

# Diversity and consequences of structural variation in the human genome

Ryan L. Collins<sup>1,2,3,4</sup> & Michael E. Talkowski<sup>1,2,3,5</sup>  

## Abstract

The biomedical community is increasingly invested in capturing all genetic variants across human genomes, interpreting their functional consequences and translating these findings to the clinic. A crucial component of this endeavour is the discovery and characterization of structural variants (SVs), which are ubiquitous in the human population, heterogeneous in their mutational processes, key substrates for evolution and adaptation, and profound drivers of human disease. The recent emergence of new technologies and the remarkable scale of sequence-based population studies have begun to crystalize our understanding of SVs as a mutational class and their widespread influence across phenotypes. In this Review, we summarize recent discoveries and new insights into SVs in the human genome in terms of their mutational patterns, population genetics, functional consequences, and impact on human traits and disease. We conclude by outlining three frontiers to be explored by the field over the next decade.

## Sections

Introduction

Mutational properties of SVs

Population genetics of SVs

Functional consequences of SVs

The contribution of SVs to human diseases and traits

Conclusions and future perspectives

<sup>1</sup>Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. <sup>2</sup>Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>3</sup>Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>4</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>5</sup>Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. ✉e-mail: [talkowsk@broadinstitute.org](mailto:talkowsk@broadinstitute.org)

## Introduction

Human genetic variation is often categorized into three major classes based on the number of DNA nucleotides altered per variant: single-nucleotide variants (SNVs), short insertion or deletion variants (indels; 1–50 bp), and structural variants (SVs;  $\geq 50$  bp)<sup>1</sup>. Among these, SNVs and indels, collectively referred to here as short variants (<50 bp), have been extensively studied due to their high mutational densities in humans and the relative ease of their detection using a multitude of technologies<sup>1–4</sup>. By contrast, SVs have proven more difficult to interrogate due to several factors, foremost among which are the technical challenges of SV ascertainment<sup>5–7</sup>. Beyond these technical challenges, the density of SVs throughout each human genome is two orders of magnitude lower (tens of thousands) than that of SNVs and indels (millions)<sup>1</sup> while encompassing a much broader spectrum of mutational classes<sup>8–10</sup>. Consequently, our nascent capabilities to capture and interpret SVs in genetic research and clinical diagnostics remains substantially less precise than for short variants<sup>11</sup>.

Despite these limitations, the last 60 years in human genetic research have witnessed remarkable progress in defining the mutational spectrum of SVs. Since the initial discovery of the human chromosome number in 1956 and the discovery of trisomy 21 in Down syndrome in 1959 (refs. 12,13), progressive generations of technological breakthroughs have reliably yielded transformative new insights into SVs. The use of microscopic techniques, such as karyotyping or fluorescence in situ hybridization, revealed that chromosome-scale rearrangements were rare in the general population but enriched in individuals with developmental disorders and in cancer genomes<sup>14–18</sup>. In the early 2000s, the application of microarray technologies led to the discovery of the widespread existence of large DNA copy number variants (CNVs) throughout the human genome<sup>19–24</sup>. These observations were extended by shallow ( $\sim 4$ – $12\times$ ) whole-genome sequencing (WGS), which revealed that every human genome harbours thousands of CNVs and other types of SVs<sup>9,10,25,26</sup> and that these SVs generally adhered to the same principles of population genetics established for short variants<sup>26,27</sup>. More recently, deep ( $\geq 30$ -fold coverage) WGS of large populations has demonstrated that there are likely tens of millions of unique SVs segregating in the human population, most of which appear in just one or a handful of individuals worldwide<sup>28–31</sup>. Furthermore, the maturation of long-read sequencing and other technologies continues to reveal thousands of cryptic SVs embedded in the most complex and repetitive sequence contexts<sup>32–37</sup>.

Advancing our nascent understanding of the forces and factors influencing where new SVs arise in the human germline and how they are subject to natural selection requires large, population-scale characterization studies. Major initiatives have been launched to catalogue SVs in the human genome, such as the Genome Aggregation Database (gnomAD)<sup>28</sup>, 1000 Genomes Project and Human Genome Structural Variation Consortium (HGSVC)<sup>9,25,31</sup>, UK Biobank<sup>38</sup>, and the All of Us Research Program<sup>39</sup>, yet there remains a limited representation of SVs that exist across global populations, especially in non-European populations. Even within European populations, the most repetitive  $\sim 10\%$  of the genome remains intractable to most technologies. Recent breakthroughs using the combination of multiple new technologies and sophisticated analytic approaches that have achieved near complete telomere-to-telomere (T2T) assemblies are beginning to dissolve these technical barriers and are poised to enable the study of these genomic loci that were inaccessible to prior technologies<sup>40,41</sup>. Another major challenge in human genetics is the prediction of the functional consequences of SVs on gene function and *cis*-regulatory networks<sup>42,43</sup>.

Likewise, the impact of SVs has been underexplored in many human diseases and traits as SVs have not been routinely incorporated in most large-scale association studies to date. New technologies will capture a greater fraction of the SV spectrum<sup>34–36,44</sup>, but long-read assembly studies have been inadequate in terms of the number of individuals surveyed that is needed to derive principles of population genetics and human disease architecture. In summary, although some general patterns of SVs in the human genome have been defined, the specifics remain opaque across disparate domains of biology and medicine and represent opportunities for dramatic advances in the coming years.

In this Review, we summarize the existing knowledge of germline SVs in the human population. First, we discuss the mutational properties of SVs, including their diversity, complexity and mutation rates. Second, we describe the patterns of SVs in the global human population, with an emphasis on the role of natural selection. Third, we summarize the modes in which SVs can result in functional alterations to protein-coding and non-coding loci throughout the genome. Fourth, we review the roles of SVs in rare, common and complex human diseases, and speculate on the role of advancing technologies in this area of study. We do not focus on the technical aspects of detecting SVs<sup>6,7,45–47</sup>, the roles of somatic SVs in cancer<sup>48</sup> or the impact of SVs on three-dimensional genome organization<sup>43</sup> as these topics have all been comprehensively detailed in other recent reviews. Similarly, we do not discuss somatic or mosaic SVs in non-cancerous tissues, which is an active area of enquiry with insufficient high-resolution data across the complete SV spectrum to definitively contrast such SVs with germline variants in the human population<sup>49,50</sup>. We conclude by enumerating three frontiers in SV research that we anticipate will be topics of intense interest in the coming years.

## Mutational properties of SVs

The term ‘structural variation’ encompasses a remarkably diverse family of mutational classes, each of which has distinct properties (Fig. 1). In this section, we recap the current taxonomy for germline SVs, summarize their mutational mechanisms, and discuss germline SV mutation rates and covariates.

## Canonical and complex SV classes

By convention, germline SVs are categorized into several canonical mutational classes. Beyond chromosome-scale gains and losses of DNA, such as aneuploidies, the most commonly surveyed SV classes are deletions and duplications, collectively known as copy number variants (CNVs). These CNVs are defined by a loss or tandem gain of sequence relative to a reference genome. Canonical CNVs in the human germline have just one alternate allele documented in the population (that is, a biallelic variant)<sup>9,28</sup>, whereas multiallelic CNVs (mCNVs) in certain repetitive loci can reach high copy numbers, sometimes accruing dozens of duplicate copies in individual genomes<sup>30,51</sup>. Although mCNVs can be some of the most mutationally diverse SVs in humans, surveys to date from existing technologies suggest that they are relatively sparse throughout the genome – between 673 and 1,356 mCNV loci have been reported<sup>31,51</sup>, representing  $<0.05\%$  of all SVs reported to date across global population studies. Improved access to complex repetitive sequences from long-read technologies will upwardly revise these estimates as they are applied across large population data sets<sup>44,52</sup>. Insertions are the second-most common class of SVs in humans and comprise numerous subclasses such as mobile element insertions (for example, *LINE1* and *Alu*)<sup>53–55</sup>, novel non-reference insertions (for example, viral DNA insertions)<sup>56–58</sup>, nuclear insertions of mitochondrial DNA<sup>59</sup>, and

expansions of tandem repeats (for example, short tandem repeats (STRs; units of 1–10 bp) and variable number of tandem repeats (units of 10–100 bp))<sup>60,61</sup>. Although there is some ambiguity in these definitions, most classes of insertion SVs differ from tandem duplication CNVs either by not involving duplication of endogenous sequences or, if duplicated, not arising in tandem with their reference paralogue.

Multiple classes of SVs result in genome reorganization but do not involve a change in copy number. These ‘balanced’ SVs involve dosage-neutral derived alleles such as inversions, chromosomal translocations and some insertions. The largest classes of balanced SVs that involve alterations to large portions of chromosomes have been surveyed in cytogenetic studies for many decades<sup>14</sup>. More recently, new sequencing methods have enabled the localization of these events at nucleotide resolution and the prediction of their influence on developmental disorders and genome organization<sup>62–65</sup>. Intriguingly, despite their relative scarcity, balanced inversions seem to have an outsized impact on the chromosomal organization of individual haploid genomes. A recent study used a combination of optical mapping, long-reads and strand-specific sequencing to show that the average human haploid genome has 11.6 Mb of inverted genomic sequence, which is fourfold more than the number of nucleotides altered by all short variants and twofold more than CNVs and insertions<sup>36</sup>.

