem, Gisli Masson, Agnar Helgason, Daniel F. Gudbjartsson, Unnur Thorsteinsdottir, and Kari Stefansson. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, 363(6425):eaau1043, January 2019. Publisher: American Association for the Advancement of Science.

[68] Agnar Helgason, Axel W. Einarsson, Valdís B. Guðmundsdóttir, Ásgeir Sigurðsson, Ellen D. Gunnarsdóttir, Anuradha Jagadeesan, S. Sunna Ebenesersdóttir, Augustine Kong, and Kári Stefánsson. The Y-chromosome point mutation rate in humans. *Nat Genet*, 47(5):453–457, May 2015. Publisher: Nature Publishing Group.

[69] Mir Henglin, Maryam Ghareghani, William T. Harvey, David Porubsky, Sergey Koren, Evan E. Eichler, Peter Ebert, and Tobias Marschall. Graphasing: phasing diploid genome assembly graphs with single-cell strand sequencing. *Genome Biology*, 25(1):265, October 2024.

[70] Kristien Hens, Hilde Peeters, and Kris Dierickx. Genetic testing and counseling in the case of an autism diagnosis: A caregivers perspective. *European Journal of Medical Genetics*, 59(9):452–458, September 2016.

[71] Susan M. Hiatt, James M. J. Lawlor, Lori H. Handley, Ryne C. Ramaker, Brianne B. Rogers, E. Christopher Partridge, Lori Beth Boston, Melissa Williams, Christopher B. Plott, Jerry Jenkins, David E. Gray, James M. Holt, Kevin M. Bowling, E. Martina Bebin, Jane Grimwood, Jeremy Schmutz, and Gregory M. Cooper. Long-read genome sequencing for the molecular diagnosis of neurodevelopmental disorders. *HGGADVANCE*, 2(2), April 2021. Publisher: Elsevier.

[72] Giles S. Holt, Lois E. Batty, Bilal K. S. Alobaidi, Hannah E. Smith, Manon S. Oud, Liliana Ramos, Miguel J. Xavier, and Joris A. Veltman. Phasing of de novo mutations using a scaled-up multiple amplicon long-read sequencing approach. *Hum Mutat*, 43(11):1545–1556, November 2022.

[73] F Horta, S Catt, P Ramachandran, B Vollenhoven, and P Temple-Smith. Female ageing affects the DNA repair capacity of oocytes in IVF using a controlled model of sperm DNA damage in mice. *Human Reproduction*, 35(3):529–544, March 2020.

[74] Austin L. Hughes. Near-Neutrality: the Leading Edge of the Neutral Theory of Molecular Evolution. *Ann N Y Acad Sci*, 1133:162–179, 2008.

[75] Dick G. Hwang and Phil Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences*, 101(39):13994–14001, September 2004. Publisher: Proceedings of the National Academy of Sciences.

[76] Ivan Iossifov, Brian J. O'Roak, Stephan J. Sanders, Michael Ronemus, Niklas Krumm, Dan Levy, Holly A. Stessman, Kali T. Witherspoon, Laura Vives, Karynne E. Patterson, Joshua D. Smith, Bryan Paeper, Deborah A. Nickerson, Jeanselle Dea, Shan Dong, Luis E. Gonzalez, Jeffrey D. Mandell, Shrikant M. Mane, Michael T. Murtha, Catherine A. Sullivan, Michael F. Walker, Zainulabedin Waqar, Liping Wei, A. Jeremy Willsey, Boris Yamrom, Yoon-ha Lee, Ewa Grabowska, Ertugrul Dalkic, Zihua Wang, Steven Marks, Peter Andrews, Anthony Leotta, Jude Kendall, Inessa Hakker, Julie Rosenbaum, Beicong Ma, Linda Rodgers, Jennifer Troge, Giuseppe Narzisi, Seungtai Yoon, Michael C. Schatz, Kenny Ye, W. Richard McCombie, Jay Shendure, Evan E. Eichler, Matthew W. State, and Michael Wigler. The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, 515(7526):216–221, November 2014. Publisher: Nature Publishing Group.

[77] P A Jacobs, T J Hassold, E Whittington, G Butler, S Collyer, M Keston, and M Lee. Klinefelter's syndrome: an analysis of the origin of the additional sex chromosome using molecular probes. *Ann Hum Genet*, 52(2):93–109, May 1988.

[78] Chirag Jain, Arang Rhie, Nancy F. Hansen, Sergey Koren, and Adam M. Phillippy. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods*, 19(6):705–710, June 2022. Publisher: Nature Publishing Group.

[79] Souraya Jaroudi, Georgia Kakourou, Suzanne Cawood, Alpesh Doshi, Domenico M. Ranieri, Paul Serhal, Joyce C. Harper, and Sioban B. SenGupta. Expression profiling of DNA repair genes in human oocytes and blastocysts using microarrays. *Human Reproduction*, 24(10):2649–2655, October 2009.

[80] Young Seok Ju, Inigo Martincorena, Moritz Gerstung, Mia Petljak, Ludmil B. Alexandrov, Raheleh Rahbari, David C. Wedge, Helen R. Davies, Manasa Ramakrishna, Anthony Fullam, Sancha Martin, Christopher Alder, Nikita Patel, Steve Gamble, Sarah O'Meara, Dilip D. Giri, Torril Sauer, Sarah E. Pinder, Colin A. Purdie, Åke Borg, Henk Stunnenberg, Marc van de Vijver, Benita K. T. Tan, Carlos Caldas, Andrew Tutt, Naoto T. Ueno, Laura J. van 't Veer, John W. M. Martens, Christos Sotiriou, Stian Knappskog, Paul N. Span, Sunil R. Lakhani, Jórunn Erla Eyfjörd, Anne-Lise Børresen-Dale, Andrea Richardson, Alastair M. Thompson, Alain Viari, Matthew E.

Hurles, Serena Nik-Zainal, Peter J. Campbell, and Michael R. Stratton. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*, 543(7647):714–718, March 2017. Publisher: Nature Publishing Group.

[81] Hákon Jónsson, Patrick Sulem, Birte Kehr, Snaedis Kristmundsdottir, Florian Zink, Eirikur Hjartarson, Marteinn T. Hardarson, Kristjan E. Hjorleifsson, Hannes P. Eggertsson, Sigurjon Axel Gudjonsson, Lucas D. Ward, Gudny A. Arnadottir, Einar A. Helgason, Hannes Helgason, Arnaldur Gylfason, Adalbjorg Jonasdottir, Aslaug Jonasdottir, Thorunn Rafnar, Mike Frigge, Simon N. Stacey, Olafur Th. Magnusson, Unnur Thorsteinsdottir, Gisli Masson, Augustine Kong, Bjarni V. Halldorsson, Agnar Helgason, Daniel F. Gudbjartsson, and Kari Stefansson. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, 549(7673):519–522, September 2017. Publisher: Nature Publishing Group.

[82] D K Kalousek, I J Barrett, and B C McGillivray. Placental mosaicism and intrauterine survival of trisomies 13 and 18. *Am J Hum Genet*, 44(3):338–343, March 1989.

[83] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, July 2002.

[84] Kazutaka Katoh and Daron M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780, April 2013.

[85] Michael D. Kessler, Douglas P. Loesch, James A. Perry, Nancy L. Heard-Costa, Daniel Taliun, Brian E. Cade, Heming Wang, Michelle Daya, John Ziniti, Soma Datta, Juan C. Celedón, Manuel E. Soto-Quiros, Lydiana Avila, Scott T. Weiss, Kathleen Barnes, Susan S. Redline, Ramachandran S. Vasan, Andrew D. Johnson, Rasika A. Mathias, Ryan Hernandez, James G. Wilson, Deborah A. Nickerson, Goncalo Abecasis, Sharon R. Browning, Sebastian Zöllner, Jeffrey R. O'Connell, Braxton D. Mitchell, National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Population Genetics Working Group, and Timothy D. O'Connor. De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proceedings of the National Academy of Sciences*, 117(5):2560–2569, February 2020. Publisher: Proceedings of the National Academy of Sciences.

[86] Evgenia V. Khokhlova, Zoia S. Fesenko, Julia V. Sopova, and Elena I. Leonova. Features of DNA Repair in the Early Stages of Mammalian Embryonic Development. *Genes (Basel)*, 11(10):1138, September 2020.

[87] Il Bin Kim, Taeyeop Lee, Junehawk Lee, Jonghun Kim, Suho Lee, In Gyeong Koh, Jae Hyun Kim, Joon-Yong An, Hyunseong Lee, Woo Kyeong Kim, Young Seok Ju, Yongseong Cho, Seok Jong Yu, Soon Ae Kim, Miae Oh, Dong Wook Han, Eunjoon Kim, Jung Kyoon Choi, Hee Jeong Yoo, and Jeong Ho Lee. Non-coding de novo mutations in chromatin interactions are implicated in autism spectrum disorder. *Mol Psychiatry*, 27(11):4680–4694, November 2022. Publisher: Nature Publishing Group.

[88] Sangtae Kim, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, Yeonbin Kim, Doruk Beyter, Peter Krusche, and Christopher T. Saunders. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*, 15(8):591–594, August 2018. Publisher: Nature Publishing Group.

[89] Motoo Kimura and Tomoko Ohta. On the rate of molecular evolution. *J Mol Evol*, 1(1):1–17, March 1971.

[90] Alexey S. Kondrashov. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Human Mutation*, 21(1):12–27, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.10147.

[91] Augustine Kong, Michael L. Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A. Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Wendy S. W. Wong, Gunnar Sigurdsson, G. Bragi Walters, Stacy Steinberg, Hannes Helgason, Gudmar Thorleifsson, Daniel F. Gudbjartsson, Agnar Helgason, Olafur Th Magnusson, Unnur Thorsteinsdottir, and Kari Stefansson. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412):471–475, August 2012. Publisher: Nature Publishing Group.

[92] Niklas Krumm, Tychele N. Turner, Carl Baker, Laura Vives, Kiana Mohajeri, Kali Witherspoon, Archana Raja, Bradley P. Coe, Holly A. Stessman, Zong-Xiao He, Suzanne M. Leal, Raphael Bernier, and Evan E. Eichler. Excess of rare, inherited truncating mutations in autism. *Nat Genet*, 47(6):582–588, June 2015. Publisher: Nature Publishing Group.

[93] Jan Krumsiek, Roland Arnold, and Thomas Rattei. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, 23(8):1026–1028, April 2007.

[94] Erdi Kucuk, Bart P. G. H. van der Sanden, Luke O'Gorman, Michael Kwint, Ronny Derks, Aaron M. Wenger, Christine Lambert, Shreyasee Chakraborty, Primo Baybayan, William J. Rowell, Han G. Brunner, Lisenka E. L. M. Vissers, Alexander Hoischen, and Christian Gilissen. Comprehensive de novo mutation discovery with HiFi long-read sequencing. *Genome Medicine*, 15(1):34, May 2023.

[95] Yoko Kuroki, Atsushi Toyoda, Hideki Noguchi, Todd D. Taylor, Takehiko Itoh, Dae-Soo Kim, Dae-Won Kim, Sang-Haeng Choi, Il-Chul Kim, Han Ho Choi, Yong Sung Kim, Yoko Satta, Naruya Saitou, Tomoyuki Yamada, Shinichi Morishita, Masahira Hattori, Yoshiyuki Sakaki, Hong-Seog Park, and Asao Fujiyama. Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat Genet*, 38(2):158–167, February 2006. Publisher: Nature Publishing Group.

[96] Hyeonjin Lee, Eun Na Kim, Ji-Young Lee, Ji Hun Kim, Ji-Hye Oh, Won-Kyung Kim, Eun Jeong Cho, Jinyeong Lim, Sung-Min Chun, and Chang Ohk Sung. Characterization of early postzygotic somatic mutations through multi-organ analysis. *J Hum Genet*, 66(8):777–784, August 2021. Publisher: Nature Publishing Group.

[97] Dan Levy, Michael Ronemus, Boris Yamrom, Yoon-ha Lee, Anthony Leotta, Jude Kendall, Steven Marks, B. Lakshmi, Deepa Pai, Kenny Ye, Andreas Buja, Abba Krieger, Seungtai Yoon, Jennifer Troge, Linda Rodgers, Ivan Iossifov, and Michael Wigler. Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders. *Neuron*, 70(5):886–897, June 2011.

[98] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, November 2011.

[99] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, May 2013. arXiv:1303.3997 [q-bio].

[100] Heng Li. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–2851, October 2014.

[101] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018.

[102] Heng Li, Jonathan M. Bloom, Yossi Farjoun, Mark Fleharty, Laura Gauthier, Benjamin Neale, and Daniel MacArthur. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods*, 15(8):595–597, August 2018. Publisher: Nature Publishing Group.

[103] Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5):589–595, March 2010.

[104] Heng Li, Xiaowen Feng, and Chong Chu. The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, 21(1):265, October 2020.

[105] Wen-Hsiung Li, Chung-I. Wu, and Chi-Cheng Luo. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J Mol Evol*, 21(1):58–71, November 1984. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Springer-Verlag.

[106] Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T. Harvey, Simon Heumos, Kerstin Howe, Miten Jain, Tsung-Yu Lu, Charles Markello, Fergal J. Martin, Matthew W. Mitchell, Katherine M. Munson, Moses Njagi Mwaniki, Adam M. Novak, Hugh E. Olsen, Trevor Pesout, David Porubsky, Pjotr Prins, Jonas A. Sibbesen, Jouni Sirén, Chad Tomlinson, Flavia Villani, Mitchell R. Vollger, Lucinda L. Antonacci-Fulton, Gunjan Baid, Carl A. Baker, Anastasiya Belyaeva, Konstantinos Billis, Andrew Carroll, Pi-Chuan Chang, Sarah Cody, Daniel E. Cook, Robert M. Cook-Deegan, Omar E. Cornejo, Mark Diekhans, Peter Ebert, Susan Fairley, Olivier Fedrigo, Adam L. Felsenfeld, Giulio Formenti, Adam Frankish, Yan Gao, Nanibaa' A. Garrison, Carlos Garcia Giron, Richard E. Green, Leanne Haggerty, Kendra Hoekzema, Thibaut Hourlier, Hanlee P. Ji, Eimear E. Kenny, Barbara A. Koenig, Alexey Kolesnikov, Jan O. Korbel, Jennifer Kordosky, Sergey Koren, HoJoon Lee, Alexandra P. Lewis, Hugo Magalhães, Santiago Marco-Sola, Pierre Marijon, Ann McCartney, Jennifer McDaniel, Jacquelyn Mountcastle, Maria Nattestad, Sergey Nurk, Nathan D. Olson, Alice B. Popejoy, Daniela Puiu, Mikko Rautiainen, Allison A. Regier, Arang Rhie, Samuel Sacco, Ashley D. Sanders, Valerie A. Schneider, Baergen I. Schultz, Kishwar Shafin, Michael W. Smith, Heidi J. Sofia, Ahmad N. Abou Tayoun, Françoise Thibaud-Nissen, Francesca Floriana Tricomi, Justin Wagner, Brian Walenz, Jonathan M. D. Wood, Aleksey V. Zimin, Guillaume Bourque, Mark J. P. Chaisson, Paul Flicek, Adam M. Phillippy, Justin M. Zook, Evan E. Eichler, David Haussler, Ting Wang, Erich D. Jarvis, Karen H. Miga, Erik Garrison, Tobias Marschall, Ira M. Hall, Heng Li, and Benedict Paten. A draft human pangenome reference. *Nature*, 617(7960):312–324, May 2023. Publisher: Nature Publishing Group.

[107] Elaine T. Lim, Mohammed Uddin, Silvia De Rubeis, Yingleong Chan, Anne S. Kamumbu, Xiaochang Zhang, Alissa M. D'Gama, Sonia N. Kim, Robert Sean Hill, Arthur P. Goldberg, Christopher Poultney, Nancy J. Minshew, Itaru Kushima, Branko Aleksic, Norio Ozaki, Mara Parellada, Celso Arango, Maria J. Penzol, Angel Carracedo, Alexander Kolevzon, Christina M. Hultman, Lauren A. Weiss, Menachem Fromer, Andreas G. Chiocchetti, Christine M. Freitag, George M. Church, Stephen W. Scherer, Joseph D. Buxbaum, and Christopher A. Walsh. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat Neurosci*, 20(9):1217–1224, September 2017. Publisher: Nature Publishing Group.

[108] Ge Liu, NISC Comparative Sequencing Program, Shaying Zhao, Jeffrey A. Bailey, S. Cenk Sahinalp, Can Alkan, Eray Tuzun, Eric D. Green, and Evan E. Eichler. Analysis of Primate Genomic Variation Reveals a Repeat-Driven Expansion of the Human Genome. *Genome Res.*, 13(3):358–368, March 2003. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[109] Glennis Logsdon. HMW gDNA purification and ONT ultra-long-read data generation. *protocols.io*, March 2022.

[110] Glennis A. Logsdon, Allison N. Rozanski, Fedor Ryabov, Tamara Potapova, Valery A. Shepelev, Claudia R. Catacchio, David Porubsky, Yafei Mao, DongAhn Yoo, Mikko Rautiainen, Sergey Koren, Sergey Nurk, Julian K. Lucas, Kendra Hoekzema, Katherine M. Munson, Jennifer L. Gerton, Adam M. Phillippy, Mario Ventura, Ivan A. Alexandrov, and Evan E. Eichler. The variation and evolution of complete human centromeres. *Nature*, 629(8010):136–145, May 2024. Publisher: Nature Publishing Group.

[111] Glennis A. Logsdon, Mitchell R. Vollger, PingHsun Hsieh, Yafei Mao, Mikhail A. Liskovykh, Sergey Koren, Sergey Nurk, Ludovica Mercuri, Philip C. Dishuck, Arang Rhie, Leonardo G. de Lima, Tatiana Dvorkina, David Porubsky, William T. Harvey, Alla Mikheenko, Andrey V. Bzikadze, Milinn Kremitzki, Tina A. Graves-Lindsay, Chirag Jain, Kendra Hoekzema, Shwetha C. Murali, Katherine M. Munson, Carl Baker, Melanie Sorensen, Alexandra M. Lewis, Urvashi Surti, Jennifer L. Gerton, Vladimir Larionov, Mario Ventura, Karen H. Miga, Adam M. Phillippy, and Evan E. Eichler. The structure, function and evolution of a complete human chromosome 8. *Nature*, 593(7857):101–107, May 2021. Publisher: Nature Publishing Group.