Advances in sequencing technologies and computational methods have facilitated the localization of the breakpoints of millions of human SVs to single-nucleotide resolution<sup>34–36,44,66,67</sup>. This improved resolution has led to the discovery that a minority of human SVs exhibit substantial rearrangement complexity, often in the form of CNVs or inversions co-occurring with one or more other breakpoints in the same mutational event<sup>9,26,38,39,62,66,68–72</sup>. This surprisingly frequent breakpoint complexity has led to the designation of a diverse class of non-canonical SVs commonly dubbed complex SVs, which includes any rearrangement of  $\geq 50$  bp involving two or more distinct genomic segments or canonical SV signatures that cannot be explained by a single end-joining or DNA exchange event<sup>9,28,73</sup>. Although complex SVs are collectively heterogeneous, they share three unifying trends. First, inversions are a common feature of complex SVs, and these complex inversions are frequently flanked by CNVs at one or both breakpoints, suggesting that inversions are one factor that can mediate complex SV formation<sup>9,72,74–77</sup>. Second, complex SVs are enriched in repetitive sequences, which may indicate that rearrangement mechanisms involving repetitive sequences are especially relevant in the creation of new complex SVs<sup>33,74,75</sup>. Third, the spectrum of rearrangement complexity is extensive; although virtually all (>99%) complex SVs characterized in the human germline to date are relatively simple – involving just two or three segments or breakpoints – dozens of extremely complex germline SVs have also been reported, collectively known as chromoanagenesis (which means ‘chromosome rebirth’ in Greek)<sup>66,77,78</sup>. These highly complex rearrangements typically involve dozens of breakpoints and two or more chromosomes interleaved in a single mutational event<sup>66,77–80</sup>. Such highly intricate genomic rearrangements can have catastrophic consequences; they were originally discovered in highly aberrant tumour genomes and are now recognized as a common feature of many cancers<sup>81–83</sup>. Germline chromoanagenesis is exceedingly rare but has been observed in children affected by severe developmental disorders<sup>64,77</sup>. The growth of population-based cohorts has revealed that these events also occur in the general population at extremely low frequencies (approximately <1:10,000 individuals), as evidenced by studies in gnomAD and the All of Us Research Program<sup>28,84</sup>. Long-read sequencing and similarly high-resolution genomic technologies will







































likely reveal even greater SV-associated complexity throughout the human genome in the years to come<sup>40,47</sup>.

## SV mechanisms and mutation rates

The mutational diversity of SVs is mirrored in the molecular mechanisms responsible for their creation<sup>85–87</sup>. In the human genome, SV mechanisms are inferred by the sequence and context of individual breakpoints and often vary by SV class<sup>85</sup>. The simplest SV breakpoints involve two blunt ends with no homology, which are typically formed by non-homologous end-joining (NHEJ) following a DNA double-stranded break<sup>85,88</sup>. NHEJ events often feature ‘scarring’ at the breakpoint, introduced by imperfect break repair in the form of small (<10 bp) deletions or non-templated insertions<sup>67,87,88</sup>. Other SV breakpoints exhibiting short (<70 bp) stretches of sequence homology (that is, ‘microhomology’) between the two break ends are commonly formed by replication fork-stalling and template-switching or microhomology-mediated break-induced replication<sup>71,89,90</sup>. Pairs of larger homologous sequences, usually hundreds or thousands of nucleotides long, can lead to non-allelic homologous recombination (NAHR) and produce CNVs, inversions or complex SVs depending on the orientation of the homologues involved<sup>86,87</sup>. Pairs of transposons, such as *Alu* or LINE1 elements, are particularly common substrates for homology-mediated SV formation<sup>86,91,92</sup>. Collectively, these four core mechanisms – NHEJ, fork-stalling and template-switching, microhomology-mediated break-induced replication, and NAHR – account for most CNVs discovered to date in the human genome<sup>25,67</sup>. However, not all classes of SVs are generated by these four mechanisms. For instance, mobile element insertions can be caused by transposition or retrotransposition of an endogenous mobile element<sup>53</sup>, whereas tandem repeat expansions are caused by DNA polymerase slippage during replication<sup>60</sup>. The mechanisms responsible for complex SVs are more diverse still, ranging from multistep mutational cascades involving inverted DNA repeats to chromatid missegregation into micronuclei during cell division<sup>72,93</sup>. Finally, certain mechanisms of SV formation might be specific to the most highly repetitive and hypermutable genomic loci, such as large (>100 kb) microsatellite tracts and DNA repeat arrays found near centromeres or on acrocentric chromosomal arms<sup>94</sup>, and therefore may be currently unknown but illuminated in the coming years by pangenome assemblies<sup>40,95</sup>.

The detailed understanding of many SV mechanisms belies our uncertainty about the rates at which these mechanisms act to generate de novo SVs in the human germline. Early microarray studies of large CNVs (usually >100 kb) in parent–child trios and unselected populations identified de novo CNVs in approximately 1–3% of individuals, with higher rates observed in children affected by developmental disorders<sup>22,96–98</sup>. Estimates of de novo SV rates have increased by an order of magnitude over the last decade due to the gradual adoption of WGS in human genetic research, which can capture most classes of SVs at base-pair resolution<sup>6,99</sup>. Short-read WGS studies have estimated a range of 0.11–0.29 de novo SVs per generation that are accessible to this technology when summed across all SV classes, with de novo CNVs and insertions appearing more frequently than other SVs<sup>9,28,100–103</sup>. However, most published estimates of SV mutation rates have not included variation in tandem repeats and other repeat-mediated sequences – among the most mutable of all human genetic elements – because they cannot be comprehensively surveyed by short-read WGS<sup>60,104</sup>. The inclusion of de novo tandem repeat mutations alone would dramatically inflate these published SV mutation rate estimates, as STRs mutate at a rate of  $10^{-3}$  de novo mutations per locus per generation and

# Review article

Nucleotides altered	Variant class	Abbreviation	Dosage change	Example subclass	Example alleles		Detectable by	Variants per genome		
					Reference	Alternate				
Short variants (<50 bp)	1 bp	Single-nucleotide variants	SNV		Transition, transversion	ATC TAG	→	AGC TCG	  	-4,000,000
	1–49 bp	Small insertions and deletions	InDel		-	ATCGT TAGCA	→	ATCACA TGTGTCA A--GT T--CA	  	-400,000
	1 bp to >10 kb	Tandem repeats	TR		STR, VNTR				   	-200,000
Structural variants (≥50 bp)	5–10 kb	Mobile element insertions	MEI		SINE, LINE, SVA, HERV				   	-2,000
	≥50 bp (Med. ~1 kb)	Copy number variants	CNV		Deletion, duplication, mCNV				   	-10,000 (~800 (>10 kb))
	≥50 bp (Med. ~5 kb)	Inversions	INV		-				  	-100
	≥50 bp (Med. ~10 kb)	Complex structural variants	CPX	 	delINVdel, INVdup, DUP-TRP/INV-DUP				    	-100
	≥5 Mb	Chromosomal abnormalities	CA	 	Reciprocal translocation, aneuploidy				    	-0.01

 Balanced     Unbalanced     Microarray     Cytogenetics     Optical mapping     Exome sequencing     Genome sequencing



**Fig. 1 | Human SVs span a broad mutational spectrum.** Genomic variants are usually divided into categories based on the number of nucleotides altered by the variant allele. Though the exact delineation of these categories is imprecise and varies based on semi-arbitrary thresholds, the field of human genetics has converged on a consensus of variants involving <50 nucleotides as ‘short’ variants (sometimes also referred to as sequence variants) and all other variation involving  $\geq 50$  nucleotides as structural variants (SVs)<sup>9</sup>. Short variants include just two distinct classes: single-nucleotide variants and short insertions or deletions (collectively known as indels). In total, there are approximately four million short variants present in the average human genome<sup>136,243</sup>. By contrast, SVs are comprised of a vastly more diverse family of mutational classes and subclasses, each with its own characteristic alternate allele structure and unique properties. A primary axis along which SVs can be further divided is whether their variant allele involves <50 bp total genomic gain or loss, which separates ‘balanced’ SVs, such as inversions, from ‘unbalanced’ SVs such as deletions,

duplications and large tandem repeats. Developing a unified taxonomy for all SVs has proven challenging in part due to the difficulty of ascertaining all SVs with a single technology or assay, as has been recently reviewed elsewhere<sup>6,45</sup>. In this figure, the technologies able to capture each SV class are indicated; lighter-shaded hexagons indicate technologies that do not reliably detect most SVs in the class using conventional methods. One universal trend is clear from the past two decades of human genomics research: there is a strong inverse relationship between variant size and abundance in the human population, with extreme cases of chromosome-scale abnormalities (that is, reciprocal translocations) being observed in <1% of individuals. delINVdel, paired-deletion inversion; DUP-TRP/INV-DUP, duplication-inverted triplication-duplication; HERV, human endogenous retroviruses; INVdup, inverted duplicated; LINE, long interspersed element; mCNV, multiallelic copy number variant; Med., median; SINE, short interspersed element; STR, short tandem repeats; SVA, SINE-variable number of tandem repeats–Alu; VNTR, variable number of tandem repeats.

therefore every human genome is expected to carry dozens of de novo STR mutations<sup>105</sup>. Resolving SVs in repetitive sequences outside of annotated tandem repeats<sup>6,45</sup>, which also have elevated SV mutation rates compared to the rest of the genome<sup>106,107</sup>, will further upwardly revise these mutation rates. Long-read sequencing technologies are ideal for interrogating these loci but have not yet been applied to large family-based cohorts to catalogue de novo SVs at scale. Thus, we expect that SV mutation rate estimates will continue to increase and be refined by variant class and genomic context in the coming years as T2T assemblies and other emerging technologies are more widely implemented. It is also likely that SV mutation rates will exhibit tissue-specific differences between the human germline and other healthy somatic tissues, which will require continued improvements in single-cell technologies to enable high-throughput profiling of SVs in the genomes of millions of individual cells<sup>108–110</sup>.

## Population genetics of SVs

As with other forms of genetic variation, SVs are ubiquitous in the global human population and adhere to most established Mendelian and population genetic principles. However, SVs differ markedly from other kinds of genetic variation in several ways, including their ability to form complex or unstable haplotypes and their tendency to undergo particularly strong natural selection. In this section, we highlight a few key patterns of SVs in individual genomes and in large human populations, with an emphasis on natural selection.

## Distributions in the global population

The average human genome harbours many thousands of SVs relative to the consensus reference sequence<sup>28,33</sup>, but the exact number and types of SVs identified per genome are heavily dependent on the technologies and algorithms used for SV detection<sup>6</sup>. (Table 1) In the last few years, the public release of several large short-read WGS studies has provided new insights into the landscape of SVs across human populations; these short-read studies have typically reported 9,000–13,000 SVs per genome on average<sup>28,29,31,39,111,112</sup>. By comparison, the three largest long-read WGS studies have been performed on smaller cohorts but with highly sensitive discovery power for SVs per genome, yielding 22,000–26,000 SVs per genome<sup>34,44,113</sup>. The vast majority of the differences across these studies and technologies are explained by small (<500 bp) SVs, insertions, tandem repeats and other variation in highly repetitive genomic contexts such as segmental duplications and satellites<sup>33,35,40,113,114</sup>. Notably, exciting new algorithmic developments

in graph-based genome assembly and haplotyping from long-read data have begun to close the gap in the disparity between short-read and long-read SV discovery. These methods enable imputation of SVs accessible to long-read WGS across much larger short-read WGS data sets, capturing approximately 18,000 SVs per short-read genome<sup>115,116</sup>. Insertions and CNVs comprise >90% of all SVs per genome irrespective of the technology used for SV detection, while recent studies using short-read and long-read WGS, optical mapping, and strand-specific sequencing have revealed that the average genome also harbours several hundred balanced and complex SVs<sup>28,33,74,75</sup>. Based on contemporary estimates from long-read WGS, the aggregate burden of SVs per genome ( $\sim 2 \times 10^4$ ) is approximately ~200-fold less than the average number of SNVs ( $\sim 4 \times 10^6$ ) and ~40-fold less than small indels ( $\sim 8 \times 10^5$ )<sup>35</sup>. Their sparse distribution notwithstanding, the imperative to incorporate SVs into human genetic studies is clear as they represent the predominant source of total nucleotide diversity between any two human genomes. Specifically, due to their large mutational footprints, SVs alter an average of 32.1 Mb per genome compared to the 6.7 Mb impacted by short variants<sup>31</sup>.

Most SVs segregating in the human population are rare (allele frequency <1%). More specifically, the two largest published short-read WGS SV catalogues indicate that approximately 49.6–51.2% of all SVs are small (<500 bp) and rare<sup>28,29</sup>. However, even the largest published sample sizes capture minuscule fractions (<0.01%) of the overall global population and have historically under-represented the most genetically diverse demographic groups such as populations in Africa. Therefore, existing catalogues of human SVs are largely incomplete due to the anticipated tens of millions of rare SVs present exclusively in genomes and populations that have yet to be sequenced, plus the myriad cryptic SVs embedded in highly repetitive sequence contexts that are inaccessible to population-scale short-read WGS<sup>35,37,115</sup>. On the other hand, a recent long-read WGS study of 15 individuals proposed that up to 97% of all major SV alleles (that is, SVs found on the majority of all human chromosomes) have already been discovered<sup>34</sup>. Perhaps counterintuitively, the majority (>95%) of SVs present in any one individual genome are common (allele frequency  $\geq 1\%$ ) polymorphic SVs. These principles of population genetics have been well described in prior studies; for example, the recent short-read WGS study of 3,622 individuals from the 1000 Genomes Project reported a total of 58,046 singleton SVs (that is, variants appearing as a heterozygote in just one individual) and 84,508 rare SVs compared to just 30,769 common SVs. The average individual genome correspondingly harboured just 29 singleton and

**Table 1 | The growth of population-scale SV catalogues**

Cohort	Year	Number of genomes	Total SVs	SVs per genome	Refs.
<b>Short-read WGS</b>					
UK Biobank	2023	490,640	1,926,132	13,102	111
All of Us	2022	97,940	1,506,805	9,686	39
gnomAD	2023	63,046	1,199,117	11,844	28,117
TOPMed	2023	138,134	355,667	Not reported	112
CCDG	2020	23,175	241,031	4,442	29
1000 Genomes Project	2022	3,202	173,366	9,679	31
<b>Long-read WGS</b>					
HGSVC	2024	65	188,500	26,115	113
Beyter et al.	2021	3,622	133,886	22,636	44
Audano et al.	2019	15	99,604	22,755	34

CCDG, Centers for Common Disease Genetics; gnomAD, Genome Aggregation Database; HGSVC, Human Genome Structural Variation Consortium; SV, structural variant; TOPMed, Trans-Omics for Precision Medicine; WGS, whole-genome sequencing.

173 rare SVs compared to 9,773 common SVs<sup>31</sup>. These trends are not unique to SVs; for example, the average genome in gnomAD carries 3.8 million total short variants, of which just 0.13 million (3.4%) are rare<sup>117</sup>. The precise allele frequency of common SVs can sometimes be difficult to estimate due to the technical challenges of SV detection in large populations that have precluded most studies from performing large-scale joint analyses of SVs<sup>6,45,46,99</sup>. Most conventional approaches require imperfect ‘clustering’ (that is, merging) of SVs observed in multiple individuals at the same locus and can incorrectly obscure multiple distinct SV alleles that arose through independent mutational events. With new genome representations, such as pangenome graphs and the precision of T2T assemblies<sup>32,118</sup>, these allele frequencies may be refined for common SVs, including those found in repetitive sequence regions that are inadequately modelled by conventional linear reference-based strategies. However, these regions remain a major technical challenge at scale and the sample sizes surveyed will remain the dominant force driving the new discovery for rare variants in population genetics and human disease studies<sup>45</sup>.

The principles of population genetics that govern the number, frequency and diversity of most SVs are well established. Population-scale SV studies using short-read and long-read WGS in large and diverse human populations<sup>44,52</sup> have shown that most SVs segregate stably on haplotypes within populations. Therefore, these SVs exhibit patterns of population stratification, Mendelian transmission, linkage disequilibrium and site-frequency spectra similar to those of short variants<sup>9,23,27–30,107,119,120</sup>. Likewise, within individual genomes, common SVs are almost always highly correlated with nearby short variants in linkage disequilibrium; for example, a landmark study in 2008 by the HapMap Consortium reported a perfect genotypic correlation ( $r^2 = 1.0$ ) for all polymorphic (allele frequency >5%), biallelic CNVs and at least one neighbouring SNV<sup>23</sup>. This general trend has been replicated many times since these early microarray-based studies<sup>9,23,25,28,121</sup>, although the adoption of population-scale sequencing has shown that linkage disequilibrium between SVs and short variants decays in repetitive sequence contexts due to both biological (for example, elevated SV mutation rates) and technical (for example, increased genotyping error)

factors<sup>28,51,121</sup>. This observation is especially true for mCNVs, whose wide distributions of copy numbers can break expected linkage disequilibrium patterns and produce ‘runaway haplotypes’ specific to individual populations that are not well tagged by short variants<sup>9,30,51</sup>.

## Natural selection

SVs have long been acknowledged to have a prominent role in evolutionary adaptation. In 1970, Susumu Ohno published *Evolution by Gene Duplication*, which claimed that duplications of individual genes and entire genomes are prominent evolutionary substrates because they bypass negative selection on deleterious coding variants through partial redundancy of paralogous duplicate genes<sup>122</sup>. Empirical support for the role of SVs in human adaptation has come from comparative genomics, which has shown that thousands of SVs are fixed in the human population and not shared with any of our closest evolutionary relatives such as chimpanzees<sup>123</sup>. Several of these human-specific SVs are thought to have contributed to the evolution of human-specific traits<sup>124</sup>. For example, a dispersed duplication of the *NOTCH2* locus in the ancestral great ape genome and subsequent gene conversion event created the human-specific *NOTCH2NL* gene family, which may have contributed to the expansion of the human neocortex<sup>125</sup>. Similarly, human-specific duplications of the *BOLA2* gene can partially explain differences in iron metabolism and erythropoiesis between modern humans and other species<sup>126</sup>. Even within the modern human population, a subset of polymorphic SVs have undergone recent positive selection for advantageous traits<sup>9,27,30</sup>. For example, deletions of hominin-specific exons in the haptoglobin gene (*HP*) have lowered blood cholesterol levels<sup>127</sup>, multiallelic CNVs at the salivary amylase (*AMY1*) locus may have aided in adapting to starch-heavy diets during the transition from hunter-gatherer to agrarian societies<sup>128,129</sup>, and complex inversion SVs at the *KANSL1* locus have increased reproductive fecundity<sup>130,131</sup>. Despite these examples, the systematic identification of adaptive SVs has been impeded by the lack of large, ancestrally diverse cohorts with deep phenotyping and genetic data. Thus, we anticipate that many more adaptive SVs will be identified over the coming years as national biobanks tied to electronic health records and diverse sequencing initiatives continue to mature.