[112] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, November 2010.

[113] Tuomo Mantere, Kornelia Neveling, Céline Pebrel-Richard, Marion Benoist, Guillaume van der Zande, Ellen Kater-Baats, Imane Baatout, Ronald van Beek, Tony Yammine, Michiel Oorsprong, Faten Hsoumi, Daniel Olde-Weghuis, Wed Majdali, Susan Vermeulen, Marc Pauper, Aziza Lebbar, Marian Stevens-Kroef, Damien Sanlaville, Jean Michel Dupont, Dominique Smeets, Alexander Hoischen, Caroline Schluth-Bolard, and Laïla El Khattabi. Optical genome mapping enables constitutional chromosomal aberration detection. *The American Journal of Human Genetics*, 108(8):1409–1422, August 2021.

[114] Christian R. Marshall, Abdul Noor, John B. Vincent, Anath C. Lionel, Lars Feuk, Jennifer Skaug, Mary Shago, Rainald Moessner, Dalila Pinto, Yan Ren, Bhooma Thiruvahindrapduram, Andreas Fiebig, Stefan Schreiber, Jan Friedman, Cees E. J. Ketelaars, Yvonne J. Vos, Can Ficicioglu, Susan Kirkpatrick, Rob Nicolson, Leon Sloman, Anne Summers, Clare A. Gibbons, Ahmad Teebi, David Chitayat, Rosanna Weksberg, Ann Thompson, Cathy Vardy, Vicki Crosbie, Sandra Luscombe, Rebecca Baatjes, Lonnie Zwaigenbaum, Wendy Roberts, Bridget Fernandez, Peter Szatmari, and Stephen W. Scherer. Structural Variation of Chromosomes in Autism Spectrum Disorder. *The American Journal of Human Genetics*, 82(2):477–488, February 2008.

[115] Marcel Martin, Murray Patterson, Shilpa Garg, Sarah O. Fischer, Nadia Pisanti, Gunnar W. Klau, Alexander Schöenhuth, and Tobias Marschall. WhatsHap: fast and accurate read-based phasing. *bioRxiv*, November 2016. Pages: 085050 Section: New Results.

[116] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, March 2011.

[117] Jason D. Merker, Aaron M. Wenger, Tam Sneddon, Megan Grove, Zachary Zappala, Laure Fresard, Daryl Waggott, Sowmi Utiramerur, Yanli Hou, Kevin S. Smith, Stephen B. Montgomery, Matthew Wheeler, Jillian G. Buchan, Christine C. Lambert, Kevin S. Eng, Luke Hickey, Jonas Korlach, James Ford, and Euan A. Ashley. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in Medicine*, 20(1):159–163, January 2018.

[118] Jacob J. Michaelson, Yujian Shi, Madhusudan Gujral, Hancheng Zheng, Dheeraj Malhotra, Xin Jin, Minghan Jian, Guangming Liu, Douglas Greer, Abhishek Bhandari, Wenting Wu, Roser Corominas, Áine Peoples, Amnon Koren, Athurva Gore, Shuli Kang, Guan Ning Lin, Jasper Estabillo, Therese Gadomski, Balvindar Singh, Kun Zhang, Natacha Akshoomoff, Christina Corsello, Steven McCarroll, Lilia M. Iakoucheva, Yingrui Li, Jun Wang, and Jonathan Sebat. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. *Cell*, 151(7):1431–1442, December 2012.

[119] Karen H. Miga and Evan E. Eichler. Envisioning a new era: Complete genetic information from routine, telomere-to-telomere genomes. *The American Journal of Human Genetics*, 110(11):1832–1840, November 2023. Publisher: Elsevier.

[120] Danny E. Miller, Arvis Sulovari, Tianyun Wang, Hailey Loucks, Kendra Hoekzema, Katherine M. Munson, Alexandra P. Lewis, Edith P. Almanza Fuerte, Catherine R. Paschal, Tom Walsh, Jenny Thies, James T. Bennett, Ian Glass, Katrina M. Dipple, Karynne Patterson, Emily S. Bonkowski, Zoe Nelson, Audrey Squire, Megan Sikes,

Erika Beckman, Robin L. Bennett, Dawn Earl, Winston Lee, Rando Allikmets, Seth J. Perlman, Penny Chow, Anne V. Hing, Tara L. Wenger, Margaret P. Adam, Angela Sun, Christina Lam, Irene Chang, Xue Zou, Stephanie L. Austin, Erin Huggins, Alexias Safi, Apoorva K. Iyengar, Timothy E. Reddy, William H. Majoros, Andrew S. Allen, Gregory E. Crawford, Priya S. Kishnani, Mary-Claire King, Tim Cherry, Jessica X. Chong, Michael J. Bamshad, Deborah A. Nickerson, Heather C. Mefford, Dan Doherty, and Evan E. Eichler. Targeted long-read sequencing identifies missing disease-causing variation. *The American Journal of Human Genetics*, 108(8):1436–1449, August 2021. Publisher: Elsevier.

[121] Jaina Mistry, Robert D. Finn, Sean R. Eddy, Alex Bateman, and Marco Punta. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, 41(12):e121, July 2013.

[122] Ileena Mitra, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E. Lohmueller, and Melissa Gymrek. Patterns of de novo tandem repeat mutations and their role in autism. *Nature*, 589(7841):246–250, January 2021. Publisher: Nature Publishing Group.

[123] Nicholas M. Murphy, Tanya S. Samarasekera, Lisa Macaskill, Jayne Mullen, and Luk J. F. Rombauts. Genome sequencing of human in vitro fertilisation embryos for pathogenic variation screening. *Sci Rep*, 10(1):1–10, March 2020. Publisher: Nature Publishing Group.

[124] Richard Musson, Łukasz Gąsior, Simona Bisogno, and Grażyna Ewa Ptak. DNA damage in preimplantation embryos and gametes: specification, clinical relevance and repair strategies. *Hum Reprod Update*, 28(3):376–399, January 2022.

[125] Francesc Muyas, Luis Zapata, Roderic Guigó, and Stephan Ossowski. The rate and spectrum of mosaic mutations during embryogenesis revealed by RNA sequencing of 49 tissues. *Genome Med*, 12(1):1–14, December 2020. Number: 1 Publisher: BioMed Central.

[126] Michael W Nachman and Susan L Crowell. Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics*, 156(1):297–304, September 2000.

[127] Thomas Nagylaki and Thomas D Petes. INTRACHROMOSOMAL GENE CONVERSION AND THE MAINTENANCE OF SEQUENCE HOMOGENEITY AMONG REPEATED GENES. *Genetics*, 100(2):315–337, February 1982.

[128] Benjamin M. Neale, Yan Kou, Li Liu, Avi Ma'ayan, Kaitlin E. Samocha, Aniko Sabo, Chiao-Feng Lin, Christine Stevens, Li-San Wang, Vladimir Makarov, Paz Polak, Seungtai Yoon, Jared Maguire, Emily L. Crawford, Nicholas G. Campbell, Evan T. Geller, Otto Valladares, Chad Schafer, Han Liu, Tuo Zhao, Guiqing Cai, Jayon Lihm, Ruth Dannenfelser, Omar Jabado, Zuleyma Peralta, Uma Nagaswamy, Donna Muzny, Jeffrey G. Reid, Irene Newsham, Yuanqing Wu, Lora Lewis, Yi Han, Benjamin F. Voight, Elaine Lim, Elizabeth Rossin, Andrew Kirby, Jason Flannick, Menachem Fromer, Khalid Shakir, Tim Fennell, Kiran Garimella, Eric Banks, Ryan Poplin, Stacey Gabriel, Mark DePristo, Jack R. Wimbish, Braden E. Boone, Shawn E. Levy, Catalina Betancur, Shamil Sunyaev, Eric Boerwinkle, Joseph D. Buxbaum, Edwin H. Cook Jr, Bernie Devlin, Richard A. Gibbs, Kathryn Roeder, Gerard D. Schellenberg, James S. Sutcliffe, and Mark J. Daly. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–245, May 2012. Publisher: Nature Publishing Group.

[129] H Newman, S Catt, B Vining, B Vollenhoven, and F Horta. DNA repair and response to sperm DNA damage in oocytes and embryos, and the potential consequences in ART: a systematic review. *Molecular Human Reproduction*, 28(1):gaab071, January 2022.

[130] Jeffrey K. Ng and Tychele N. Turner. HAT: de novo variant calling for highly accurate short-read and long-read sequencing data, January 2023. Pages: 2023.01.27.525940 Section: New Results.

[131] Michelle D. Noyes, William T. Harvey, David Porubsky, Arvis Sulovari, Ruiyang Li, Nicholas R. Rose, Peter A. Audano, Katherine M. Munson, Alexandra P. Lewis, Kendra Hoekzema, Tuomo Mantere, Tina A. Graves-Lindsay, Ashley D. Sanders, Sara Goodwin, Melissa Kramer, Younes Mokrab, Michael C. Zody, Alexander Hoischen, Jan O. Korbel, W. Richard McCombie, and Evan E. Eichler. Familial long-read sequencing increases yield of de novo mutations. *The American Journal of Human Genetics*, 109(4):631–646, April 2022. Publisher: Elsevier.

[132] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Nae-Chyun Chen, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G. S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias

Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Ying Sims, Arian F. A. Smit, Daniela C. Soto, Ivan Sović, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, and Adam M. Phillippy. The complete sequence of a human genome. *Science*, 376(6588):44–53, April 2022. Publisher: American Association for the Advancement of Science.

[133] Tomoko Ohta. The neutral theory is dead. The current significance and standing of neutral and nearly neutral theories. *BioEssays*, 18(8):673–677, 1996. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.950180811.

[134] Brian J. O'Roak, Pelagia Deriziotis, Choli Lee, Laura Vives, Jerrod J. Schwartz, Santhosh Girirajan, Emre Karakoc, Alexandra P. MacKenzie, Sarah B. Ng, Carl Baker, Mark J. Rieder, Deborah A. Nickerson, Raphael Bernier, Simon E. Fisher, Jay Shendure, and Evan E. Eichler. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet*, 43(6):585–589, June 2011. Publisher: Nature Publishing Group.

[135] Seongyeol Park, Nanda Maya Mali, Ryul Kim, Jeong-Woo Choi, Junehawk Lee, Joonoh Lim, Jung Min Park, Jung Woo Park, Donghyun Kim, Taewoo Kim, Kijong Yi, June Hyug Choi, Seong Gyu Kwon, Joo Hee Hong, Jeonghwan Youk, Yohan An, Su Yeon Kim, Soo A. Oh, Youngoh Kwon, Dongwan Hong, Moonkyu Kim, Dong Sun Kim, Ji Young Park, Ji Won Oh, and Young Seok Ju. Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature*, 597(7876):393–397, September 2021.

[136] Marc Pauper, Erdi Kucuk, Aaron M. Wenger, Shreyasee Chakraborty, Primo Baybayan, Michael Kwint, Bart van der Sanden, Marcel R. Nelen, Ronny Derks, Han G. Brunner, Alexander Hoischen, Lisenka E. L. M. Vissers, and Christian Gilissen. Long-read trio sequencing of individuals with unsolved intellectual disability. *Eur J Hum Genet*, 29(4):637–648, April 2021.

[137] Brock A. Peters, Bahram G. Kermani, Oleg Alferov, Misha R. Agarwal, Mark A. McElwain, Natali Gulbahce, Daniel M. Hayden, Y. Tom Tang, Rebecca Yu Zhang, Rick Tearle, Birgit Crain, Renata Prates, Alan Berkeley, Santiago Munné, and Radoje

Drmanac. Detection and phasing of single base de novo mutations in biopsies from human in vitro fertilized embryos by advanced whole-genome sequencing. *Genome Res*, 25(3):426–434, March 2015.

[138] Susanne P. Pfeifer. Direct estimate of the spontaneous germ line mutation rate in african green monkeys. *Evolution*, 71(12):2858–2870, 12 2017.

[139] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T. Afshar, Sam S. Gross, Lizzie Dorfman, Cory Y. McLean, and Mark A. DePristo. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*, 36(10):983–987, November 2018. Publisher: Nature Publishing Group.

[140] Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J. Daly, Ben Neale, Daniel G. MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, July 2018. Pages: 201178 Section: New Results.

[141] David Porubsky, Harriet Dashnow, Thomas A. Sasani, Glennis A. Logsdon, Pille Hallast, Michelle D. Noyes, Zev N. Kronenberg, Tom Mokveld, Nidhi Koundinya, Cillian Nolan, Cody J. Steely, Andrea Guarracino, Egor Dolzhenko, William T. Harvey, William J. Rowell, Kirill Grigorev, Thomas J. Nicholas, Keisuke K. Oshima, Jiadong Lin, Peter Ebert, W. Scott Watkins, Tiffany Y. Leung, Vincent C. T. Hanlon, Sean McGee, Brent S. Pedersen, Michael E. Goldberg, Hannah C. Happ, Hyeonsoo Jeong, Katherine M. Munson, Kendra Hoekzema, Daniel D. Chan, Yanni Wang, Jordan Knuth, Gage H. Garcia, Cairbre Fanslow, Christine Lambert, Charles Lee, Joshua D. Smith, Shawn Levy, Christopher E. Mason, Erik Garrison, Peter M. Lansdorp, Deborah W. Neklason, Lynn B. Jorde, Aaron R. Quinlan, Michael A. Eberle, and Evan E. Eichler. A familial, telomere-to-telomere reference for human de novo mutation and recombination from a four-generation pedigree, August 2024. Pages: 2024.08.05.606142 Section: New Results.

[142] David Porubsky and Evan E. Eichler. A 25-year odyssey of genomic technology advances and structural variant discovery. *Cell*, 187(5):1024–1037, February 2024. Publisher: Elsevier.

[143] David Porubsky, Shilpa Garg, Ashley D. Sanders, Jan O. Korbel, Victor Guryev, Peter M. Lansdorp, and Tobias Marschall. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat Commun*, 8(1):1–10, November 2017. Publisher: Nature Publishing Group.

[144] David Porubsky, Xavi Guitart, DongAhn Yoo, Philip C. Dishuck, William T. Harvey, and Evan E. Eichler. SVbyEye: A visual tool to characterize structural variation among whole genome assemblies, September 2024. Pages: 2024.09.11.612418 Section: New Results.

[145] David Porubsky, Ashley D Sanders, Aaron Taudt, Maria Colomé-Tatché, Peter M Lansdorp, and Victor Guryev. breakpointR: an R/Bioconductor package to localize strand state changes in Strand-seq data. *Bioinformatics*, 36(4):1260–1261, February 2020.

[146] David Porubský, Ashley D. Sanders, Niek van Wietmarschen, Ester Falconer, Mark Hills, Diana C. J. Spierings, Marianna R. Bevova, Victor Guryev, and Peter M. Lansdorp. Direct chromosome-length haplotyping by single-cell sequencing. *Genome Res.*, 26(11):1565–1574, November 2016. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[147] Jennifer E. Posey, Jill A. Rosenfeld, Regis A. James, Matthew Bainbridge, Zhiyv Niu, Xia Wang, Shweta Dhar, Wojciech Wiszniewski, Zeynep H. C. Akdemir, Tomasz Gambin, Fan Xia, Richard E. Person, Magdalena Walkiewicz, Chad A. Shaw, V. Reid Sutton, Arthur L. Beaudet, Donna Muzny, Christine M. Eng, Yaping Yang, Richard A. Gibbs, James R. Lupski, Eric Boerwinkle, and Sharon E. Plon. Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genetics in Medicine*, 18(7):678–685, July 2016.

[148] G. David Poznik, Brenna M. Henn, Muh-Ching Yee, Elzbieta Sliwerska, Ghia M. Euskirchen, Alice A. Lin, Michael Snyder, Lluis Quintana-Murci, Jeffrey M. Kidd, Peter A. Underhill, and Carlos D. Bustamante. Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males Versus Females. *Science*, 341(6145):562–565, August 2013. Publisher: American Association for the Advancement of Science.

[149] The 1000 Genomes Project. Variation in genome-wide mutation rates within and between human families. *Nat Genet*, 43(7):712–714, July 2011. Publisher: Nature Publishing Group.

[150] Raheleh Rahbari, Arthur Wuster, Sarah J. Lindsay, Robert J. Hardwick, Ludmil B. Alexandrov, Saeed Al Turki, Anna Dominiczak, Andrew Morris, David Porteous, Blair Smith, Michael R. Stratton, and Matthew E. Hurles. Timing, rates and spectra of human germline mutation. *Nat Genet*, 48(2):126–133, February 2016. Publisher: Nature Publishing Group.

[151] Mikko Rautiainen, Sergey Nurk, Brian P. Walenz, Glennis A. Logsdon, David Porubsky, Arang Rhie, Evan E. Eichler, Adam M. Phillippy, and Sergey Koren. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol*, 41(10):1474–1482, October 2023. Publisher: Nature Publishing Group.

[152] Jennifer Reiner, Laura Pisani, Wanqiong Qiao, Ram Singh, Yao Yang, Lisong Shi, Wahab A. Khan, Robert Sebra, Ninette Cohen, Arvind Babu, Lisa Edelmann, Ethylin Wang Jabs, and Stuart A. Scott. Cytogenomic identification and long-read single molecule real-time (SMRT) sequencing of a Bardet–Biedl Syndrome 9 (BBS9) deletion. *npj Genomic Med*, 3(1):1–5, January 2018. Publisher: Nature Publishing Group.

[153] Jingwen Ren and Mark J. P. Chaisson. lra: A long read aligner for sequences and contigs. *PLOS Computational Biology*, 17(6):e1009078, June 2021. Publisher: Public Library of Science.

[154] Arang Rhie, Sergey Nurk, Monika Cechova, Savannah J. Hoyt, Dylan J. Taylor, Nicolas Altemose, Paul W. Hook, Sergey Koren, Mikko Rautiainen, Ivan A. Alexandrov, Jamie Allen, Mobin Asri, Andrey V. Bzikadze, Nae-Chyun Chen, Chen-Shan Chin, Mark Diekhans, Paul Flicek, Giulio Formenti, Arkarachai Fungtammasan, Carlos Garcia Giron, Erik Garrison, Ariel Gershman, Jennifer L. Gerton, Patrick G. S. Grady, Andrea Guarracino, Leanne Haggerty, Reza Halabian, Nancy F. Hansen, Robert Harris, Gabrielle A. Hartley, William T. Harvey, Marina Haukness, Jakob Heinz, Thibaut Hourlier, Robert M. Hubley, Sarah E. Hunt, Stephen Hwang, Miten Jain, Rupesh K. Kesharwani, Alexandra P. Lewis, Heng Li, Glennis A. Logsdon, Julian K. Lucas, Wojciech Makalowski, Christopher Markovic, Fergal J. Martin, Ann M. Mc Cartney, Rajiv C. McCoy, Jennifer McDaniel, Brandy M. McNulty, Paul Medvedev, Alla Mikheenko, Katherine M. Munson, Terence D. Murphy, Hugh E. Olsen, Nathan D. Olson, Luis F. Paulin, David Porubsky, Tamara Potapova, Fedor Ryabov, Steven L. Salzberg, Michael E. G. Sauria, Fritz J. Sedlazeck, Kishwar Shafin, Valery A. Shepelev, Alaina Shumate, Jessica M. Storer, Likhitha Surapaneni, Angela M. Taravella Oill, Françoise Thibaud-Nissen, Winston Timp, Marta Tomaszkiewicz, Mitchell R. Vollger, Brian P. Walenz, Allison C. Watwood, Matthias H. Weissensteiner, Aaron M. Wenger, Melissa A. Wilson, Samantha Zarate, Yiming Zhu, Justin M. Zook, Evan E. Eichler, Rachel J. O'Neill, Michael C. Schatz, Karen H. Miga, Kateryna D. Makova, and Adam M. Phillippy. The complete sequence of a human Y chromosome. *Nature*, 621(7978):344–354, September 2023. Publisher: Nature Publishing Group.

[155] Arang Rhie, Brian P. Walenz, Sergey Koren, and Adam M. Phillippy. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, 21(1):245, September 2020.

[156] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R. F. Twigg, Andrew O. M. Wilkie, Gil McVean, and Gerton Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet*, 46(8):912–918, August 2014. Publisher: Nature Publishing Group.

[157] Jared C. Roach, Gustavo Glusman, Arian F. A. Smit, Chad D. Huff, Robert Hubley, Paul T. Shannon, Lee Rowen, Krishna P. Pant, Nathan Goodman, Michael Bamshad, Jay Shendure, Radoje Drmanac, Lynn B. Jorde, Leroy Hood, and David J. Galas. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science*, 328(5978):636–639, April 2010. Publisher: American Association for the Advancement of Science.

[158] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nat Biotechnol*, 29(1):24–26, January 2011. Publisher: Nature Publishing Group.

[159] Nicole B. Rockweiler, Avinash Ramu, Liina Nagirnaja, Wing H. Wong, Michiel J. Noordam, Casey W. Drubin, Ni Huang, Brian Miller, Ellen Z. Todres, Katinka A. Vigh-Conrad, Antonino Zito, Kerrin S. Small, Kristin G. Ardlie, Barak A. Cohen, and Donald F. Conrad. The origins and functional effects of postzygotic mutations throughout the human life span. *Science*, 380(6641):eabn7113, 2023.

[160] Ashley D. Sanders, Ester Falconer, Mark Hills, Diana C. J. Spierings, and Peter M. Lansdorp. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat Protoc*, 12(6):1151–1176, June 2017. Publisher: Nature Publishing Group.

[161] Stephan J. Sanders, Xin He, A. Jeremy Willsey, A. Gulhan Ercan-Sencicek, Kaitlin E. Samocha, A. Ercument Cicek, Michael T. Murtha, Vanessa H. Bal, Somer L. Bishop, Shan Dong, Arthur P. Goldberg, Cai Jinlu, John F. Keaney, Lambertus Klei, Jeffrey D. Mandell, Daniel Moreno-De-Luca, Christopher S. Poultney, Elise B. Robinson, Louw Smith, Tor Solli-Nowlan, Mack Y. Su, Nicole A. Teran, Michael F. Walker, Donna M. Werling, Arthur L. Beaudet, Rita M. Cantor, Eric Fombonne, Daniel H. Geschwind, Dorothy E. Grice, Catherine Lord, Jennifer K. Lowe, Shrikant M. Mane, Donna M. Martin, Eric M. Morrow, Michael E. Talkowski, James S. Sutcliffe, Christopher A. Walsh, Timothy W. Yu, David H. Ledbetter, Christa Lese Martin, Edwin H. Cook, Joseph D. Buxbaum, Mark J. Daly, Bernie Devlin, Kathryn Roeder, and Matthew W. State. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron*, 87(6):1215–1233, September 2015.

[162] Thomas A Sasani, Brent S Pedersen, Ziyue Gao, Lisa Baird, Molly Przeworski, Lynn B Jorde, and Aaron R Quinlan. Large, three-generation human families reveal post-

zygotic mosaicism and variability in germline mutation accumulation. *eLife*, 8:e46922, September 2019. Publisher: eLife Sciences Publications, Ltd.

[163] Valerie A. Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, Kim D. Pruitt, Françoise Thibaud-Nissen, Derek Albracht, Robert S. Fulton, Milinn Kremitzki, Vincent Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, William Chow, Joanna Collins, Glenn Harden, Timothy Hubbard, Sarah Pelan, Jared T. Simpson, Glen Threadgold, James Torrance, Jonathan M. Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Paul Peluso, Heng Li, Chen-Shan Chin, Adam M. Phillippy, Richard Durbin, Richard K. Wilson, Paul Flicek, Evan E. Eichler, and Deanna M. Church. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*, 27(5):849–864, May 2017.

[164] Jonathan Sebat, B. Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese-Martin, Tom Walsh, Boris Yamrom, Seungtai Yoon, Alex Krasnitz, Jude Kendall, Anthony Leotta, Deepa Pai, Ray Zhang, Yoon-Ha Lee, James Hicks, Sarah J. Spence, Annette T. Lee, Kaija Puura, Terho Lehtimäki, David Ledbetter, Peter K. Gregersen, Joel Bregman, James S. Sutcliffe, Vaidehi Jobanputra, Wendy Chung, Dorothy Warburton, Mary-Claire King, David Skuse, Daniel H. Geschwind, T. Conrad Gilliam, Kenny Ye, and Michael Wigler. Strong Association of De Novo Copy Number Mutations with Autism. *Science*, 316(5823):445–449, April 2007. Publisher: American Association for the Advancement of Science.

[165] Arian F. A. Smit, Robert Hubley, and Phil Green. RepeatMasker Open-3.0, 1996.

[166] Tegan B. Smith, Matthew D. Dun, Nathan D. Smith, Ben J. Curry, Haley S. Connaughton, and Robert J. Aitken. The presence of a truncated base excision repair pathway in human spermatozoa that is mediated by OGG1. *Journal of Cell Science*, 126(6):1488–1497, March 2013.

[167] R M Smits, M J Xavier, M S Oud, G D N Astuti, A M Meijerink, P F de Vries, G S Holt, B K S Alobaidi, L E Batty, G Khazeeva, K Sablauskas, L E L M Vissers, C Gilissen, K Fleischer, D D M Braat, L Ramos, and J A Veltman. De novo mutations in children born after medical assisted reproduction. *Human Reproduction*, 37(6):1360–1369, June 2022.

[168] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, May 2014. Publisher: Oxford Academic.

[169] John A. Stamatoyannopoulos, Ivan Adzhubei, Robert E. Thurman, Gregory V. Kryukov, Sergei M. Mirkin, and Shamil R. Sunyaev. Human mutation rate associated with DNA replication timing. *Nat Genet*, 41(4):393–395, April 2009. Publisher: Nature Publishing Group.

[170] Cody J. Steely, W. Scott Watkins, Lisa Baird, and Lynn B. Jorde. The mutational dynamics of short tandem repeats in large, multigenerational families. *Genome Biol*, 23(1):253, December 2022.

[171] Jessica M. Stringer, Amy Winship, Nadeen Zerafa, Matthew Wakefield, and Karla Hutt. Oocytes can efficiently repair DNA double-strand breaks to restore genetic integrity and protect offspring health. *Proceedings of the National Academy of Sciences*, 117(21):11513–11522, May 2020. Publisher: Proceedings of the National Academy of Sciences.

[172] Peter H. Sudmant, Swapan Mallick, Bradley J. Nelson, Fereydoun Hormozdiari, Niklas Krumm, John Huddleston, Bradley P. Coe, Carl Baker, Susanne Nordenfelt, Michael Bamshad, Lynn B. Jorde, Olga L. Posukh, Hovhannes Sahakyan, W. Scott Watkins, Levon Yepiskoposyan, M. Syafiq Abdullah, Claudio M. Bravi, Cristian Capelli, Tor Hervig, Joseph T. S. Wee, Chris Tyler-Smith, George van Driem, Irene Gallego Romero, Aashish R. Jha, Sena Karachanak-Yankova, Draga Toncheva, David Comas, Brenna Henn, Toomas Kivisild, Andres Ruiz-Linares, Antti Sajantila, Ene Metspalu, Jüri Parik, Richard Villems, Elena B. Starikovskaya, George Ayodo, Cynthia M. Beall, Anna Di Rienzo, Michael F. Hammer, Rita Khusainova, Elza Khusnutdinova, William Klitz, Cheryl Winkler, Damian Labuda, Mait Metspalu, Sarah A. Tishkoff, Stanislav Dryomov, Rem Sukernik, Nick Patterson, David Reich, and Evan E. Eichler. Global diversity, population stratification, and selection of human copy-number variation. *Science*, 349(6253):aab3761, September 2015. Publisher: American Association for the Advancement of Science.

[173] Arvis Sulovari, Ruiyang Li, Peter A. Audano, David Porubsky, Mitchell R. Vollger, Glennis A. Logsdon, Wesley C. Warren, Alex A. Pollen, Mark J. P. Chaisson, and Evan E. Eichler. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc Natl Acad Sci U S A*, 116(46):23243–23253, November 2019.

[174] Amalio Telenti, Levi C. T. Pierce, William H. Biggs, Julia di Iulio, Emily H. M. Wong, Martin M. Fabani, Ewen F. Kirkness, Ahmed Moustafa, Naisha Shah, Chao Xie, Suzanne C. Brewerton, Nadeem Bulsara, Chad Garner, Gary Metzker, Efren Sandoval, Brad A. Perkins, Franz J. Och, Yaron Turpaz, and J. Craig Venter. Deep sequencing of 10,000 human genomes. *Proceedings of the National Academy of Sciences*,

113(42):11901–11906, October 2016. Publisher: Proceedings of the National Academy of Sciences.

[175] Kosuke M Teshima and Hideki Innan. The Coalescent with Selection on Copy Number Variants. *Genetics*, 190(3):1077–1086, March 2012.

[176] Tychele N. Turner, Bradley P. Coe, Diane E. Dickel, Kendra Hoekzema, Bradley J. Nelson, Michael C. Zody, Zev N. Kronenberg, Fereydoun Hormozdiari, Archana Raja, Len A. Pennacchio, Robert B. Darnell, and Evan E. Eichler. Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell*, 171(3):710–722.e12, October 2017.

[177] Arikuni Uchimura, Hirotaka Matsumoto, Yasunari Satoh, Yohei Minakuchi, Sayaka Wakayama, Teruhiko Wakayama, Mayumi Higuchi, Masakazu Hashimoto, Ryutaro Fukumura, Atsushi Toyoda, Yoichi Gondo, and Takeshi Yagi. Early embryonic mutations reveal dynamics of somatic and germ cell lineages in mice. *Genome Res*, 32(5):945–955, May 2022.

[178] Joris A. Veltman and Han G. Brunner. De novo mutations in human genetic disease. *Nat Rev Genet*, 13(8):565–575, August 2012. Publisher: Nature Publishing Group.

[179] Manuel Viotti. Preimplantation Genetic Testing for Chromosomal Abnormalities: Aneuploidy, Mosaicism, and Structural Rearrangements. *Genes*, 11(6):602, June 2020. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[180] Mitchell R. Vollger, Philip C. Dishuck, William T. Harvey, William S. DeWitt, Xavi Guitart, Michael E. Goldberg, Allison N. Rozanski, Julian Lucas, Mobin Asri, Katherine M. Munson, Alexandra P. Lewis, Kendra Hoekzema, Glennis A. Logsdon, David Porubsky, Benedict Paten, Kelley Harris, PingHsun Hsieh, and Evan E. Eichler. Increased mutation and gene conversion within human segmental duplications. *Nature*, 617(7960):325–334, May 2023. Publisher: Nature Publishing Group.

[181] Mitchell R. Vollger, Philip C. Dishuck, Melanie Sorensen, AnneMarie E. Welch, Vy Dang, Max L. Dougherty, Tina A. Graves-Lindsay, Richard K. Wilson, Mark J. P. Chaisson, and Evan E. Eichler. Long-read sequence and assembly of segmental duplications. *Nat Methods*, 16(1):88–94, January 2019. Publisher: Nature Publishing Group.

[182] Mitchell R. Vollger, Xavi Guitart, Philip C. Dishuck, Ludovica Mercuri, William T. Harvey, Ariel Gershman, Mark Diekhans, Arvis Sulovari, Katherine M. Munson, Alexandra P. Lewis, Kendra Hoekzema, David Porubsky, Ruiyang Li, Sergey Nurk, Sergey Koren, Karen H. Miga, Adam M. Phillippy, Winston Timp, Mario Ventura, and Evan E. Eichler. Segmental duplications and their variation in a complete human

genome. *Science*, 376(6588):eabj6965, April 2022. Publisher: American Association for the Advancement of Science.

[183] Ludmila Volozonoka, Anna Miskova, and Linda Gailite. Whole Genome Amplification in Preimplantation Genetic Testing in the Era of Massively Parallel Sequencing. *International Journal of Molecular Sciences*, 23(9):4819, January 2022. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[184] L E Voullaire, G C Webb, and M A Leversha. Chromosome deletion at 11q23 in an abnormal child from a family with inherited fragility at 11q23. *Hum Genet*, 76(2):202–204, June 1987.

[185] Jeremiah Wala, Cheng-Zhong Zhang, Matthew Meyerson, and Rameen Beroukhim. VariantBam: filtering and profiling of next-generational sequencing data using region-specific rules. *Bioinformatics*, 32(13):2029–2031, July 2016.

[186] Cheng Wang, Hong Lv, Xiufeng Ling, Hong Li, Feiyang Diao, Juncheng Dai, Jiangbo Du, Ting Chen, Qi Xi, Yang Zhao, Kun Zhou, Bo Xu, Xiumei Han, Xiaoyu Liu, Meijuan Peng, Congcong Chen, Shiyao Tao, Lei Huang, Cong Liu, Mingyang Wen, Yangqian Jiang, Tao Jiang, Chuncheng Lu, Wei Wu, Di Wu, Minjian Chen, Yuan Lin, Xuejiang Guo, Ran Huo, Jiayin Liu, Hongxia Ma, Guangfu Jin, Yankai Xia, Jiahao Sha, Hongbing Shen, and Zhibin Hu. Association of assisted reproductive technology, germline de novo mutations and congenital heart defects in a prospective birth cohort study. *Cell Res*, 31(8):919–928, August 2021.

[187] Nan Wang, Peng Chen, Yuanyuan Xu, Lingxia Guo, Xianxin Li, Hualin Yi, Robert M Larkin, Yongfeng Zhou, Xiuxin Deng, and Qiang Xu. Phased genomics reveals hidden somatic mutations and provides insight into fruit development in sweet orange. *Horticulture Research*, 11(2):uhad268, February 2024.

[188] Krystyna Wasilewska, Tomasz Gambin, Małgorzata Rydzanicz, Krzysztof Szczałuba, and Rafał Płoski. Postzygotic mutations and where to find them – recent advances and future implications in the field of non-neoplastic somatic mosaicism. *Mutation Research/Reviews in Mutation Research*, 790:108426, 2022.

[189] Amy B. Wilfert, Tychele N. Turner, Shwetha C. Murali, PingHsun Hsieh, Arvis Sulovari, Tianyun Wang, Bradley P. Coe, Hui Guo, Kendra Hoekzema, Trygve E. Bakken, Lara H. Winterkorn, Uday S. Evani, Marta Byrska-Bishop, Rachel K. Earl, Raphael A. Bernier, Michael C. Zody, and Evan E. Eichler. Recent ultra-rare inherited variants implicate new autism candidate risk genes. *Nat Genet*, 53(8):1125–1134, August 2021. Publisher: Nature Publishing Group.

[190] Raf Winand, Kristien Hens, Wybo Dondorp, Guido de Wert, Yves Moreau, Joris Robert Vermeesch, Inge Liebaers, and Jan Aerts. In vitro screening of embryos by whole-genome sequencing: now, in the future or never? *Human Reproduction*, 29(4):842–851, April 2014.

[191] C. F. Wright, E. Prigmore, D. Rajan, J. Handsaker, J. McRae, J. Kaplanis, T. W. Fitzgerald, D. R. FitzPatrick, H. V. Firth, and M. E. Hurles. Clinically-relevant postzygotic mosaicism in parents and children with developmental disorders in trio exome sequencing data. *Nat Commun*, 10(1):2985, July 2019. Publisher: Nature Publishing Group.

[192] Felix L. Wu, Alva I. Strand, Laura A. Cox, Carole Ober, Jeffrey D. Wall, Priya Moorjani, and Molly Przeworski. A comparison of humans and baboons suggests germline mutation rates do not track cell divisions. *PLOS Biology*, 18(8):e3000838, August 2020. Publisher: Public Library of Science.