Although a handful of prominent examples demonstrate how positive selection can act on SVs, most SVs instead experience negative selection in the human population (Fig. 2 and Supplementary Information)<sup>22,96,107</sup>. Two key observations underpin this conclusion: first, most SVs appear at low allele frequencies in the general population<sup>9,28,29</sup>; second, this excess of rare SVs cannot be explained by the slow rate at which new SVs arise de novo in the human germline nor by genetic drift based on known human demographic history<sup>9,22,27,107</sup>. Multiple covariates influence the strength of selection on SVs. First, while the average SV typically experiences relatively weak negative selection, the number of nucleotides rearranged per SV seems to be a critical factor in this relationship because SV size and allele frequency are inversely correlated, with larger SVs appearing at lower allele frequencies in the population<sup>28,29,121</sup>. For example, 73.0% of all SVs  $\geq 1$  Mb catalogued in gnomAD are singletons compared to just 44.1% of all small (<200 bp) SVs<sup>28</sup>. Large SVs are also enriched in severe diseases<sup>14,132–134</sup> and are less likely than smaller SVs to arise de novo in healthy children<sup>96,97,103</sup>. Second, although the strength of negative selection is proportional to SV size across all classes of SVs<sup>28</sup>, not all classes experience the same average magnitude of selection. For instance, deletions seem to be under stronger negative selection than duplications on average, given that duplications tend to be larger<sup>9</sup>, have higher allele frequencies<sup>27,28</sup>,

have smaller effect sizes in association studies of human traits<sup>135</sup>, and are 2.3-fold more abundant over human evolutionary history than deletions<sup>123</sup>. Third, complex SVs typically occur at lower allele frequencies than canonical SVs of similar sizes, indicating that rearrangement complexity may also contribute to negative selection on SVs<sup>28,77</sup>. These differences between SV classes are at least partially attributable to their relative potential to disrupt protein-coding genes<sup>42</sup> because, across all SV classes, gene-disruptive SVs are under dramatically stronger negative selection than non-gene-disruptive SVs from the same SV class<sup>24,28,54</sup>. Interestingly, SVs and short variants exhibit a similar excess proportion of singletons when restricted to variants predicted to result in loss-of-function of protein-coding genes (13–16% more singletons than expected in the absence of any selection)<sup>28,136</sup>, implying that the functional consequences of a variant are a stronger determinant of negative selection than its mutational class or type.

## Functional consequences of SVs

Due to their size and mutational diversity, SVs can cause a broader range of functional consequences than other categories of genetic variation<sup>42</sup>. SVs can alter genome function by directly disrupting coding genes but also by disrupting *cis*-regulatory elements (CREs), regulatory networks and genome organization<sup>42,43,65</sup>. By virtue of their size and properties, SVs also have the unique propensity to disrupt multiple genes or CREs in a single mutational event. Below, we summarize the various ways in which SVs can affect genome function and explain how these functional consequences can reveal mutationally intolerant genes and other biologically important loci.

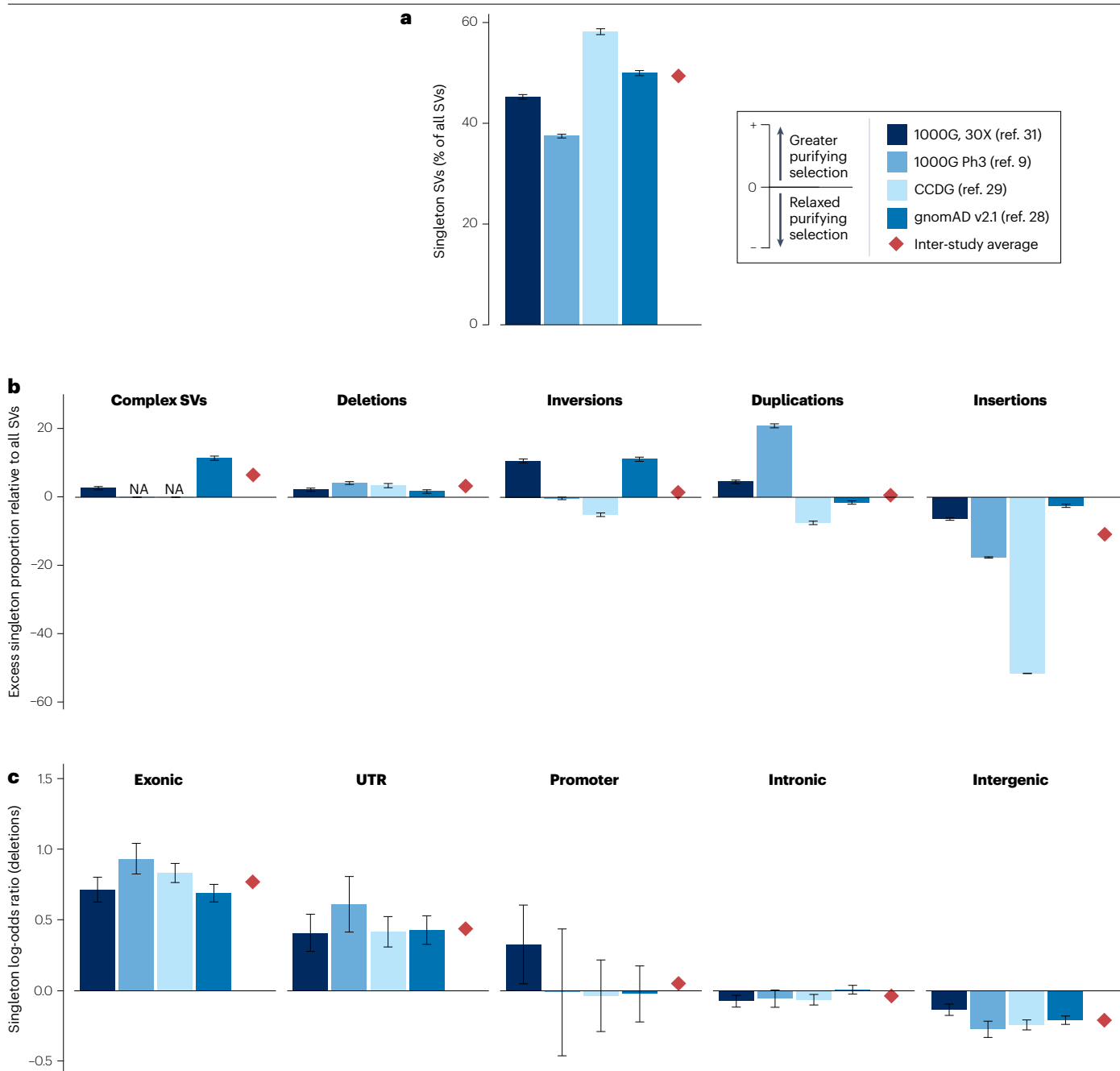
## Spectrum of coding and *cis*-regulatory effects

SVs can alter genome function most directly by disrupting one or more genes. SVs can disrupt genes through context-specific mechanisms, the simplest of which is by deleting or duplicating one copy of an entire gene, which leads to loss-of-function (LoF) or whole-gene copy-gain, respectively. A recent short-read WGS-based estimate predicted that SVs collectively contribute 25–29% of all rare LoF events per genome<sup>28</sup>, which is likely an underestimate given that many small SVs and at least 395 genes with potential biomedical relevance remain incompletely mapped by short-read WGS but are becoming accessible to long-read technologies<sup>6,34,41,45</sup>. Genic SVs often have profound influences on transcription and RNA processing. Simple changes in gene copy number usually correspond to linear changes in RNA expression levels as do LoF SVs other than whole-gene CNVs<sup>137–140</sup>. Partial-gene CNVs adhere to this same trend on average, although the transcriptional effects of partial-gene duplications are more variable and can sometimes lead to decreased – rather than increased – RNA expression levels due to nonsense-mediated decay<sup>137</sup>. SVs can cause myriad coding effects other than LoF and copy-gain, such as gene fusions<sup>141</sup>, polypeptide repeats produced by STRs<sup>142</sup>, and the creation of novel exons<sup>143</sup>, or can cause abnormal mRNA processing without altering total expression levels, such as by disrupting mRNA splicing<sup>139</sup>. Mobile element insertions seem particularly prone to altering mRNA splicing, as many families of mobile elements encode transcription factor-binding sites and can create novel splice sites, cause exon skipping, or lead to the inclusion of cryptic exons<sup>143–145</sup>. Recent discoveries have also revealed that mobile elements might indirectly influence gene expression by scaffolding interactions between CREs and their target gene promoters<sup>146</sup>. Complex SVs can have remarkable transcriptional consequences as they can simultaneously rearrange multiple genes to produce chimaeric transcripts in addition to LoF, copy-gain and other conventional effects<sup>147,148</sup>.

The genomic rearrangements introduced by SVs are also a major source of non-coding regulatory variation. Landmark advances in mapping CREs across human tissues and cell types<sup>149</sup> have enabled the prioritization of potentially functional non-coding SVs, which has revealed that SVs are broadly but weakly depleted over most CREs in the general population<sup>28,29,100,137,150</sup>, and a subset of CRE-disruptive SVs likely contribute to risk for a range of diseases<sup>151–155</sup>. Linking non-coding SVs to their affected genes is an unsolved challenge in most cases but recent studies using paired short-read WGS and transcriptome sequencing from matched donors, such as the Genotype-Tissue Expression (GTEx) Project, have made inroads into this problem<sup>137–139,156</sup>. The most recent GTEx study of SVs identified 1,271 common SVs (72% of which were non-coding) that were significantly correlated with the RNA expression levels of neighbouring genes, also known as expression quantitative trait loci (eQTLs)<sup>157</sup>. Similar observations have also been reported by SV eQTL studies in cohorts other than GTEx<sup>156</sup>. Collectively, three main patterns have emerged from human SV eQTL studies to date. First, the average common SV is 2.6–10.8 times more likely than the average short variant to be a significant eQTL for at least one gene; for example, one study reported that 4.8% of all common SVs were the ‘lead’ eQTL (that is, exhibiting the strongest association among all common variants) for at least one gene compared to just 1.9% of common SNVs and 2.0% of common indels<sup>157</sup>. Furthermore, these enrichments were highly non-uniform across SV types, ranging from 2.0% of common mobile element insertions to 15.0% of common duplications in the GTEx data set<sup>157</sup>. Second, SVs are more capable than short variants of acting over long distances to influence gene expression, with recent studies reporting that 22.6–40.5% of all SV eQTLs are associated with genes  $\geq 250$  kb away compared to just 16.4% of SNVs and 16.7% of indels<sup>156</sup>. Third, SV length influences the strength of its effects on gene expression levels in *cis*<sup>138,139,156</sup>; for example, the lead eQTL effect sizes of large ( $\geq 50$  kb) SVs were 3.0-fold greater than those of lead short variants and 1.6-fold greater than small (<500 bp) SVs in GTEx<sup>157</sup>. However, these same eQTL studies have also estimated that SVs are the causal variant for a small minority (3.5–7.2%) of all eQTLs and explain just 8.4% of the total genetic variance in RNA expression levels<sup>34,138</sup> owing to the smaller number of SVs in the human genome and their linkage disequilibrium with many short variants. To date, no similarly powered studies have been conducted using long-read WGS paired with gene expression; thus, the contribution of SVs in the most highly repetitive ~10% of the genome to gene expression remains relatively unknown. Finally, while many lines of evidence now underscore the strong impact of SVs on gene expression and *cis*-regulation, the mechanisms underpinning these effects are still being determined and have been recently reviewed elsewhere in detail<sup>43</sup>.

## Dosage sensitivity and genic constraint against SVs

Many SVs alter gene function but not all genes are equally sensitive to such disruptions. There is a continuum of relative intolerance to – or ‘constraint’ against – LoF variation<sup>136</sup>, and a special form of mutational constraint, known as dosage sensitivity, is especially relevant to CNVs and other unbalanced SVs. Dosage sensitivity describes the relationship between copy number and fitness for a gene or locus, which includes not only haploinsufficiency (intolerance to decreased copy number) caused by LoF SVs but also triplosensitivity (intolerance to increased copy number) caused by SVs that result in copy-gain. Hundreds of individual genes are known to be dosage sensitive in the context of human disease<sup>158</sup>, virtually all of which are annotated as haploinsufficient. For example, the Clinical Genome Consortium Dosage Sensitivity Map, one



of the leading resources for the diagnostic interpretation of CNVs, currently includes a total of 374 genes with 'sufficient evidence' for pathogenic dosage sensitivity, just three of which (*APP*, *LMNB1* and *PLP1*) are known to be triplosensitive<sup>11</sup>. This paucity of recognized triplosensitive genes is attributable to the formidable challenges of interpreting copy-gain duplications, including duplications having weaker average effects than deletions<sup>135</sup>, their duplicate copies being able to arise in a variety of orientations relative to the endogenous locus or being translocated to entirely different chromosomes, and their transcriptional consequences often being ambiguous and hard to predict in silico<sup>137,138</sup>. A handful of genic dosage sensitivity metrics and models have been proposed over the last decade to fill this void<sup>29,120,159-161</sup>, beginning with a

seminal study in 2010 that produced some of the first gene-level probabilities of haploinsufficiency for a majority ( $n = 12,443$ ) of all human genes<sup>161</sup>. More recently, a meta-analysis of rare CNVs detected by microarrays in nearly one million individuals enabled new metrics reflecting both haploinsufficiency (pHaplo) and triplosensitivity (pTriplo) for essentially all ( $n = 18,641$ ) autosomal protein-coding genes<sup>160</sup>. Owing to the comparatively large sample size, these pHaplo and pTriplo metrics have been applied to define high-confidence sets of 2,997 predicted haploinsufficient genes and predict a dramatic increase to 1,557 triplosensitive genes wherein rare copy-gain CNVs experience a magnitude of negative selection roughly equal to gene truncation by short variants in established LoF-constrained genes in gnomAD<sup>136</sup>. While these lists of



## Fig. 2 | Most SVs experience purifying selection in the general population.

Over the past decade, population-scale sequencing studies of structural variants (SVs) have unanimously concluded that most human SVs are held at low allele frequencies owing to purifying selection<sup>9,28,29,31</sup>. **a**, Given that the apparent SV mutation rate in humans is extremely low (fewer than one newly arising de novo SV per genome on average)<sup>103</sup>, a simple approach for estimating the strength of negative selection on SVs is to calculate the fraction of all SVs in a population that appears as ‘singletons’ observed as a heterozygous genotype in one individual out of the whole population. This singleton proportion will vary between studies and different classes of SVs owing to both technical and biological factors, but most large-scale population-based SV sequencing studies ( $n > 2,500$  individuals) have reported that roughly half of all SVs are observed as singletons<sup>9,28,29,31</sup>. Here, singleton proportion estimates are provided from four prominent population SV studies that employed short-read genome sequencing (bars with 95% confidence intervals) as well as an inverse-variance weighted meta-analysis of the four studies (red diamond). **b**, Purifying selection acts more strongly against complex SVs and deletions than on other forms of SV, including inversions, duplications

and insertions. This trend can be clearly observed when comparing the singleton proportions for a single class of SV to all other SVs not belonging to that class. Positive values indicate a greater proportion of singletons for a given class of SVs relative to the totality of all classes of SVs, which is evidence for negative selection against SVs of this class. **c**, The strength of selection also varies by the genomic context and predicted consequences of each SV. For example, when focusing exclusively on deletions for interpretability, all four population sequencing studies find elevated rates of singleton deletions overlapping protein-coding exons and untranslated regions (UTRs) as compared to all deletion SVs. Notably, SVs that impact coding-proximal sequences, such as promoters and introns, also exhibit slightly elevated rates of singleton variants compared to strictly intergenic SVs, suggesting a mild negative selection weaker than for deletions of protein-coding exons. More information on the data sets and code used to generate these plots can be found in the Supplementary Information. CCDG, Centers for Common Disease Genetics; gnomAD, Genome Aggregation Database; NA, not available.

predicted dosage-sensitive genes represent promising starting points for prioritizing likely triplosensitive genes for follow-up studies, it is important to underscore that – as with all predictive computational algorithms in clinical genomics – expert manual curation and clinical assessments will be required before implicating any of these prioritized triplosensitive genes in specific disease aetiologies at the level of confidence necessary for diagnostic screens and medical practice.

The gradual improvement of genic dosage sensitivity models and metrics over the last decade has progressively unlocked more insights into the biology underpinning dosage sensitivity. First among these insights was the intriguing observation that both deletions and duplications are depleted in genes that are also intolerant to protein-truncating short variants<sup>28,100,120</sup>. This finding implied that most human genes do not have independent sensitivities specific to increased or decreased gene dosage but instead exhibit a general intolerance to all variations altering gene expression or function. However, while this trend emerges on average across all human genes, there are hundreds of exceptions where a gene is predicted to be either triplosensitive or haploinsufficient but not bidirectionally dosage sensitive. For example, the pHaplo and pTriplo metrics exhibit a strong positive correlation ( $R = 0.55$ ) across all genes en masse but have also been leveraged to define initial sets of 63 and 111 genes that exceed high-confidence thresholds for haploinsufficiency or triplosensitivity, respectively, with virtually no evidence for sensitivity to the reciprocal copy number change<sup>160</sup>. These metrics have also enabled early data-driven insights into the features that distinguish haploinsufficiency from triplosensitivity, which have highlighted gene size, expression levels and *cis*-regulatory complexity as some of the factors partially responsible for determining the sensitivity of a gene to increased versus decreased dosage. Despite this progress, deeper exploration of these patterns in larger data sets and more sophisticated computational models paired with experimental validation will be necessary to clarify the specific aspects responsible for genic dosage sensitivity beyond the broad, coarse trends that have emerged from recent studies. Similarly, population-scale applications of long-read WGS will be essential to understanding the dosage sensitivity of the several hundred protein-coding genes embedded in highly repetitive genomic loci, which are not detectable by microarrays and short-read WGS and are therefore absent from virtually all existing data sets. As large-scale reference resources, such as All of Us and gnomAD, begin to aggregate large-scale long-read data sets, insights into how natural selection acts on gene dosage changes induced by SVs will become increasingly accessible.