[193] Chunlin Xiao, Zhong Chen, Wanqiu Chen, Cory Padilla, Michael Colgan, Wenjun Wu, Li-Tai Fang, Tiantian Liu, Yibin Yang, Valerie Schneider, Charles Wang, and Wenming Xiao. Personalized genome assembly for accurate cancer somatic mutation discovery using tumor-normal paired reference samples. *Genome Biology*, 23(1):237, November 2022.

[194] Yali Xue, Qiuju Wang, Quan Long, Bee Ling Ng, Harold Swerdlow, John Burton, Carl Skuce, Ruth Taylor, Zahra Abdellah, Yali Zhao, Asan, Daniel G. MacArthur, Michael A. Quail, Nigel P. Carter, Huanming Yang, and Chris Tyler-Smith. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr Biol*, 19(17):1453–1457, September 2009.

[195] Yaping Yang, Donna M. Muzny, Fan Xia, Zhiyv Niu, Richard Person, Yan Ding, Patricia Ward, Alicia Braxton, Min Wang, Christian Buhay, Narayanan Veeraraghavan, Alicia Hawes, Theodore Chiang, Magalie Leduc, Joke Beuten, Jing Zhang, Weimin He, Jennifer Scull, Alecia Willis, Megan Landsverk, William J. Craigen, Mir Reza Bekheirnia, Asbjorg Stray-Pedersen, Pengfei Liu, Shu Wen, Wendy Alcaraz, Hong Cui, Magdalena Walkiewicz, Jeffrey Reid, Matthew Bainbridge, Ankita Patel, Eric Boerwinkle, Arthur L. Beaudet, James R. Lupski, Sharon E. Plon, Richard A. Gibbs, and Christine M. Eng. Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. *JAMA*, 312(18):1870–1879, November 2014.

[196] Taedong Yun, Helen Li, Pi-Chuan Chang, Michael F Lin, Andrew Carroll, and Cory Y McLean. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics*, 36(24):5582–5589, April 2021.

[197] Jian Zhou, Christopher Y. Park, Chandra L. Theesfeld, Aaron K. Wong, Yuan Yuan, Claudia Scheckel, John J. Fak, Julien Funk, Kevin Yao, Yoko Tajima, Alan Packer, Robert B. Darnell, and Olga G. Troyanskaya. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat Genet*, 51(6):973–980, June 2019. Publisher: Nature Publishing Group.

[198] Justin M. Zook, Jennifer McDaniel, Nathan D. Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A. Irvine, Len Trigg, Rebecca Truty, Cory Y. McLean, Francisco M. De La Vega, Chunlin Xiao, Stephen Sherry, and Marc Salit. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol*, 37(5):561–566, May 2019. Publisher: Nature Publishing Group.

[199] Emile Zuckerkandl and Linus Pauling. Molecular Disease, Evolution, and Genic Heterogeneity. *Horizons in Biochemistry*, pages 189–225, 1962.

Appendix A

CHAPTER 2 SUPPLEMENT

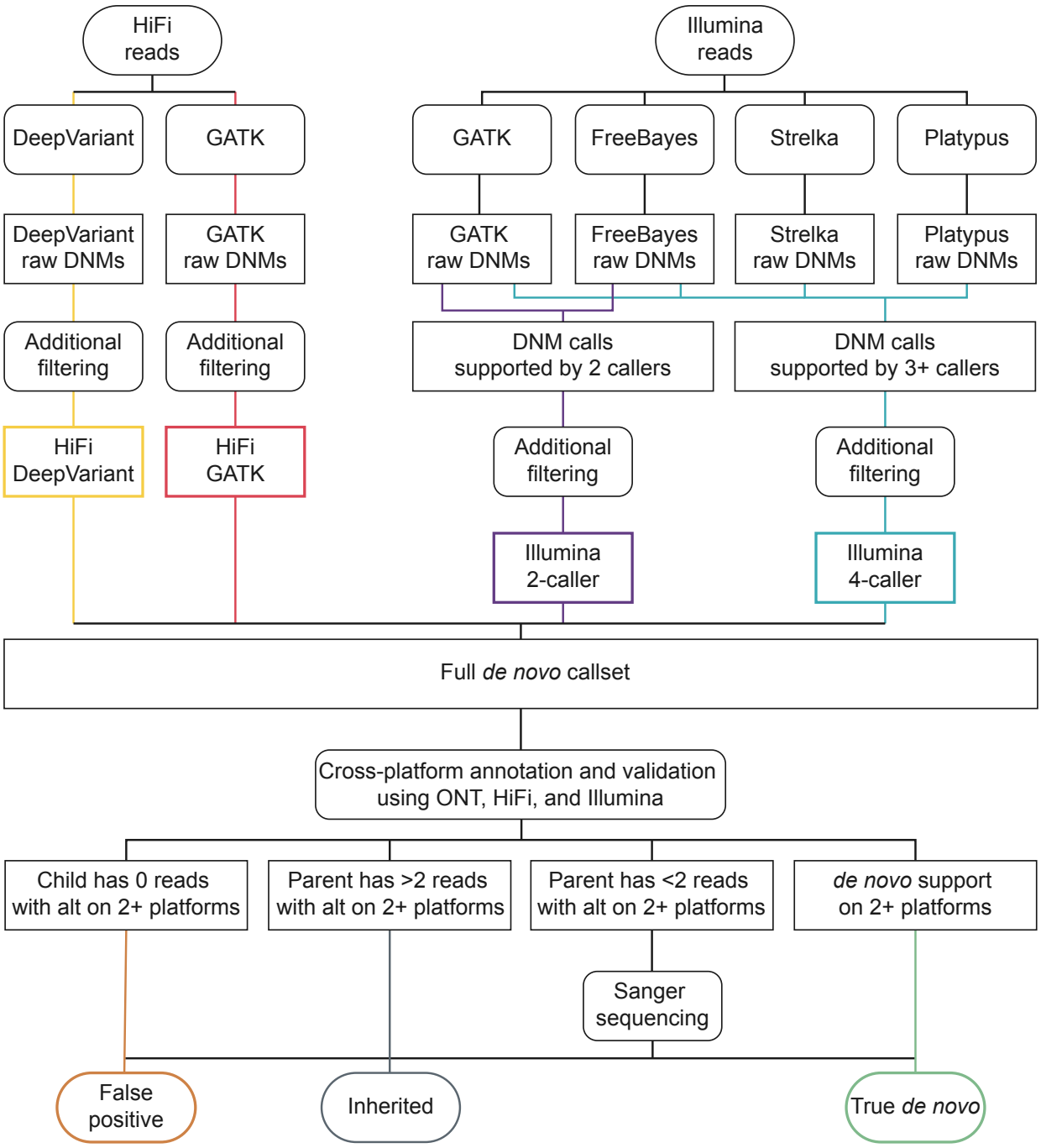


Figure A.1: **Pipeline for *de novo* SNV and small (<20 bp) indel identification.** PacBio HiFi and Illumina reads were used for DNM discovery. Both technologies were used in addition to Oxford Nanopore Sequencing (ONT) data for validation of *de novo* candidate sites.

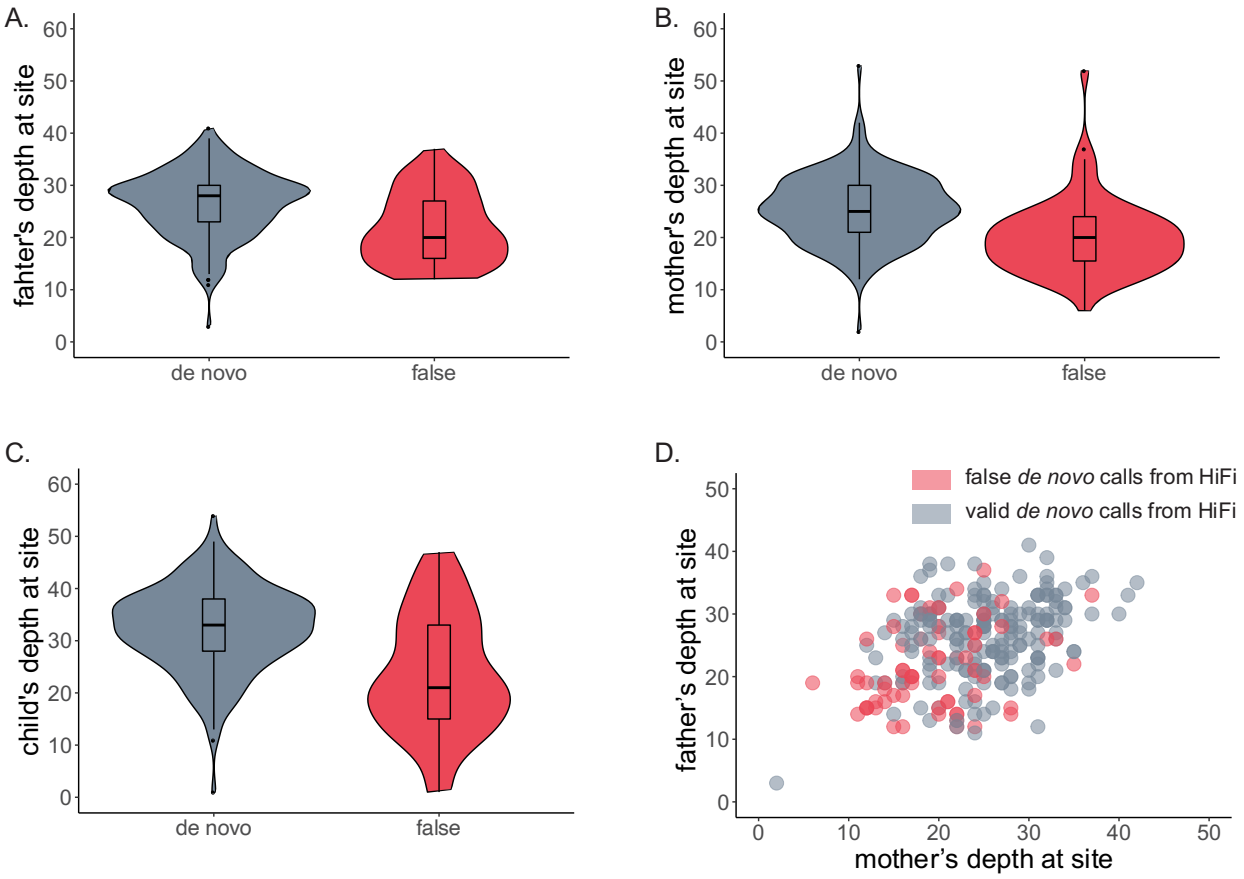


Figure A.2: **Read depth comparison between true and false calls made by HiFi callers.** For all true *de novo* and false calls made by HiFi callers, a comparison of read depth at the site of the call in (A) the father, (B) the mother, and (C) the child with the *de novo* call. In (D), the father's depth at the site plotted against the mother's depth, false calls show significantly lower parental read depth than true calls.

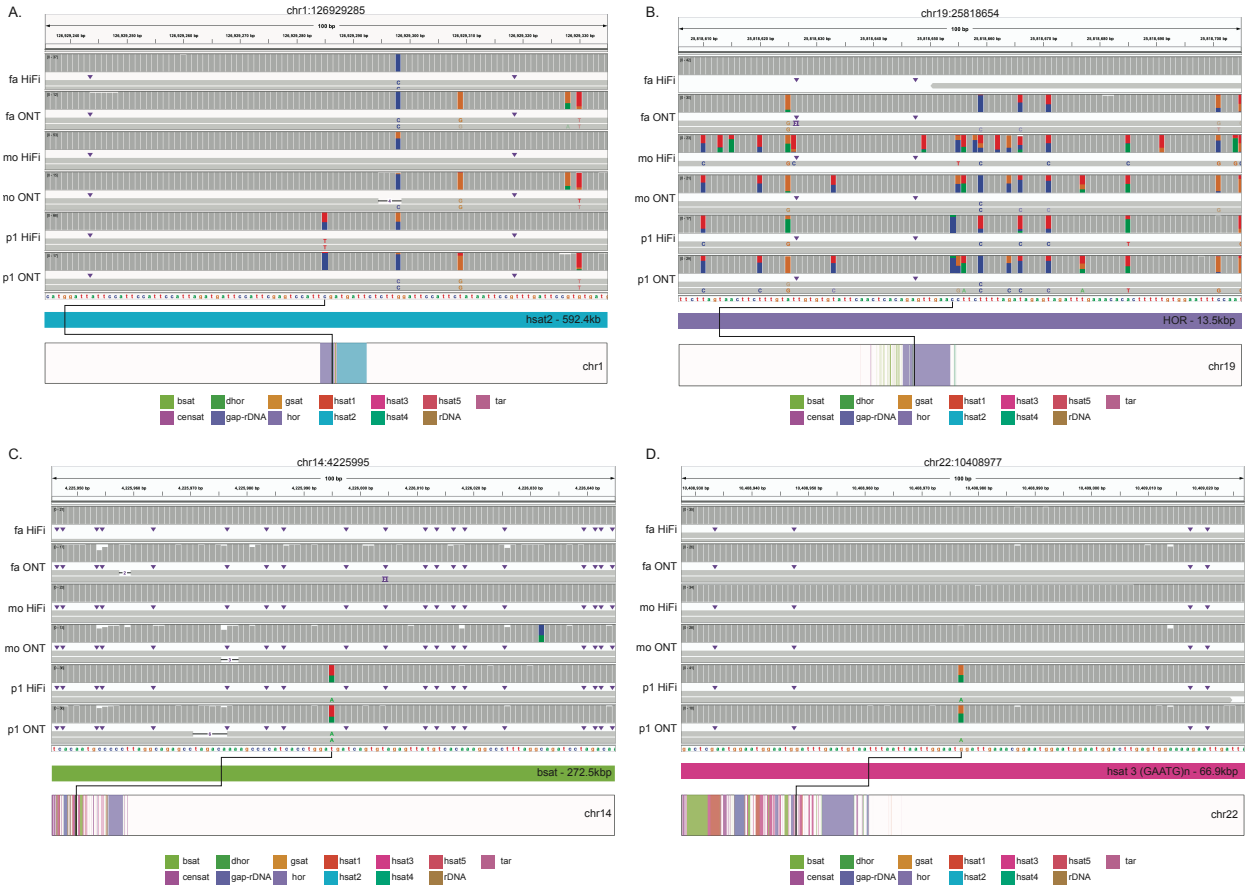


Figure A.3: **IGV shots of centromeric DNMs.** (A-D) IGV shots of PacBio HiFi and ONT reads aligned to centromeric heterochromatic satellite regions. Underneath the reads is the location of the variant in its repetitive context, and below that is the chromosome with centromeric repeats annotated. All DNMs except B (chr19_25818654_C_A) are considered true positive events.

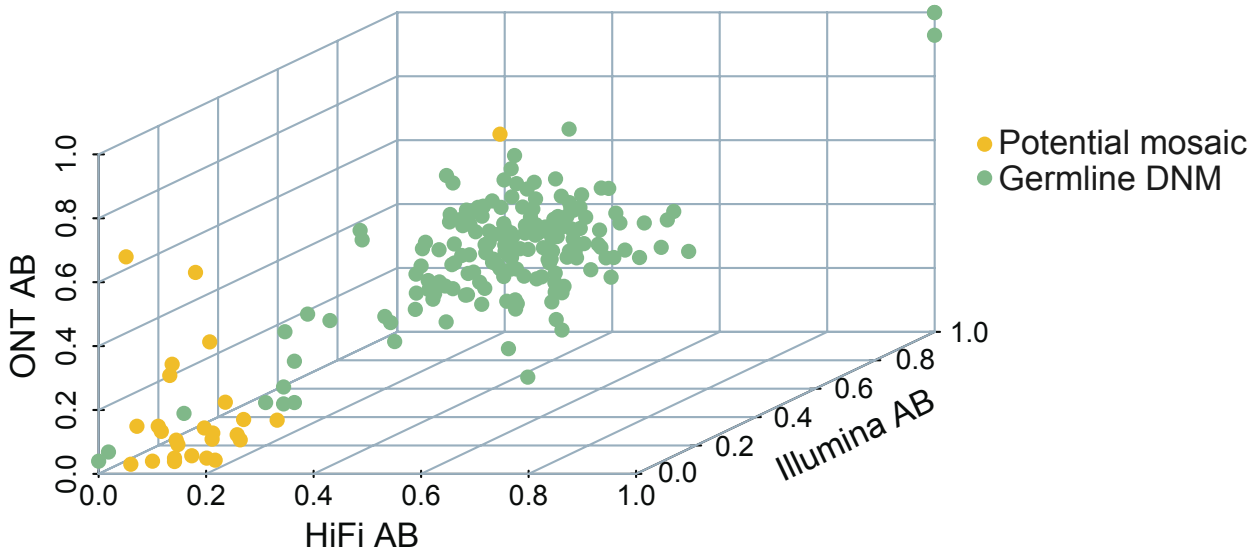


Figure A.4: **Allele balance in validated DNMs and potential mosaic mutations.** The allele balance in the child with the mutation is shown across three sequencing platforms: ONT, PacBio HiFi, and Illumina. The potential mosaic mutations are in yellow, and validated *de* novo SNVs are shown in green. The germline sites that are clustered with the mosaic mutations were only observed in GRCh38-aligned reads, and accordingly have very low allele balance (AB) in T2T-CHM13 aligned data.