## The contribution of SVs to human diseases and traits

The varying roles of SVs in human disease have been studied for over sixty years<sup>162</sup>. Starting with the discovery of trisomy 21 as the cause of Down syndrome in 1959 (ref. 12), SVs have been increasingly recognized as important contributors to the aetiologies of countless human diseases, ranging from common and complex diseases to severe Mendelian disorders. Correspondingly, SVs have become prominent targets of diagnostic screens for specific phenotypes and scenarios in clinical practice. The adoption of sequencing in disease association and clinical genetics has led to a surge of new insights into SVs in disease over the last decade. However, the medical relevance of SVs in the most repetitive ~10% of the genome remains tantalizingly underexplored due to technical limitations of conventional technologies<sup>40,41</sup>, leaving our understanding of SVs in disease far from complete. In this section, we review current knowledge of SVs in common and rare diseases as well as the utility of ascertaining SVs in clinical diagnostics.

### Common and complex diseases

Most prior research in the genetics of common and complex diseases has focused on short variants, especially in the context of conventional genome-wide association studies (GWAS) that do not typically include SVs. The reasons for this disparity are largely technical<sup>163</sup>: most GWAS have relied on microarrays or exome sequencing, both of which are imperfect modalities for accurately discovering and genotyping common CNVs and are effectively unable to detect other classes of SVs<sup>164</sup>. Nevertheless, the comparatively few GWAS that have systematically discovered and genotyped CNVs have successfully identified hundreds of CNVs provisionally associated with common and complex diseases<sup>159,160,163,165–169</sup>. Similarly, population-based studies of unselected individuals have indirectly linked thousands of SVs to disease based on strong linkage disequilibrium between each SV and short variants at previously reported GWAS loci<sup>9,28,138,156,170</sup>. These SVs in linkage disequilibrium with GWAS loci are 1.6–1.9 times more likely than other SVs to disrupt genes or annotated CREs and are up to threefold enriched for certain SV classes such as large deletions, mobile element insertions and STRs<sup>9,28,156,171</sup>. Furthermore, SVs that affect the expression of multiple genes seem especially likely to be linked to GWAS loci: a recent WGS study found that >40% of SVs that were eQTLs for two or more genes were also in linkage disequilibrium with at least one GWAS locus, as compared to just 20% for single-gene eQTL SVs and

≤10% for SVs that were not eQTLs<sup>156</sup>. In a handful of rare examples, SVs in linkage disequilibrium with common disease GWAS loci have been meticulously dissected to unearth remarkable new biological insights, as is the case for the complement component 4 (*C4*) locus. Multiallelic complex SVs at the *C4* locus have been associated with increased risk for schizophrenia<sup>140</sup> but these same SV alleles simultaneously confer protection against multiple autoimmune diseases such as systemic lupus erythematosus<sup>172</sup>. Astonishingly, these complex SVs at *C4* act in a sex-dependent manner, which appears to contribute to the sex biases observed in these diseases – schizophrenia is more common in men than in women, whereas systemic lupus erythematosus is more prevalent in women<sup>172</sup>. Despite exceptional examples such as *C4*, common SVs linked to GWAS loci are predicted to be the causal variant for just 3.2–14.2% of all GWAS loci<sup>138</sup>. Thus, while SVs play a significant role in the genetic architecture of common and complex diseases and can expose previously unknown aspects of disease biology, the contributions of SVs are likely modest in most diseases, and SVs are unlikely to explain the many thousands of GWAS loci with no known causal variant.

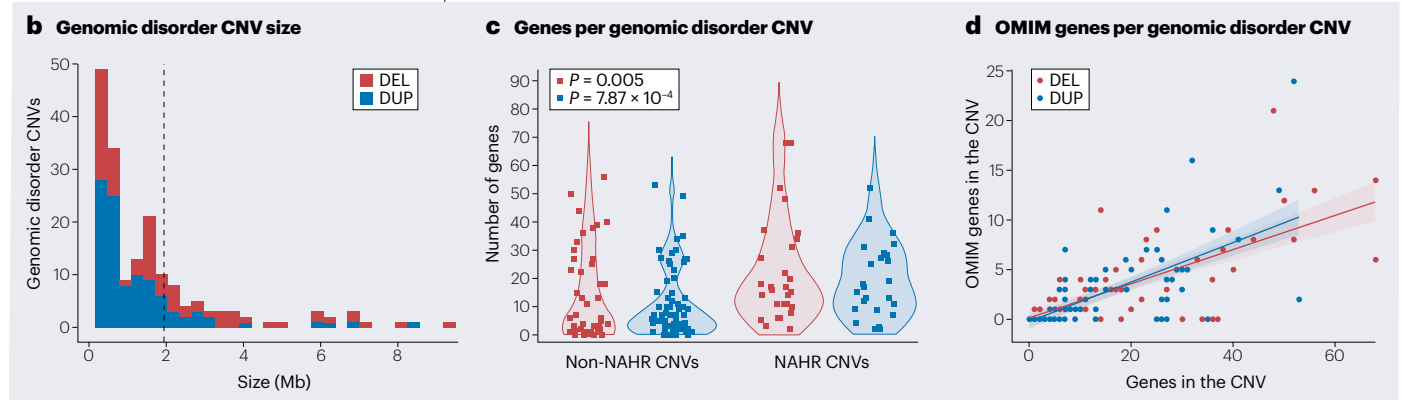
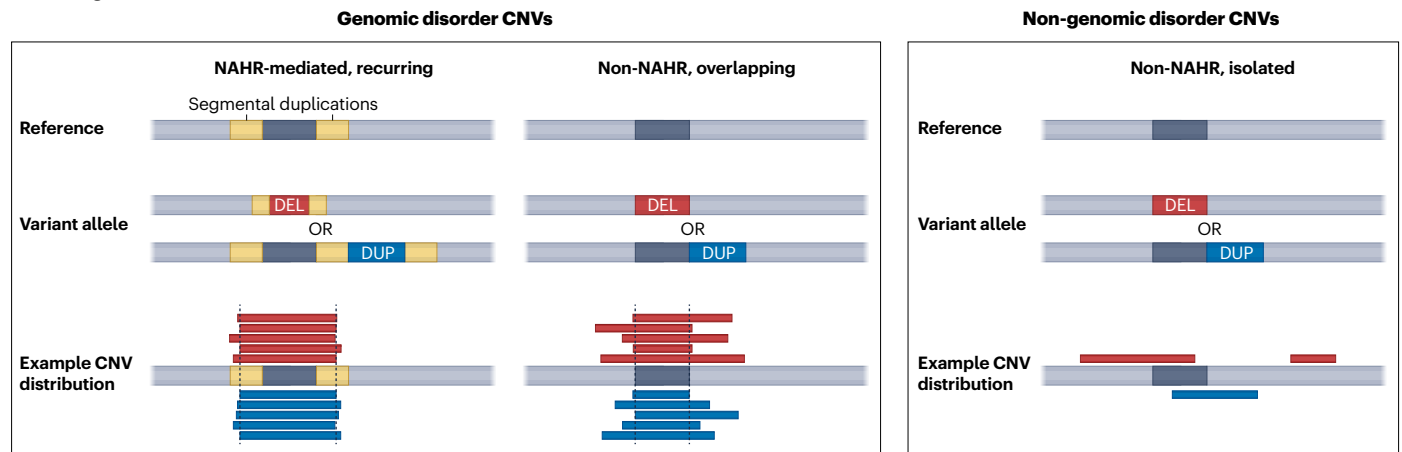
## Mendelian, developmental and genomic disorders

In contrast to common diseases, SVs have been prominent throughout the Mendelian disease literature for decades. SVs are perhaps best understood in the context of developmental disorders<sup>162</sup>, which collectively affect 17.8% of children in the USA<sup>173</sup>. Classic cytogenetic studies from 1970 to 2000 identified gross chromosomal abnormalities in 4.1–13.3% of individuals with developmental disorders, although such events were also identified in developmentally typical children albeit at rates an order of magnitude lower than in those with developmental disorders<sup>14,16–18,174–176</sup>. The adoption of chromosomal microarrays in the late 2000s further extended these observations to the resolution of tens of kilobases by showing that rare and de novo CNVs were also enriched in developmental disorders and other rare Mendelian diseases<sup>98,132,133,162,168,177–179</sup>. For example, early studies of autism spectrum disorder showed that 2.2–10.1% of affected children carried at least one large de novo CNV at microarray resolution (typically >100 kb), which was fivefold greater than the rate of 0.5–1.3% found in their unaffected siblings<sup>96–98,132,180,181</sup>. The application of microarrays in clinical genetics and translational research also led to the realization that many stereotypic developmental syndromes were in fact caused by recurrent CNVs at specific loci such as deletions of chromosome 15q11–q13 in Angelman syndrome and deletions of chromosome 22q11.2 in DiGeorge syndrome<sup>182,183</sup>. Since then, genetic association studies in developmental disorders and neuropsychiatric disorders have used microarrays to identify dozens of loci where large (>100 kb) CNVs are associated with syndromic disorders, which are now collectively known as genomic disorders and are typically referenced by their chromosomal cytoband such as 22q11.2 (ref. 184) (Fig. 3 and Supplementary Information). A recent meta-analysis of microarray data from nearly one million individuals across 54 disease phenotypes reported a total of 178 distinct, large (>100 kb), rare (<1% frequency) CNVs associated with at least one phenotype, 75% (134/178) of which were associated with developmental disorders or other similarly severe paediatric-onset diseases<sup>160</sup>. Large sample sizes such as these have likely documented most of the penetrant genomic disorders in abnormal human development that occur at appreciable (>0.1%) frequencies, although it is highly likely that ultra-rare, incompletely penetrant or endophenotype-specific genomic disorders remain to be identified in even larger and more detailed cohorts.