Chr	Position	Type	Motif	Calling approach	Fa repeat copies	Mo repeat copies	P1 repeat copies	S1 repeat copies	Sanger validation
Chr1	69781685	STR	AAAG	LPT	68 | 77	67 | 71	72 | 79	67 | 78	True positive
Chr1	103525498	VNTR	ACGGCGGGGCGGGGCGC	ExpansionHunter Denovo	9 | 14	11 | 11	11 | 15	11 | 14	True positive
Chr2	220546187	STR	TTC	LPT	131 | 132	112 | 124	124 | 125	112 | 132	True positive
Chr10	46271376	STR	CTAACT	30-mer	30 | 44	23 | 37	30 | 37	23 | 46	True positive
Chr2	133895459	VNTR	AAGAGAGAGGGGAGG	ExpansionHunter Denovo	12 | 18	6 | 18	17 | 18	12 | 17	Inherited
Chr3	111780258	STR	AAG	LPT	66 | 67	69 | 71	67 | 69	67 | 71	Inherited
Chr5	38824434	STR	CCACCA	30-mer	18 | 21	18 | 19	18 | 19	19 | 21	Inherited
Chr13	44142133	STR	CTCGG	LPT	18 | 25	18 | NA	18 | 25	18 | 18	Inherited
Chr17	51831668	STR	AGC	LPT	22 | 23	19 | 22	19 | 22	19 | 23	Inherited
Chr1	101657855	STR	TTC	LPT	NA	NA	NA	NA	Not supported
Chr7	84690930	STR	GAA	LPT	79 | NA	71 | NA	NA	71| 72	Not supported
Chr12	111257196	VNTR	AAGAAGTGGGAGGG	ExpansionHunter Denovo	33 | 37	36 | 36	37 | NA	37 | NA	Not supported
Chr14	44005178	STR	AGA	LPT	NA	71 | NA	76 | 77	76 | 77	Not supported
Chr14	99927754	STR	GTG	LPT	NA	NA	NA	NA	Not supported
Chr16	20041114	STR	AGGAG	LPT	NA	NA	NA	NA	Not supported

Table A.1: **All candidate STR and VNTR events.** Events were identified by one of three approaches: a custom k-mer based approach (30-mer), an approach based on the longest pure tandem repeat (LPT), and the tool ExpansionHunter Denovo.

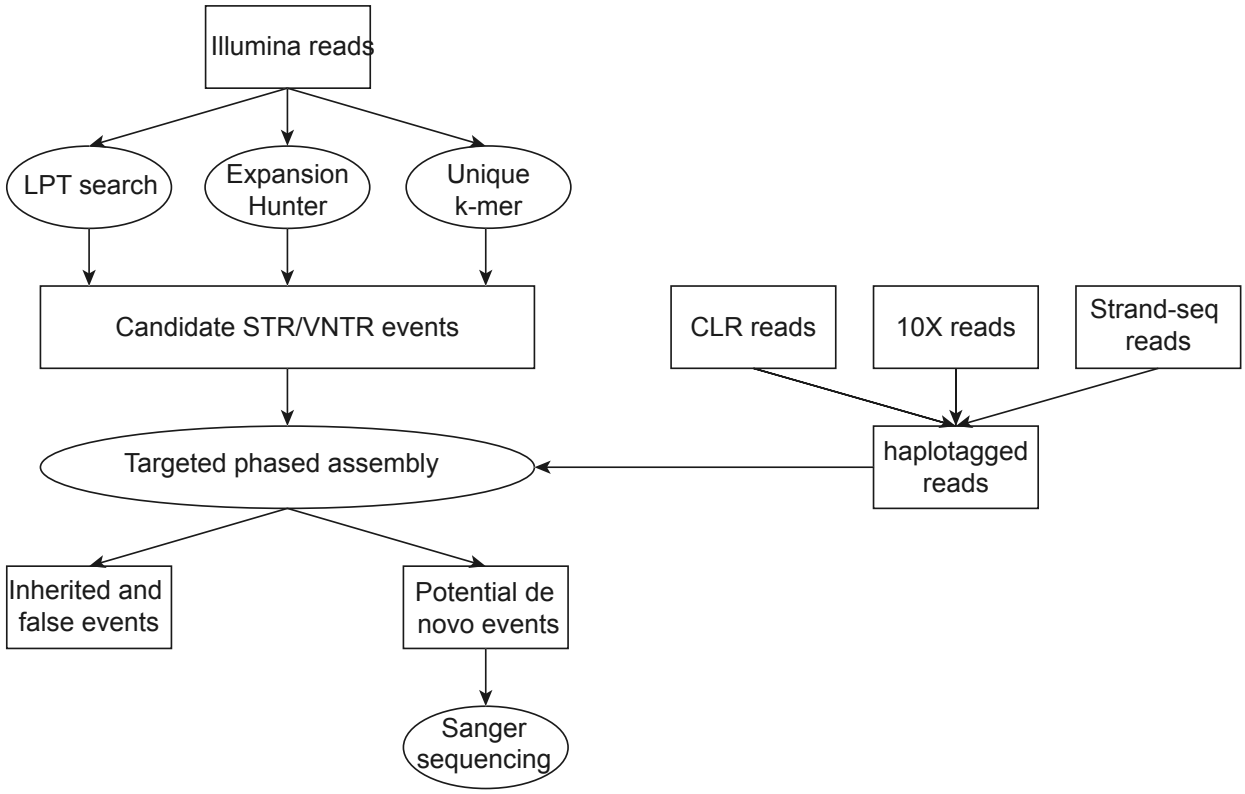


Figure A.5: **STR and VNTR calling pipeline.** This pipeline identifies candidate STR and VNTR mutations in Illumina data. HiFi continuous long-read (CLR) data were assigned haplotypes using Chromium 10X genomic sequencing (10X) and single-cell DNA template strand sequencing (Strand-seq) data. These haplotagged reads were used for targeted phased assembly in order to validate candidate events; all were present in the assemblies.

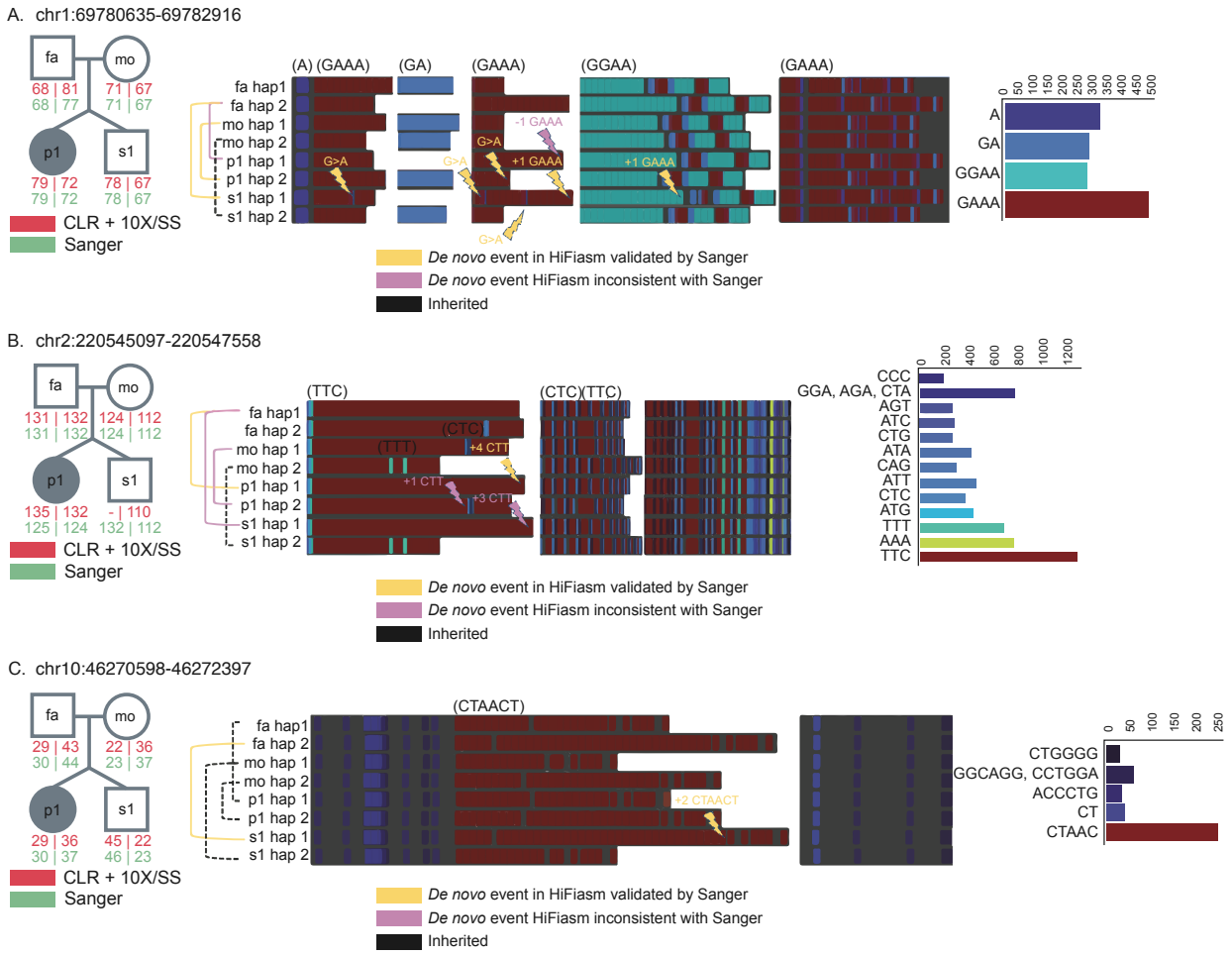


Figure A.6: **Identified *de novo* STR events.** The structure of the family annotated with the number of STR copies detected in PacBio CLR haplotagged with 10X and Strand-seq data, and the assembled haplotypes for each individual, with variants highlighted by lightning bolts, depicted for an STR event in the (A) proband and sibling, (B) in the proband, and (C) the sibling.

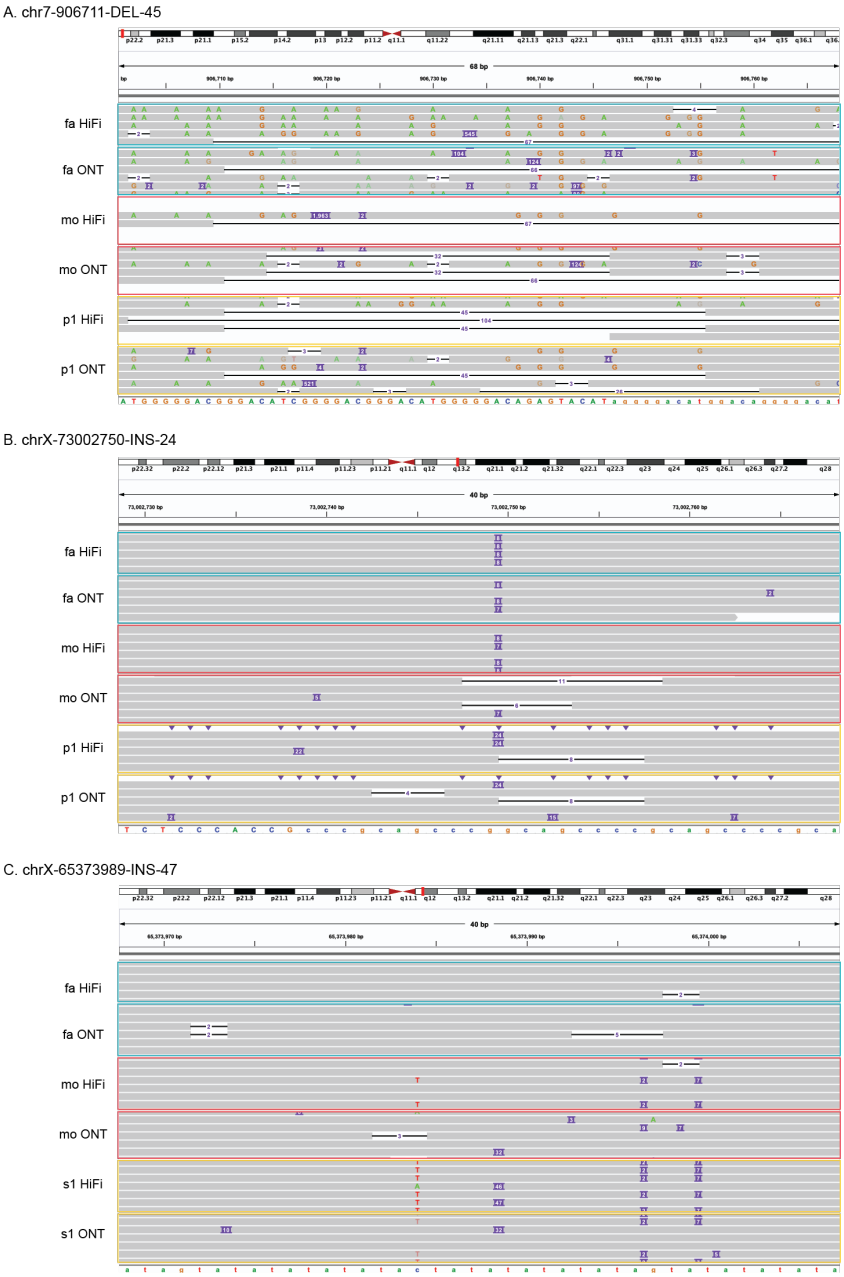


Figure A.7: **IGV shots of *de novo* 20-50bp indels.** IGV views of indel regions in the father, mother, and affected child across PacBio HiFi and ONT sequencing reads for (A) a 45bp *de novo* deletion in the proband, (B) a 16bp insertion in the proband, and (C) a 47bp insertion in the sibling.

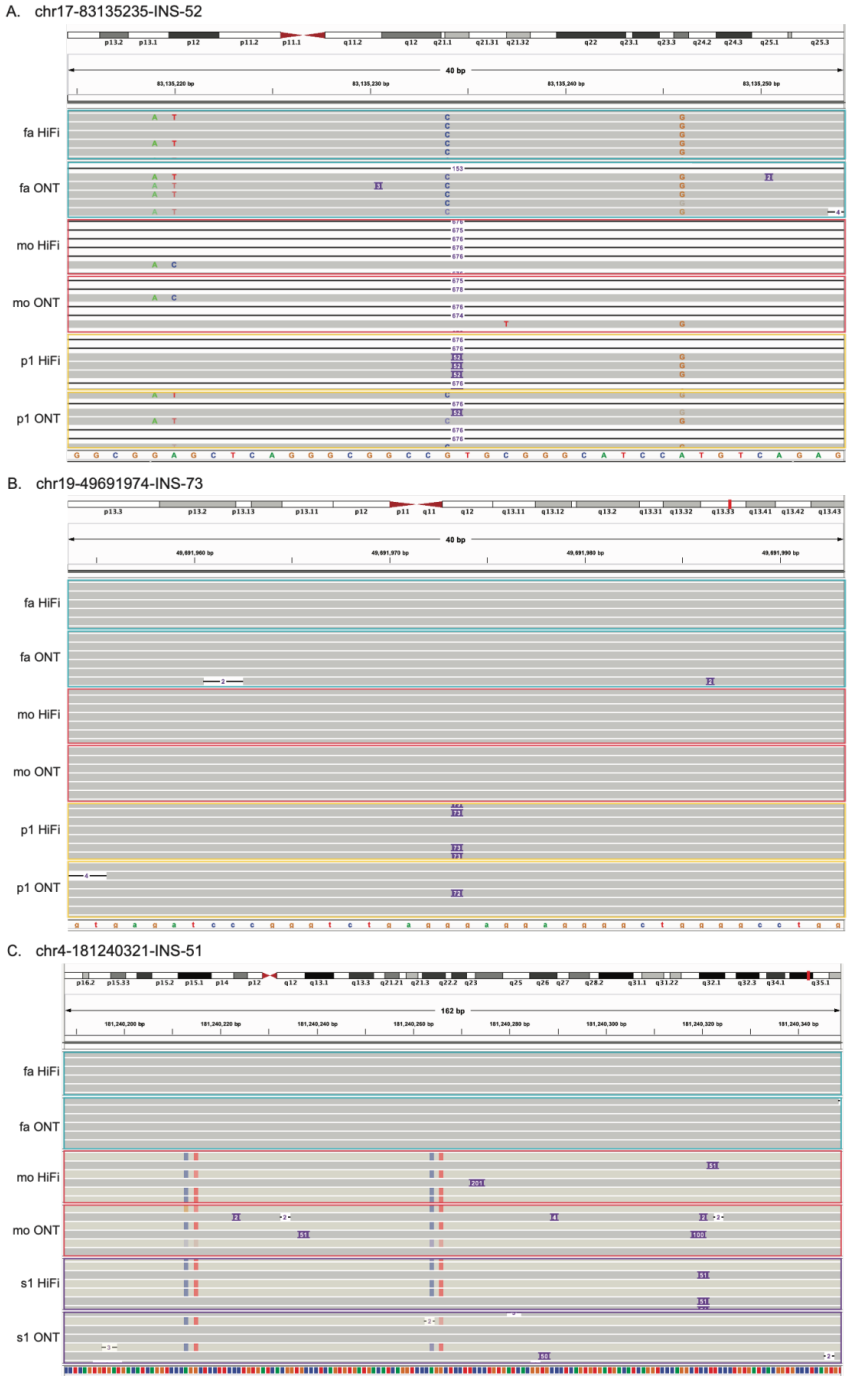


Figure A.8: **IGV shots of *de novo* SVs.** IGV views of SV regions in the father, mother, and affected child across PacBio HiFi and ONT sequencing reads for (A) the 52 bp insertion in the proband, (B) the 73 bp insertion in the proband, and (C) the 51 bp insertion in the sibling.

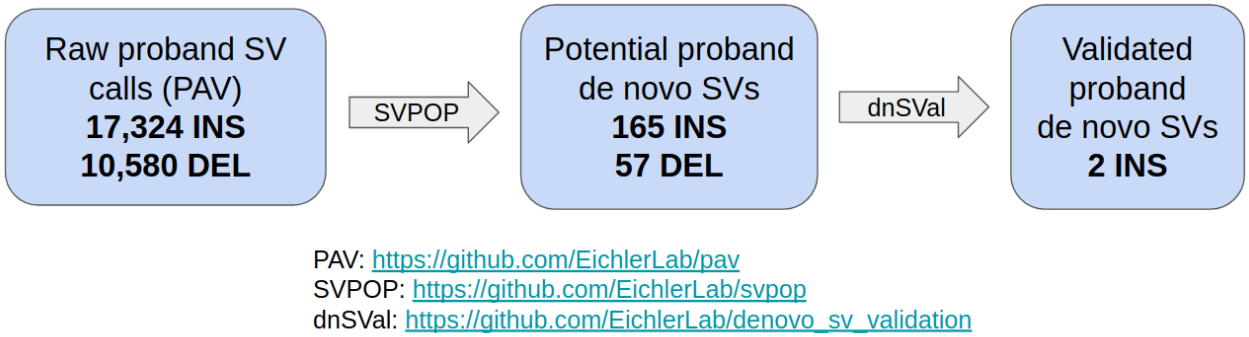


Figure A.9: **Overview of automated SV filtering process.** From the initial 27,904 *de novo* SV calls made in the proband, automated SVPOP filtering removed all but 232. The dnSVal validation uses subseq and multiple sequence alignment to further filter candidate *de novo* SV calls, resulting in a total of 2 validated *de novo* events in the proband – the same 2 that passed the manual filtering process.