A remarkable feature of many genomic disorders is their genomic context. The most frequent genomic disorders are often mediated by NAHR between long flanking tracts of segmental duplications and have been shown to recur de novo as independent mutational events in unrelated patients, even sometimes resulting in perfectly identical breakpoints<sup>86,132,185–187</sup>. The average genomic disorder CNV spans approximately one million nucleotides and thus impacts roughly 10 protein-coding genes in a single mutational event<sup>160,184</sup>. The involvement of numerous genes in many genomic disorders has sparked a debate over whether the genetic architecture involves individual ‘driver’ genes principally responsible for particular phenotypes (that is, a monogenic model) or the combined effects of multiple incompletely penetrant genes distributed throughout the CNV (that is, an oligogenic or polygenic model)<sup>188</sup>. As of 2024, there were a few convincing examples in support of each model but no single model could explain the pathogenic effects of all genomic disorders<sup>160</sup>. For example, LoF mutations in the *NSDI* and *SHANK3* genes have been recognized for two decades as the dominant causes of Sotos syndrome (chromosome 5q35 deletions) and Phelan–McDermid syndrome (chromosome 22q terminal deletions), respectively<sup>189,190</sup>. Some genomic disorders seem to have multiple independent drivers of different aspects of the combined syndromic phenotype. For example, mutations in *CRKL* and *TBX1* cause the kidney and heart abnormalities observed in DiGeorge syndrome, respectively<sup>191,192</sup>. Conversely, although oligogenic or polygenic models are difficult to identify in human data sets at current sample sizes, recent combinatorial gene knockouts in model organisms have indicated that the combined effects of multiple genes may be responsible for the syndromic phenotypes of some genomic disorders such as deletions of chromosomes 3q29, 16p11.2 and 16p12.1 (refs. 193–195). The effects of individual genomic disorder CNVs also need to be considered in the context of the entire genome, which can involve modifier variants in *trans*<sup>196</sup>. For example, short variants in the *RBM8A* gene in *trans* of 1q21.1 deletions are required for the manifestation of TAR syndrome<sup>197</sup>, and cumulative genome-wide polygenic risk from short variants has been shown to modify penetrance and phenotype severity for 22q11.2 deletions<sup>198</sup>. However, these examples cover a minority of all known or suspected genomic disorders reported to date, many of which appear at extremely low frequencies, are incompletely penetrant and present with variable phenotypic spectra. Moreover, several genomic disorder CNVs are observed in unselected individuals from the general population and have been shown to influence physiological traits, such as height and blood pressure, even in the absence of obvious clinical disease<sup>159,166,169,199</sup>. Therefore, we expect that nation-scale biobanks with deep phenotyping will be critical for further understanding the range of phenotypes and genetic architectures associated with genomic disorders throughout the genome.

Large rearrangements have been unassailably implicated in developmental disorders and other Mendelian diseases but smaller CNVs and other balanced SVs can also represent penetrant genetic risk factors for these same phenotypes<sup>100,200–204</sup>. These pathogenic SVs are usually identified by their predicted alteration of established disease genes previously implicated in the same phenotypes such as recurrent L1 insertions into exon 14 of the factor VIII gene in haemophilia A<sup>205</sup>. Sequencing studies of autism, developmental disorders, congenital anomalies and schizophrenia have identified enrichments of rare and de novo SVs in genes recurrently disrupted in independent patients by de novo short variants<sup>64,160,180,206–208</sup>. For example, recent work by the Autism Sequencing Consortium reported a striking enrichment of de novo CNVs in affected children that impacted a set of 373 genes previously

## a Pathogenic CNVs



**Fig. 3 | Properties of large pathogenic CNVs associated with genomic disorders.** **a**, Loci associated with large pathogenic copy number variants (CNVs) can be roughly assigned to one of three scenarios. Some of the most prominent large CNVs associated with disease to date occur through non-allelic homologous recombination (NAHR) between pairs of segmental duplications on the same chromosome, leading to recurrent deletions or duplications bookended by the paired segmental duplications<sup>187</sup>. The second scenario involves mutationally independent CNVs with different breakpoints detected in unrelated patients that all overlap a common critical region. Collectively, these two scenarios – NAHR-mediated and non-NAHR-mediated recurrent CNVs – are often referred to as ‘genomic disorders’. The third scenario involves the incidental observation of large, rare CNVs that do not exhibit striking enrichments in cohorts of patients but happen to overlap one or more genes known to be pathogenic in the disease of interest. **b**, Most established genomic disorder CNVs are typically large (>1 Mb), although this observation is likely

limited by the reliance on chromosomal microarray for the discovery of virtually all known genomic disorders to date. **c**, Nearly all known genomic disorder CNVs overlap multiple genes, with some directly impacting dozens of genes in a single mutational event. However, genomic disorder CNVs mediated by NAHR typically encompass a greater total number of genes than non-NAHR genomic disorder CNVs. It is not currently clear whether this difference is due to the inherent genetic architecture of these regions or whether the identical CNV breakpoints of patients carrying NAHR-mediated genomic disorder CNVs have simply precluded the identification of a minimal critical region within the larger NAHR segment. **d**, Not only do most genomic disorder CNVs overlap multiple genes but they also frequently overlap multiple known disease genes, defined here as genes reported to be disease-associated in the Online Mendelian Inheritance in Man (OMIM) database<sup>246</sup>. DEL, deletion; DUP, duplication. More information on the data sets and code used to generate the plots in parts **b–d** can be found in the Supplementary Information.

implicated in neurodevelopmental disorders by exome-based short variant analyses<sup>209</sup>. It is now clear that pathogenic SVs and short variants frequently converge onto the same set of critical genes and biological pathways in many diseases. However, not all pathogenic SVs in Mendelian disorders act by directly disrupting established disease genes. For example, sequencing studies of balanced translocations and inversions have discovered that three-dimensional chromatin domains – known as topologically associating domains (TADs) – are recurrently disrupted by non-coding balanced SVs in multiple unrelated patients<sup>64,152,153</sup>.

Such SV ‘positional effects’ identified to date have generally involved TADs that encompass recognized dominant disease genes such as *MEF2C*, *SOX9* or *KCNJ2* (ref. 43), although this trend is not universally true and at least 9.4% of TADs containing similar genes are disrupted by polymorphic SVs in healthy individuals in the general population<sup>28,65,210</sup>. Similarly, careful molecular studies have implicated non-coding deletions in disease at several loci, many of which have unique pathogenic mechanisms<sup>100,151,211</sup>. For example, non-coding homozygous deletions 300 kb upstream of the *ENI* gene locus result in severe congenital



limb malformations by disabling a previously unannotated long non-coding RNA, *Maenli*, required to activate *ENI* transcription during limb development<sup>211</sup>. Even a handful of non-coding mobile element insertions are now known to be pathogenic for certain rare diseases – a short interspersed element–variable number of tandem repeats–Alu mobile element insertion into an intron of the *TAFI* gene led to abnormal mRNA splicing and aberrant intron retention in individuals affected with a rare form of X-linked dystonia–Parkinsonism<sup>212</sup>. Thus, while the pathogenic mechanism for the majority of SVs responsible for severe and Mendelian disorders involves direct disruption of dosage-sensitive disease genes, all classes of SVs both within and outside of coding regions have the potential to contribute to a wide range of disorders.