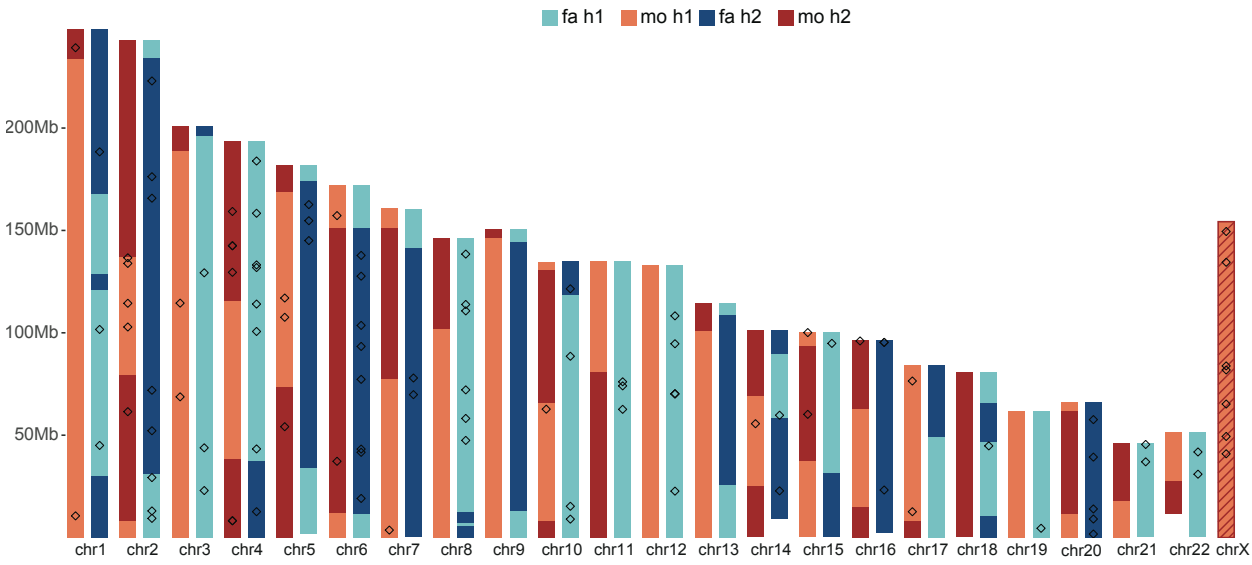


Figure A.10: **Meiotic crossovers and DNM.** A genome-wide overview of detected meiotic recombination breakpoints for the sibling. Inherited segments of maternal homologs (H1-light red, H2-dark red) appear on the left side of each chromosome while inherited segments of paternal homologs (H1-light blue, H2-dark blue) appear on the right side of each chromosome. Recombination breakpoints are visible as changes from H1 to H2 segments and vice versa. Detected DNMs (n=105) that could have been assigned to a single parental homolog are shown as empty boxes over maternal (left) and paternal (right) homologs. This individual is a male meaning that maternal chromosome X does not recombine (empty red box).

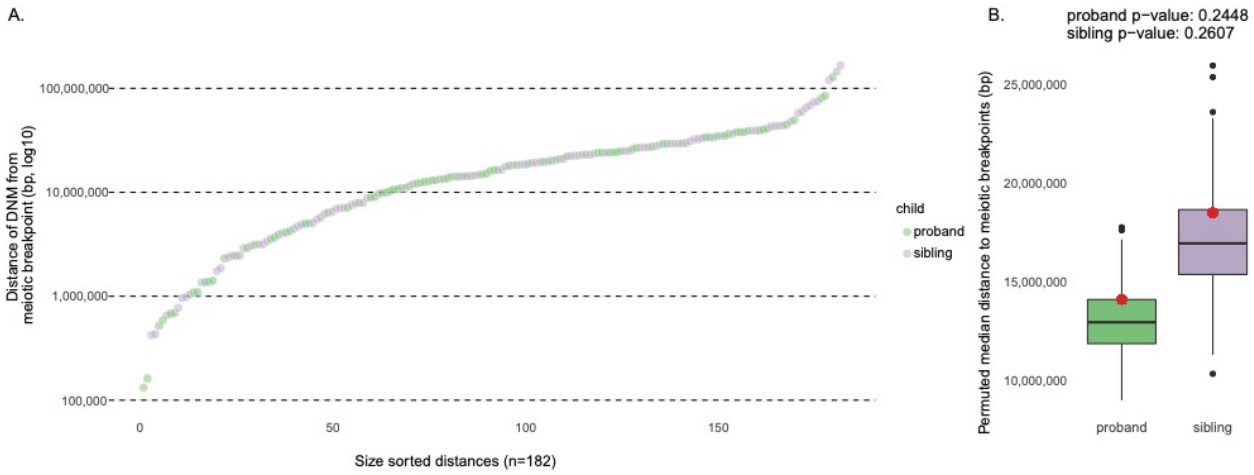


Figure A.11: **Meiotic recombination distance to DNMs.** A. Sorted distances of DNMs to the closest meiotic breakpoint reported for both proband (green) and sibling (purple). B. An enrichment analysis comparing observed median distance of DNMs to meiotic breakpoints in comparison to permuted meiotic breakpoints (1000 permutations) separately for proband- and sibling-specific DNMs.

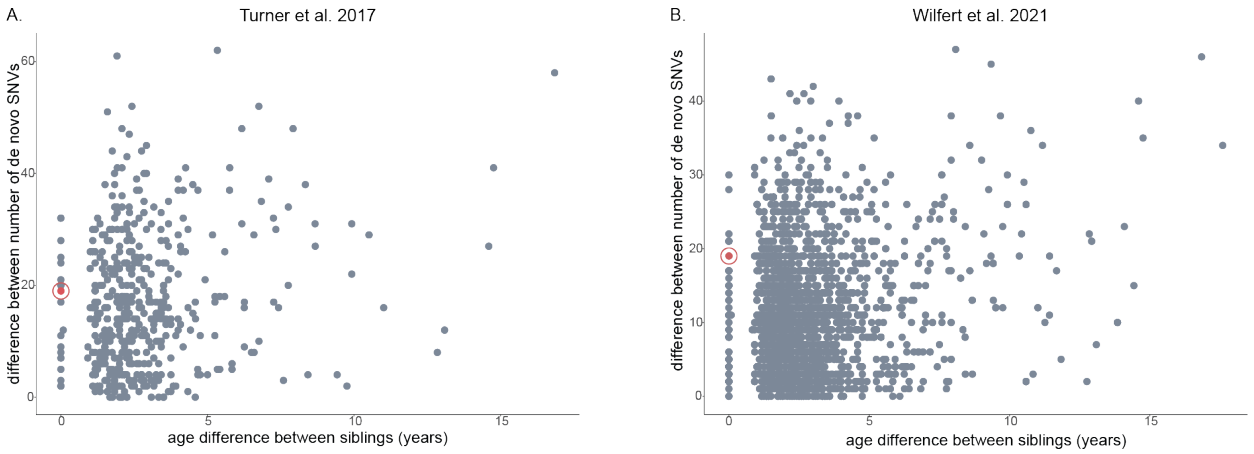


Figure A.12: **Intersibling *de novo* mutation difference.** The differences in the DNM count are compared between proband and sibling as a function of the age difference between the siblings based on two previous studies. Red dots indicate Illumina-based estimates of intersibling DNM difference for the family studied here (SSC14455).

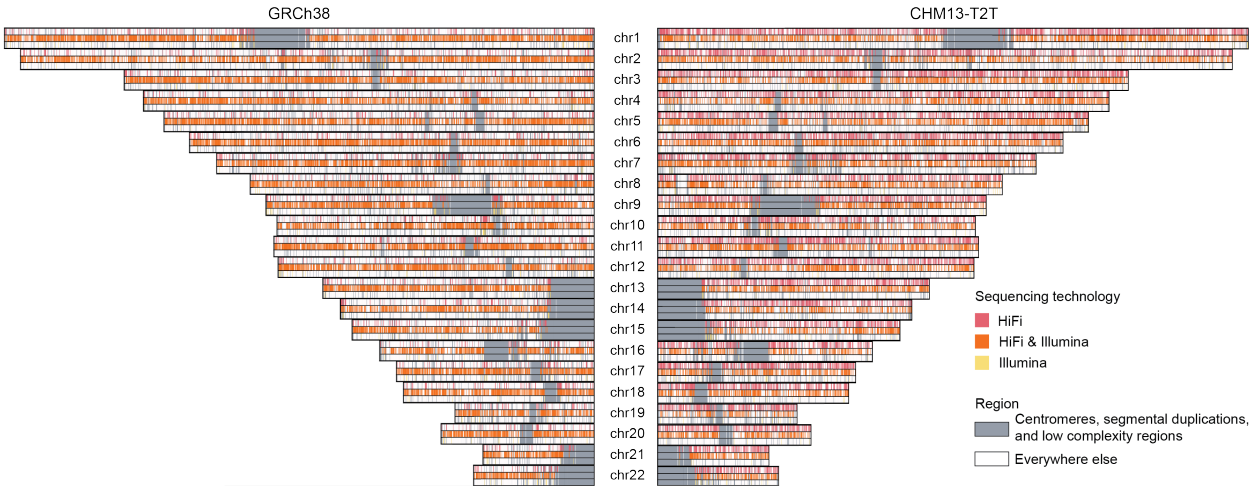


Figure A.13: **Distribution of rare inherited SNVs.** Variant calls were made using GATK on GRCh38- (left) and T2T-CHM13- (right) aligned reads. Inherited variants were identified using a modification of the *de novo* pipeline to select all variants with genotype 0/1 or 1/1 in exactly one parent, and 0/1 in at least one child. Inherited candidates were filtered for depth >10 in both parents and children, genotype quality >25 in both parents and children, and allele balance >0.2 in all individuals with the variant. Variants were annotated with VEP for their frequency in gnomAD, and all variants with allele frequency less than 0.1% were classified as rare. Any variant that was confirmed to be present in the child's ONT data was retained for the final callset.

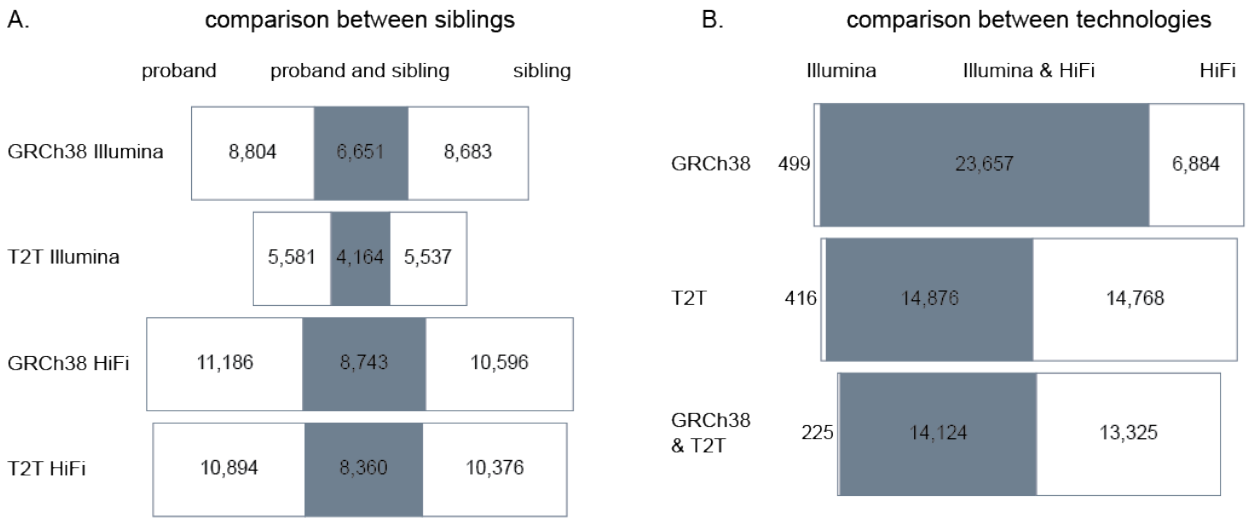


Figure A.14: **Inherited variants by child and technology.** ONT-validated rare inherited SNV (<0.1% frequency in gnomAD) discovery comparing (A) proband and sibling callsets generated by Illumina or HiFi reads aligned to the GRCh38 or T2T-CHM13 references or (B) comparing discovery based on use of different sequencing technologies. In (B), sites identified in the same sample[s] were considered to be common to both callsets, whereas the same site identified in different samples would be considered unique to each callset. Gray bars in the Venn diagram represent shared SNVs based on platform or assembly while white bars represent SNVs unique to each. Inherited callsets generated using HiFi were larger than their Illumina counterparts, and nearly every site in the short-read callsets was also present in long-read callsets. The number of novel inherited T2T-CHM13 calls was fewer than GRCh38 callsets for both short and long reads—driven in part by a failure to liftover T2T-CHM13 to GRCh38 genomic coordinates.

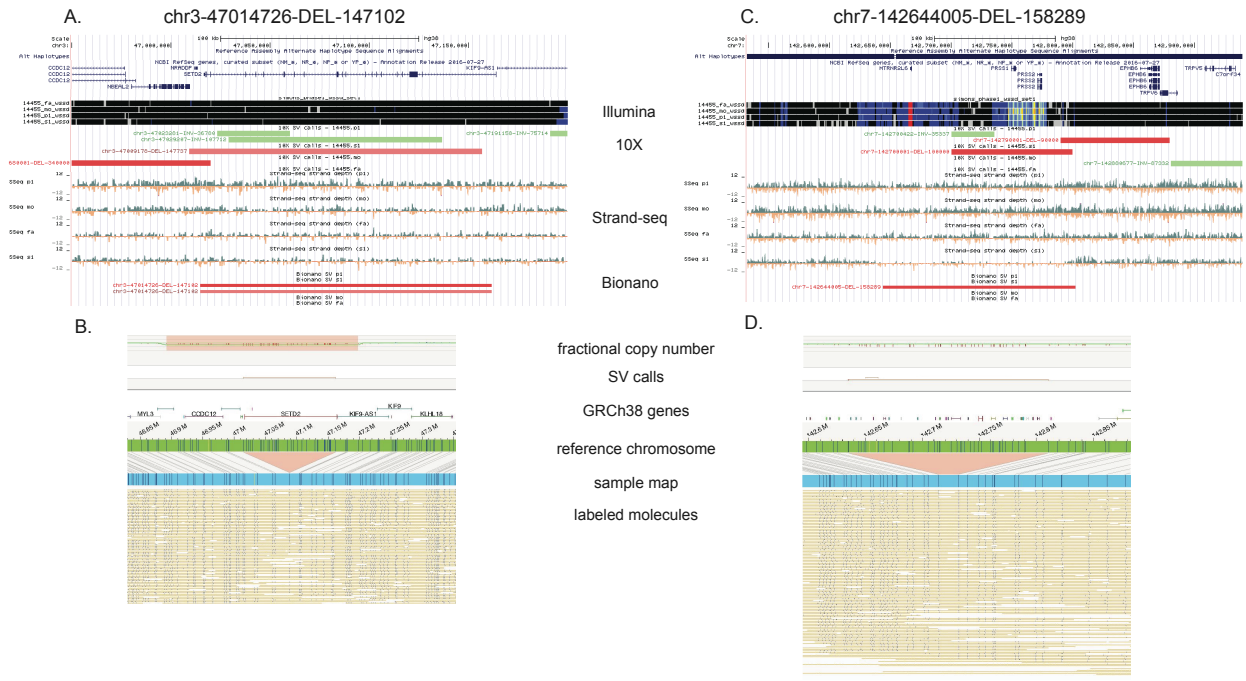


Figure A.15: **Support for cell line artifacts.** A. chr3-47014726-DEL-147102. Large sibling SETD2 deletion cell line artifact discovered by 10X (LCL derived) and Bionano (LCL derived), but not supported by Strand-seq (LCL derived) or WSSD (blood derived). B. Bionano support for SETD2 deletion. The top bar represents the fractional copy number (green line is copy number two); the second bar is SV calls based on the assembly map. The red lines compare the green reference region to the sample map (the label pattern observed over the region). The bottom represents all the labeled molecules over the region. C. chr7-142644005-DEL-158289. Large deletion over PRSS1 and PRSS2 is likely a cell line artifact with 10X, Bionano, and Strand-seq support. D. Bionano support for PRSS1 and PRSS2 deletion, in the same format as B.

Child	Chr	Position	Indel type	Length	HiFi/ONT Validation
14455.p1	chr7	906710	DEL	45	**potential *de novo***
14455.p1	chr7	132561922	DEL	28	false positive
14455.p1	chr11	82293621	DEL	24	inherited from mom
14455.p1	chr16	14928921	DEL	27	inherited from mom
14455.p1	chr16	61425944	DEL	35	false positive
14455.p1	chr19	15778633	DEL	40	false positive
14455.p1	chr21	39583858	DEL	33	inherited from dad
14455.p1	chr2	90033846	INS	21	false positive
14455.p1	chr2	114396849	INS	44	inherited from mom
14455.p1	chr6	95411818	INS	29	inherited from mom
14455.p1	chr8	57949332	INS	44	inherited from mom
14455.p1	chr9	71816230	INS	43	false positive
14455.p1	chr9	137350578	INS	33	inherited
14455.p1	chr12	3978627	INS	29	false positive
14455.p1	chr13	84254692	INS	25	inherited from dad
14455.p1	chr15	99927049	INS	22	false positive
14455.p1	chr18	32773926	INS	32	inherited from mom
14455.p1	chr20	31890154	INS	23	inherited
14455.p1	chr22	25093859	INS	38	false positive
14455.p1	chr22	28867115	INS	25	false positive
14455.p1	chrX	73002749	INS	24	**potential *de novo***
14455.s1	chr8	67056650	DEL	24	inherited from mom
14455.s1	chr10	128976652	DEL	32	inherited from dad
14455.s1	chrX	40557339	DEL	48	inherited from mom
14455.s1	chr1	19020904	INS	45	inherited from mom
14455.s1	chr4	182833096	INS	42	inherited from mom

14455.s1	chr5	29815038	INS	28	inherited from mom
14455.s1	chr8	108938105	INS	44	inherited from mom
14455.s1	chr8	128825889	INS	32	false positive
14455.s1	chr8	143218041	INS	21	inherited
14455.s1	chr9	42090783	INS	44	inherited from mom
14455.s1	chr10	131625735	INS	27	inherited from mom
14455.s1	chr11	33614584	INS	21	inherited from dad
14455.s1	chr12	76955213	INS	24	inherited from mom
14455.s1	chr13	87253488	INS	23	inherited from dad
14455.s1	chr17	71886772	INS	37	inherited
14455.s1	chr18	22204288	INS	48	inherited
14455.s1	chrX	65373988	INS	47	**potential *de novo***
14455.s1	chrY	1735802	INS	34	no read data

Table A.2: All 39 20-50bp indel calls identified using assembly-driven variant discovery.

Child	Chr	Position	SV type	Length	Population Frequency (%)
14455.p1	chr3	90544141	DEL	4,656	2.5
14455.p1	chr14	105847409	DEL	1,461	1
14455.p1	chr2	89995213	INS	812	9.3
14455.s1	chr3	47014725	DEL	147,102	0
14455.s1	chr7	142635453	DEL	158,289	0
14455.s1	chr9	66003864	DEL	905	0
14455.s1	chr21	8706195	DEL	442,533	9.8
14455.s1	chr19	8729966	INS	114,606	37.7

Table A.3: ***De novo* SVs identified by Bionano Genomics.** Summary of *de novo* structural variants detected by Bionano analysis of the proband and unaffected sibling generated from cell line DNA. Population frequency is determined by Bionano controls.

Mutation type	Affected child	Count
VNTR >50bp (SV INS)	14455.p1	2
	14455.s1	0
VNTR <50bp	14455.p1	1
	14455.s1	0
STR	14455.p1	4
	14455.s1	3
INDEL	14455.p1	5
	14455.s1	9
SNV	14455.p1	81
	14455.s1	97

Table A.4: **Distribution of DNMs by variant class.** The number of mutations identified in each category - variable number tandem repeat (VNTR) expansions >50bp (also represent structural variant insertions) and <50bp, short tandem repeat (STR) expansions, short indels <20 bp (indels), and SNVs.

Appendix B

CHAPTER 3 SUPPLEMENT

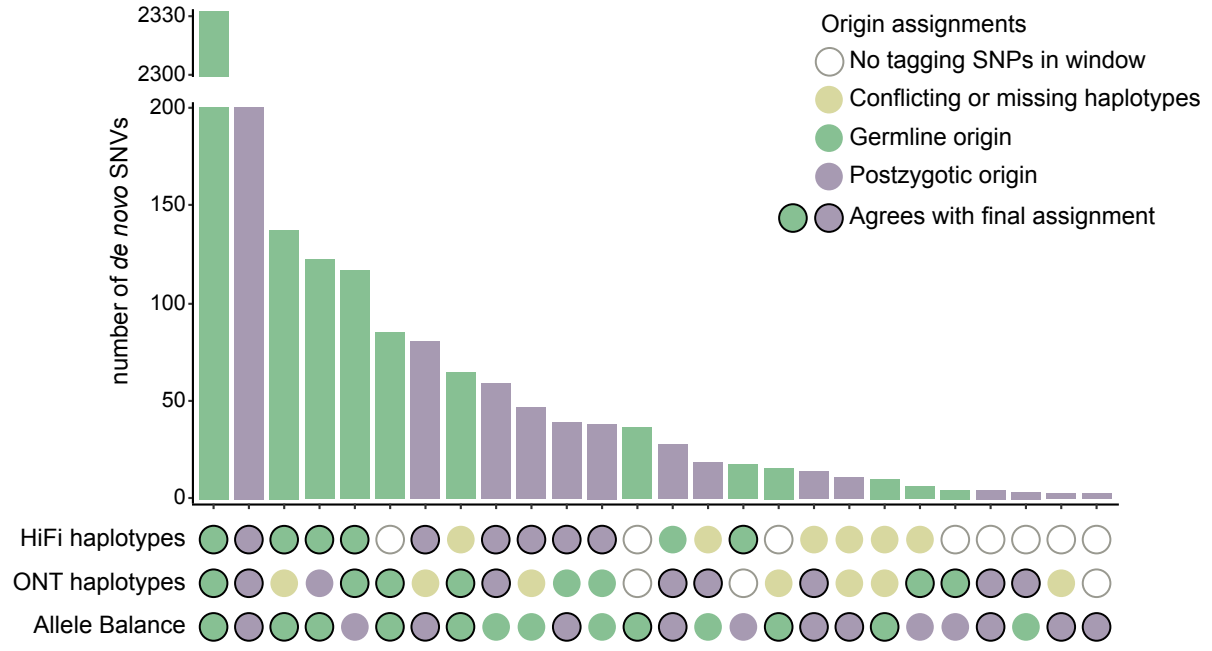


Figure B.1: **HiFi, ONT, and AB origin assignments.** Upset plot of origin assignment shows concordance between HiFi haplotypes, ONT haplotypes, and allele balance.

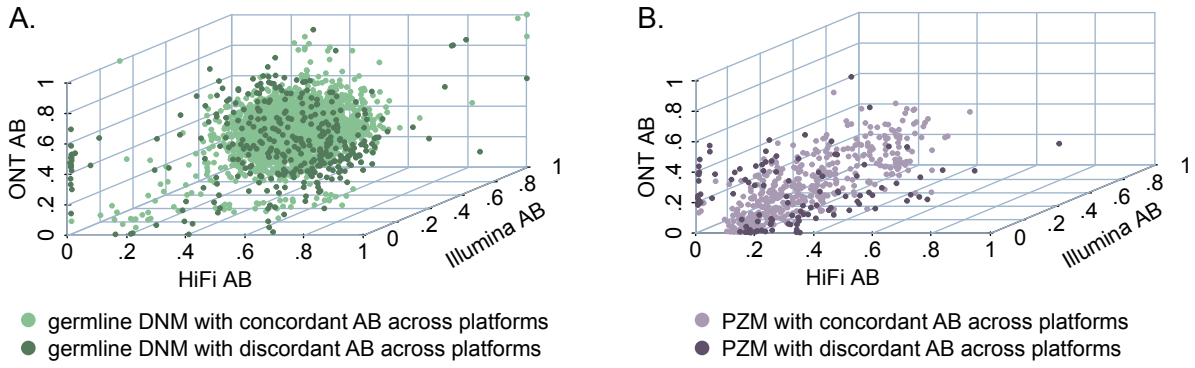


Figure B.2: **Allele balance scatter plots.** A. Allele balance (AB) of germline DNMs across all three platforms for variants with concordant and discordant AB by chi-squared shows that most variants have AB around 0.5. B. PZMs tend to have low AB across platforms.

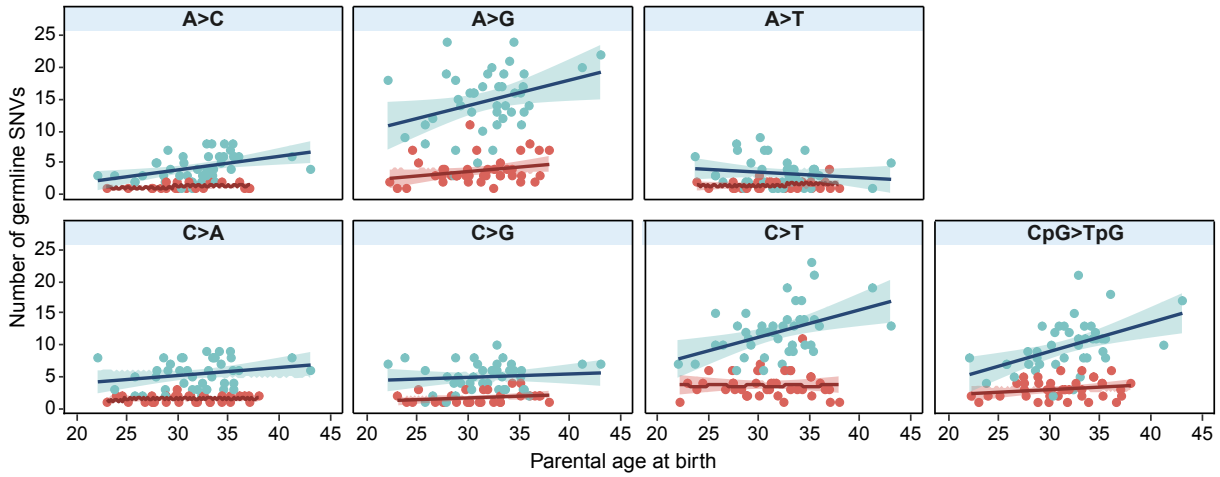


Figure B.3: **Parental age effect by mutation class.** The number of each type of substitution is not correlated with maternal age, but A>G, C>T, and CpG>TpG mutations become more prevalent as paternal age increases.

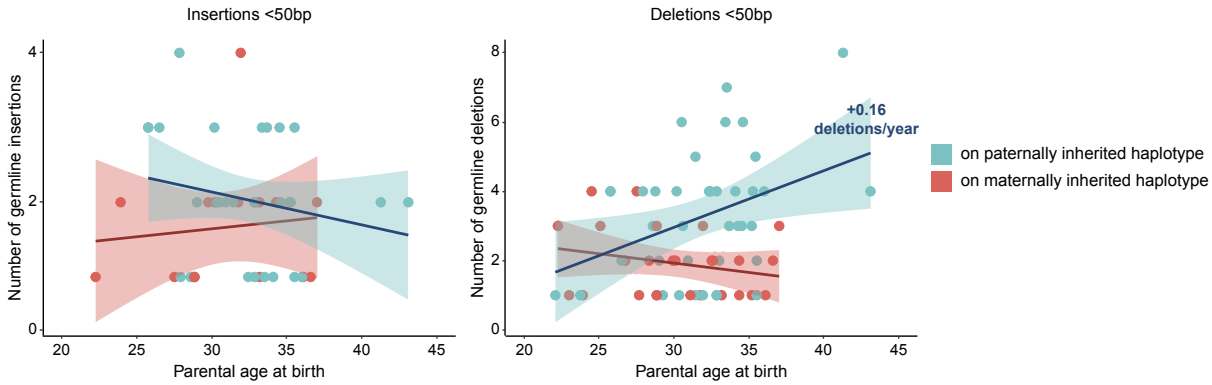


Figure B.4: **Indel parental age effect.** The number of insertions is not correlated with parental age, but deletions show a weak 0.16 increase with every year of paternal age.

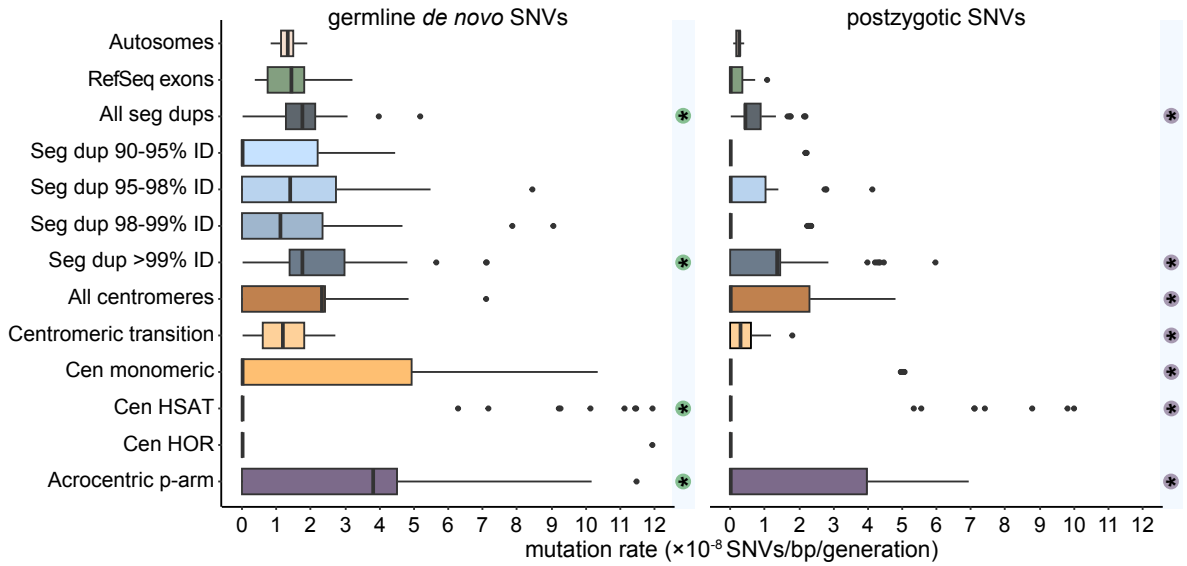


Figure B.5: **Mutation rates in SDs and centromeric regions.** The germline and postzygotic mutation rates are significantly enriched in the highest-identity SDs and acrocentric p-arms.

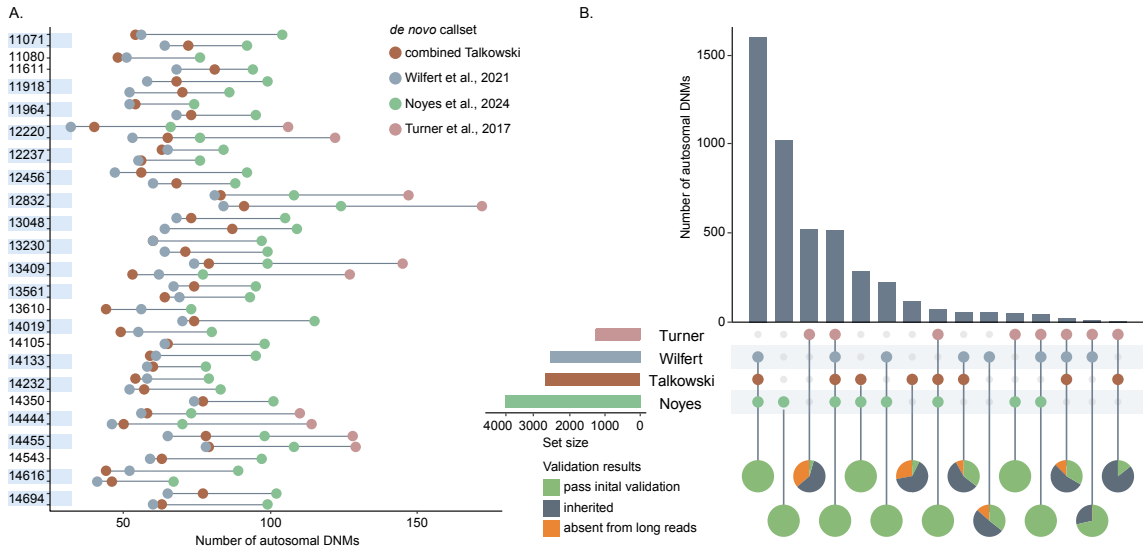


Figure B.6: **Comparison to other studies.** A. We identified more DNMs per sample than most previous Illumina-based studies [189, 4], with the exception of Turner et al. [176]. B. DNMs identified by multiple studies have the highest true positive rates, while DNMs exclusive to a single study tend to be false positives.

Appendix C

CHAPTER 4 SUPPLEMENT

C.1 Supplementary Note 1

Postzygotic SNV false positive rate

As a final filter for false positive postzygotic mutations (PZMs), we used tagging SNPs (tSNPs) to construct the haplotype on which a PZM arose. We first selected tSNPs unique to a parent to determine if a mutation arose on paternal or maternal DNA, and then we further refined our set of tSNPs, eliminating any SNP that was not heterozygous in the parent. For samples in G3, who have sequenced grandparents in our dataset, we further refined our tSNPs, until we had a set that was unique to the grandparent from whom the haplotype was inherited. Across 36 PZMs in G2 individuals, we assigned all but two to a unique parental haplotype. Across 93 PZMs in G3 individuals, we assigned 88 to unique grandparental haplotypes. For each sample in the direct lineage, we reexamined HiFi data, including every read aligned to a PZM coordinate, regardless of mapping or base quality. If a read matched the tSNP alleles surrounding the PZM coordinate, we determined it was on the same inherited haplotype. We counted the number of reads with the PZM alternate allele from both the inherited and other haplotype present in a sample, as well as on any unphased reads (Figure C.1). Finally, we determined that a PZM was an inherited event if a parent and grandparent had any reads with the alternate allele on the inherited haplotype, or if a parent or grandparent had more than one read with the alternate allele on the inherited haplotype. It is worth noting that we did not see any examples of the alternate allele on a different haplotype. In total, we found six PZMs that failed our filters, including three events from one G2 individual, NA12877.

In addition to looking at direct ancestors, we were able to evaluate whether the alternate

allele was transmitted to children for 62 PZMs across four samples (n=2 G2; n=2 G3). We determined that a PZM was transmitted if the alternate allele was present on at least one HiFi read in a child, for a total of 40 PZMs transmitted 120 times. For each transmitted variant, we expected to see the alternate allele on every read attributed to the inherited haplotype, which we saw in 87.5% of transmissions (n=105/120) (Figure C.1). However, there were 15 transmissions over 5 PZMs that were not present on every read from the parent. We also examined a variant's allele balance (AB) across HiFi, Illumina, and ONT (if available) reads in a transmitted child. We observe 29 transmissions (across 13 PZMs) where the AB was consistent across all data types and significantly different from 0.5. Of those 13 PZMs, 7 have other transmissions with AB of 0.5, leaving only 6 PZMs that deviate from expectation. We determined a PZM was a false positive event, likely caused by a recurrent sequencing error, if for every transmission, it was both incompletely linked to the inherited haplotype and had AB different from 0.5 across all sequencing platforms. In total, there were four PZMs that appear to be false positive events, including another two PZMs from NA12877.

We excluded all ten PZMs that failed to validate either in ancestors or descendants, resulting in a final callset of 109 PZMs. This final set includes 55 PZMs from samples with sequenced children, with a 60Between both validation strategies, there were another three PZMs that pass filtration thresholds but have inconsistent transmission profiles across children. These three PZMs may represent likely false positives in our dataset; given that we analyzed 55 PZMs for transmission, we can calculate a false positive rate (FPR) of 5.1%. Compared to the previous study of this family [162], there are no previously identified PZMs that we did not recover, yielding a false negative rate (FNR) of 0%. As such, we can revise our estimate of the PZM rate (μ) to be $\mu \times \frac{1-FPR}{1-FNR}$, or $2.23\times10^{-9} \times \frac{1-0.051}{1-0.0} = 1.94\times10^{-9}$.

Germline SNV false positive rate

For germline SNVs, we used Element data for the final round of validation. We determined that a DNM was false positive if it was not supported in one of four sequencing technologies

(HiFi, ONT, Illumina, or Element) and it had AB<0.1 in at least one other sequencing technology. For example, a variant with AB of 0.39 in HiFi, 0.05 in ONT, 0.2 in Illumina, and 0 in Element was considered a false positive event. In total, we observed eight such false positives out of 626 DNMs, for a FPR of 1.28%. Like for PZMs, we compared our callset to that of Sasani et al. [162]. After validating their variant calls with our pipeline, there were only four previously discovered DNMs that were absent from our callset. We can calculate our FNR as (false negative)/(true positive + false negative), giving an FNR of 0.64%. We can revise our estimate of the DNM rate (μ) to be $\mu \times \frac{1-FPR}{1-FNR}$, or $1.17\times10^{-8} \times \frac{1-0.0128}{1-0.0064} = 1.16\times10^{-8}$. We can combine these germline and postzygotic rates for an overall *de novo* SNV rate of 1.354×10^{-8} mutations per base pair per generation.

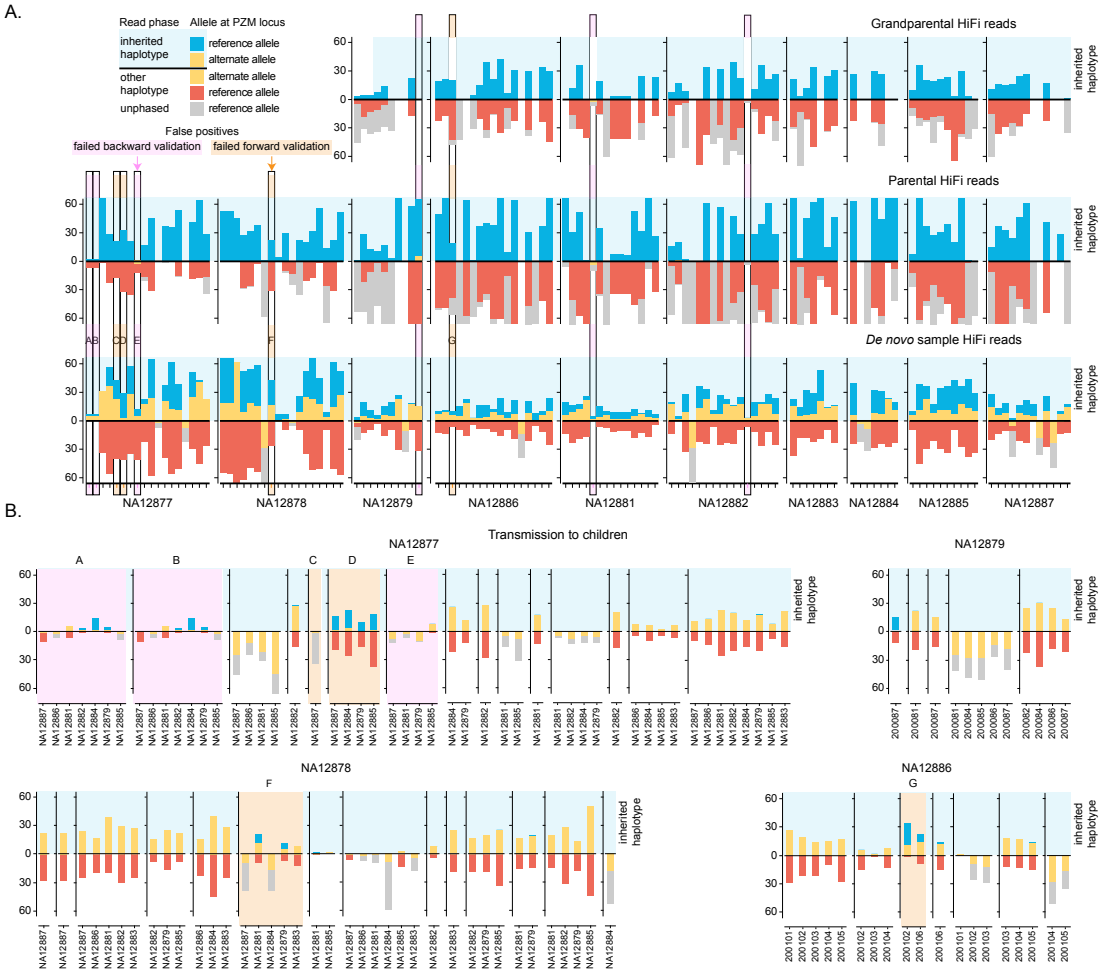


Figure C.1: **Phased haplotypes and allele counts.** A. Phased HiFi read counts for a *de novo* sample (bottom row), the parent from which they inherited the *de novo* haplotype (middle row), and the grandparent from which they inherited the haplotype (top row). Each column corresponds to a PZM, and missing read data indicates that a haplotype could not be uniquely assigned to a parent or grandparent. Reads assigned to the *de novo* haplotype are shown above the X-axis in blue, reads from the other haplotype are below the X-axis in red, and unphased reads are below the X-axis in grey. Reads with the alternate allele are shown in yellow. Variants in boxes are false positives - a pink highlight indicates that the variant failed backward validation by examining ancestors, and an orange highlight indicates that it failed forward validation by transmission. B. PZM transmissions to the next generation are shown. Transmissions are grouped by mutation, and one bar is shown for every transmission event.

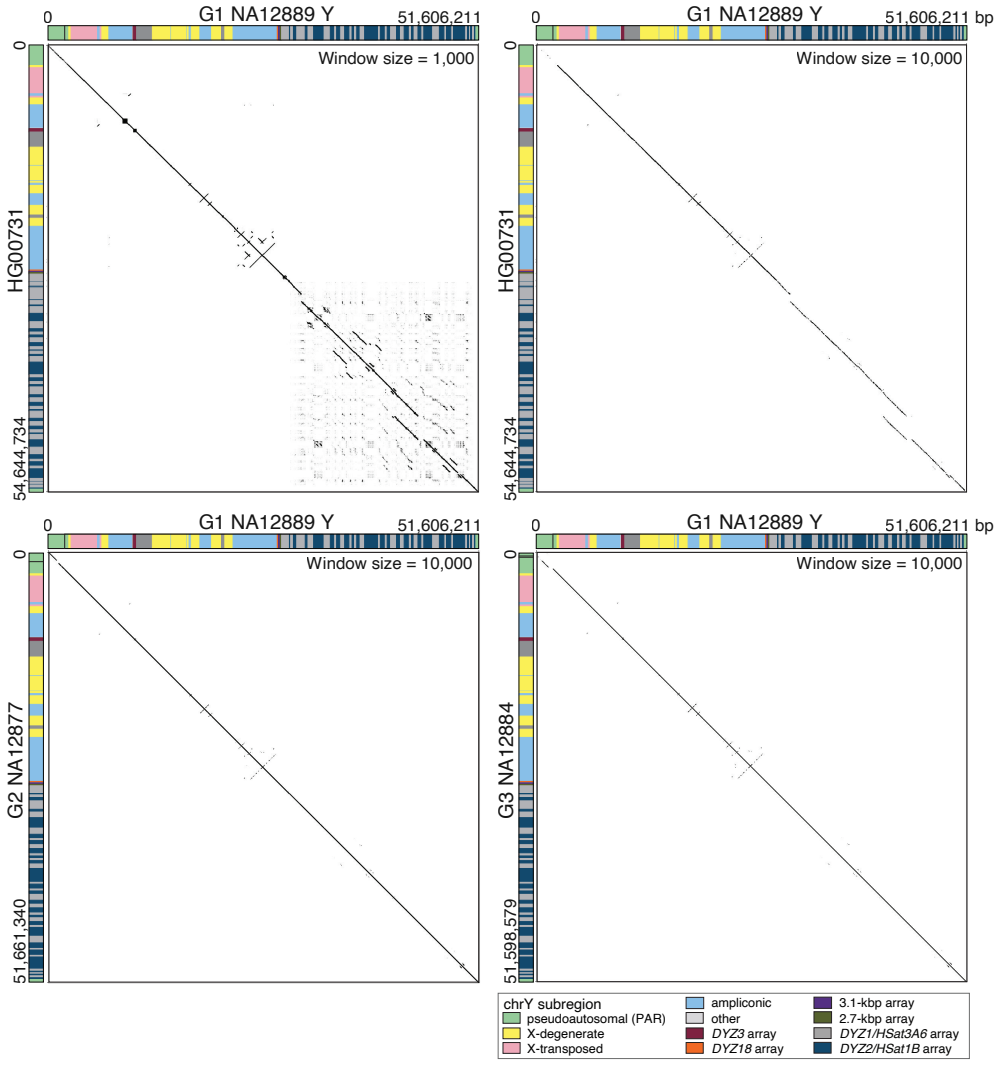


Figure C.2: **Comparison of chrY assemblies.** Dot plots of sequence similarity between the G1 NA12889 Y assembly, evolutionarily closely related Y chromosome from HG00731 (top), G2-NA12877 (bottom left), and G3-NA12884 (bottom right) assemblies. The HG00731 Y (R1b1a-Z225 Y haplogroup) last shared a common ancestor with the pedigree R1b1a-Z302 Y chromosome approximately 5,700 years ago, 95% HPD interval = 4,800–6,700 years ago, Figure 4.4A. Window sizes of 1,000 and 10,000 bp are shown for HG00731, and 10,000 bp for G2 and G3 males, indicating high levels of sequence similarity between the Y assemblies. Y-chromosomal sequence classes are shown as colored bars, with assembly breaks in the PAR1 indicated by black lines.

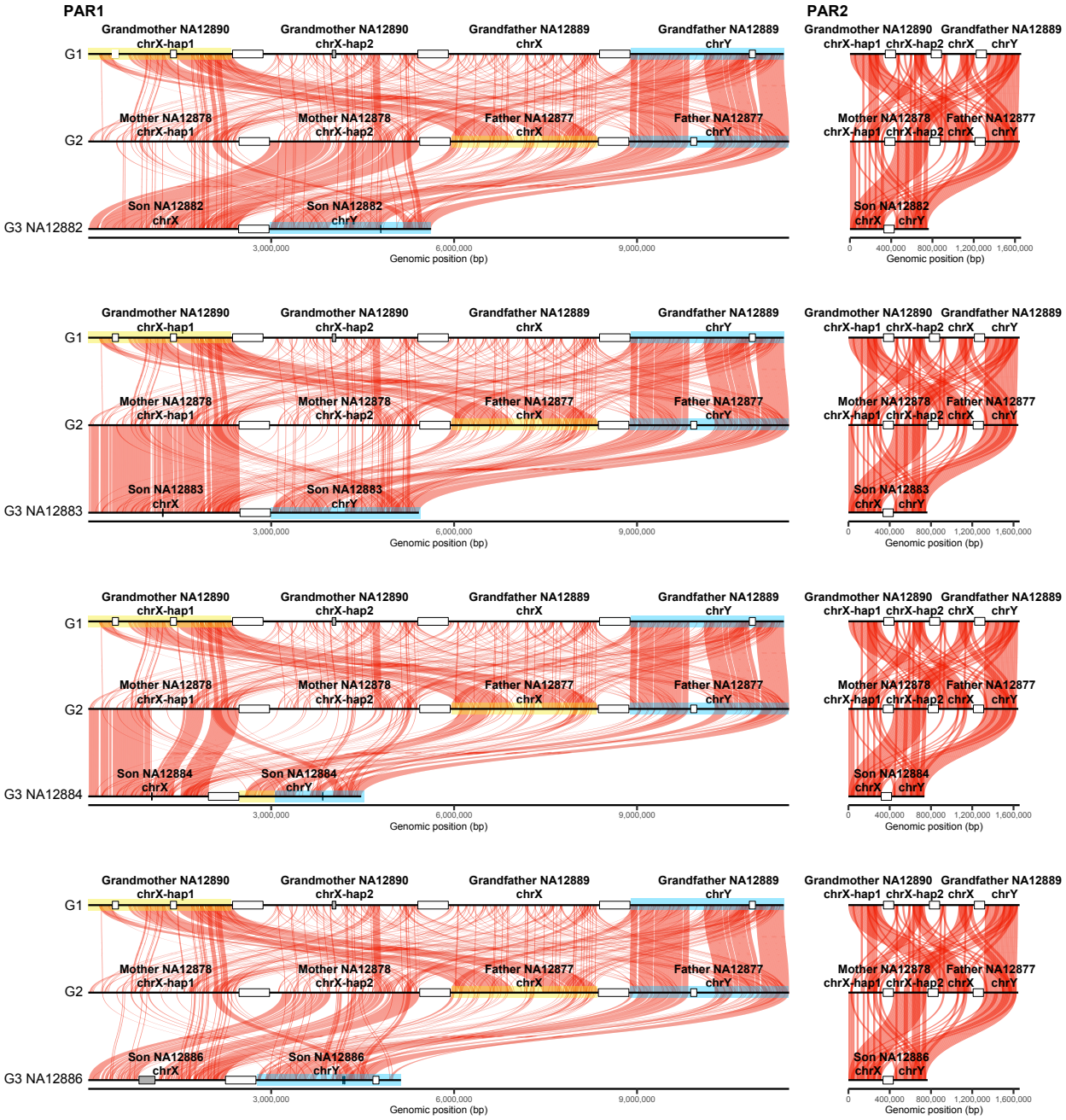

PAR1
PAR2
G1
G2
G3 NA12882
G3 NA12883
G3 NA12884
G3 NA12886
Grandmother NA12890 chrX-hap1
Grandmother NA12890 chrX-hap2
Grandfather NA12889 chrX
Grandfather NA12889 chrY
Mother NA12878 chrX-hap1
Mother NA12878 chrX-hap2
Father NA12877 chrX
Father NA12877 chrY
Son NA12882 chrX
Son NA12882 chrY
Son NA12883 chrX
Son NA12883 chrY
Son NA12884 chrX
Son NA12884 chrY
Son NA12886 chrX
Son NA12886 chrY
3,000,000
6,000,000
9,000,000
0
400,000
800,000
1,200,000
1,600,000
Genomic position (bp)

Figure C.3: **Assembled chrX and chrY pseudoautosomal regions (PAR) across three generations.** Binned alignments (10 kbp bin size) of ≥99.9% sequence identity are shown for PAR1 (left) and PAR2 (right) across generations. For G1 and G2, maternal sequences for chrX haplotypes (hap) 1 and 2, and paternal chrX and chrY PAR sequences are shown. For each of the G3 males, chrX and chrY PAR sequences are shown. Large white rectangles (equal in size to 500 kbp) separate the haplotypes, while 100 kbp-sized rectangles indicate where joints were made if several contigs represented the region for a specific individual. Gray rectangles indicate blocks of N's in the contigs. Yellow and blue rectangles indicate chrX and chrY PAR1 haplotypes that show evidence of recombination in G3 male NA12884. No recombination events were identified in PAR2. The following contigs were included for PAR1 (in the order as visualized from left to right): G1 NA12890 chrX haplotype 1 - haplotype1-0000079, haplotype1- 0000078 and haplotype1-0000010, chrX haplotype 2 - haplotype2-0000100; G1 NA12889 chrX - haplotype1-0000018 and chrY - haplotype2-0000082 and haplotype2-0000081; G2 NA12878 chrX haplotype 1 - mat-0000002, chrX haplotype 2 - pat-0000758; G2 NA12877 chrX - mat-0000005 and chrY - pat-0000406 and pat-0000383; G3 NA12882 chrX - mat-0000046 and chrY - pat-0000587; G3 NA12883 chrX - mat-0000038 and chrY - pat-0000576; G3 NA12884 chrX - mat-0000008 and chrY - pat-0000224; G3 NA12886 chrX - mat0000010 and chrY - pat-0000781 and pat-0-001035; and for PAR2: G1 NA12890 chrX haplotype 1 - haplotype1-0000010, chrX haplotype 2 - haplotype2-0000096, G1 NA12889 chrX - haplotype1-0000018, chrY - haplotype2-0000081; G2 NA12878 chrX haplotype 1 - mat- 0000002, chrX haplotype 2 - pat-0000758; G2 NA12877 chrX - mat-0000005, chrY - pat-0000383; G3 NA12882 chrX - mat-0000046 and chrY - pat-0000570; G3 NA12883 chrX - mat-0000045 and chrY - pat- 0000567; G3 NA12884 chrX - mat-0000008 and chrY - pat-0000224; G3 NA12886 chrX - mat-0000010 and chrY - pat-0000749.

VITA

Michelle Noyes was born in Southern California in 1997, and spent the first 13 years of her life in Rancho Cucamonga. She moved to Claremont, residing in the same house her father grew up in while she attended high school at the Vivian Webb School for girls (now simply known as the Webb Schools). In 2015, she headed off to the University of Chicago, where she worked in the lab of Dr. Marcus Kronforst, studying monarch butterfly evolution. She graduated with her B.S. in Biological Sciences with a specialization in genetics in 2018, and headed straight to the University of Washington to start her PhD. She joined the lab of Evan Eichler, where, as I'm sure you've gathered by now, she studied the *de novo* mutation rate in humans. After receiving her PhD, she hopes to take a nap, and then find a job that lets her stay in Seattle with her loved ones.