## SVs in clinical diagnostics

The aetiological impact of SVs in rare and Mendelian diseases has suggested that SV ascertainment is critical for clinical genetic screening. The diagnostic yield of SV testing varies widely by phenotype and technology used for ascertainment<sup>203</sup>. For example, in intellectual disability, which affects 1.2% of all individuals in the USA<sup>173</sup>, it has been estimated that roughly 15% of all cases are attributable to a gross chromosomal abnormality and another 11–14% are attributable to a genomic disorder CNV or other pathogenic large CNV<sup>213</sup>. The 26–29% diagnostic yield of these two subsets of SVs alone exceeds the yield from all gene-disruptive coding short variants, which is estimated to be 16–25% in individuals with intellectual disabilities<sup>214,215</sup>. The contributions of large CNVs to intellectual disability and other developmental disorders are substantial enough that the American College of Medical Genetics currently recommends microarray-based screening for CNVs as the first-tier diagnostic test for patients with unexplained developmental disorders or congenital anomalies<sup>216,217</sup>. However, large-scale sequencing studies have also underscored that not all SV classes are equally impactful in a clinical diagnostic setting. For example, recent studies have shown that the diagnostic yield from mobile element insertions in coding regions is extremely low in developmental disorders (just 0.02–0.06% of all patients have a pathogenic insertion)<sup>218,219</sup>. Beyond the role of large CNVs in the genetic architecture of cognitive impairments, the diagnostic yield of small coding CNVs below microarray resolution has reached 5.4% for some developmental disorders, with estimates varying greatly by phenotype, cohort, study design and other technical factors<sup>204,220–223</sup>. A landmark WGS-based study of 13,037 patients with a rare disease in the UK National Health System reported genetic diagnoses for 16.1% of all patients, with 9.8% of pathogenic variants being SVs (primarily large deletions)<sup>203</sup>.

More esoteric classes of SVs that are strongly selected against in the population, such as balanced inversions and smaller complex SVs, are too sparse throughout the genome to robustly estimate their contributions to diagnostic yields but hundreds of these rare rearrangements have been reported as pathogenic in clinical settings for a wide variety of Mendelian diseases<sup>62,64,224–228</sup>. For example, an expanded analysis of the aforementioned UK National Health System cohort estimated that pathogenic inversions can explain only 1/750 (0.1%) families affected by rare Mendelian diseases<sup>229</sup>. However, even if gene-disruptive inversions per genome are approximately two orders of magnitude less abundant than the number of gene-disruptive CNVs, emerging evidence from long-read WGS studies has suggested that a greater proportion of inversion SVs may have disease relevance<sup>36,230</sup>. The utility of ascertaining SVs in clinical diagnostics outside of diseases with a suspected Mendelian genetic aetiology is less clear as the diagnostic yields have not been firmly established. In one example of a WGS study

of 2,081 patients hospitalized with early-onset myocardial infarction, eight patients were observed with LoF variants in the *LDLR* gene, which is known to cause familial hypercholesterolaemia<sup>231</sup>. One of these eight pathogenic LoF variants was a 7.9 kb deletion that clinically correlated with the patient's blood cholesterol levels, which might suggest that incorporating SVs at sequence resolution in diagnostic testing could provide a relative ~13% increase in genetic diagnoses for this specific patient population. Therefore, the evaluation of WGS in systematic clinical trials will be vital for establishing the diagnostic value of SVs for diseases beyond paediatric Mendelian disorders.

## Conclusions and future perspectives

Sixty years of research have proven that SVs are mutationally diverse, ubiquitous in every human genome, affect myriad functional consequences, and are broadly relevant in human evolution, health and disease. Yet, despite this wide-ranging appreciation for the roles of SVs in the human genome, much remains unknown. Below, we outline three main frontiers in SV research that we anticipate will be the topics of intense effort and the source of major breakthroughs over the coming decade.

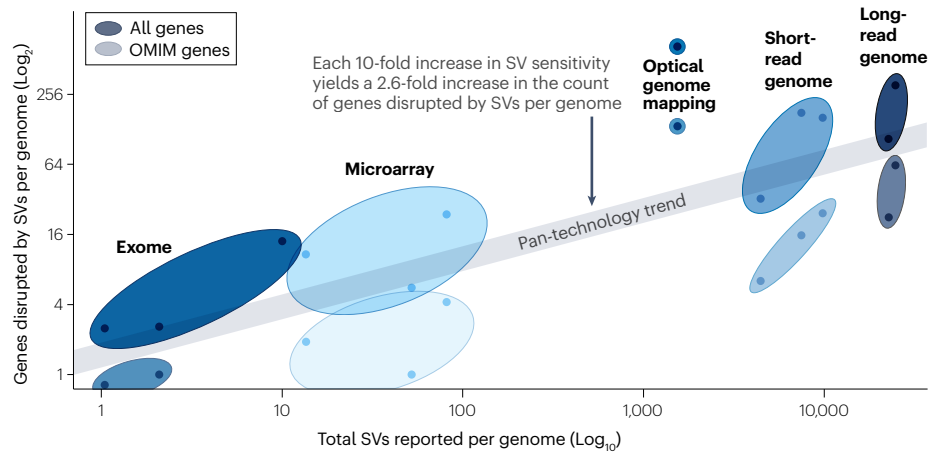
### Discovery and characterization of SVs in structurally diverse and repetitive loci

The totality of all contemporary SV data sets captures a slim fraction of the true extent of SVs in the modern human population. This deficiency is especially true for SVs localized to the most repetitive sequence contexts due to the technical limitations of detecting SVs in these regions with conventional technologies, including short-read WGS<sup>6,99</sup>. This limitation is critical for the field to surmount in the coming years, as SVs localized to many highly repetitive and biomedically relevant regions of the genome, such as the major histocompatibility complex locus, *HTT*, *SMN1*, *FMRI* and many other loci, clearly demonstrate that SVs in repetitive sequences can exert profound influences on common and rare diseases alike<sup>47,232</sup>. Long-read technologies offer the tremendous potential to resolve SVs in these loci, which comprise just ~10% of total genomic nucleotides per genome but harbour more than half of all SVs per genome<sup>40,45</sup>. Thus, concerted efforts over the next decade will be required to improve the SV reference catalogue in two critical aspects.

First, long-read WGS and genome assembly algorithms must facilitate the construction of 'pangenome' graphs that can accurately convey a population-level representation of the structural diversity present at these complex loci<sup>32,45,95,118</sup>. Indeed, graph-based computational methods have already been developed to enable pangenome representations<sup>118,233,234</sup>, but the subsequent challenge not yet surmounted is to adapt the vast reference-based genomic annotations to this new graph-based representation, ranging from carefully curated gene and transcript definitions to biochemical or chromatin data sets<sup>149,235,236</sup>.

Second, despite the justifiable optimism surrounding the emergence of new technologies and pangenome reference data sets to improve variation discovery in the human genome, these approaches alone will not improve our ability to interpret the relevance of SVs in most human traits and diseases (Fig. 4 and Supplementary Information). For example, comparisons of SVs generated by reference-based short-read and long-read WGS on the same individuals have consistently observed that long-read WGS captures more than double the total number of SVs per genome but the vast majority (>91%) of these novel variants uniquely detected by long-read WGS localize to non-coding repetitive sequences<sup>33,34,36,114</sup>. Conversely, long-read and short-read





**Fig. 4 | Yield from SV analyses across genomic technologies.** Progressively sophisticated technological breakthroughs have enabled studies of human structural variants (SVs) with increasing resolution and sensitivity, with newer technologies such as short-read and long-read genome sequencing<sup>28,29,31,35,44,113</sup> or optical genome mapping<sup>75</sup> routinely detecting orders of magnitude more SVs in each human genome than could be captured by more traditional approaches such as chromosomal microarray<sup>133,247,248</sup> or exome sequencing<sup>120,249,250</sup>. A curious corollary to this trend is that the number of gene-disruptive SVs identified in each human genome has not increased at remotely the same rate in newer technologies despite the massive gains in the absolute number of SVs per genome. This paradox is even more pronounced for genes with reported roles in disease, defined broadly here as any gene with any disease association reported in the Online Mendelian Inheritance in Man (OMIM) database<sup>346</sup>. The explanation for this trend appears to be that the exons of protein-coding genes are strongly enriched for unique, non-repetitive DNA sequence<sup>114</sup>, making the identification of SVs comparatively much easier. Accordingly, the dramatic increase in unique

SVs detected in each genome by 'third-generation' sequencing technologies such as long-read genome sequencing is driven almost entirely by tandem repeats and other complex variations in highly repetitive and non-coding sequence contexts<sup>35,114</sup>. This trend has two profound implications. First, these data do not support the notion that continued technological or algorithmic improvements in SV discovery capabilities will dramatically increase the rate of clinically diagnostic pathogenic SVs impacting currently recognized disease genes. Second, in direct contrast to the first point, these data underscore the dire need for improved mapping and annotation of SVs in the most repetitive genomic loci, which have been virtually invisible to prior large-scale efforts in human genomics, a subset of which will invariably have important roles in human disease. In this figure, each point represents the yield of SVs detected by one prominent study (referenced above) using each technology, with lighter points corresponding to those same studies if filtered to only consider OMIM genes. More information on the data sets and code used to generate this plot can be found in the Supplementary Information.

WGS exhibit strong concordance (>93%) for deletions in the remaining 90% of the genome comprised of sequences that are less repetitive and include 96% of all currently annotated protein-coding exons. Therefore, most pathogenic SVs that can currently be interpreted based on existing knowledge, annotations and clinical guidelines are captured by short-read WGS<sup>11,114</sup>. As such, SVs must be discovered from long-read WGS and genome assemblies in large populations with clinical and medical information in order to provide a foothold for comprehensive annotation of functional elements and prediction of disease association from these repetitive loci. Such efforts are under way but will likely take years to mature to the scales necessary for well-powered association studies<sup>44,52</sup>. In the short term, methods to genotype these SVs (originally identified by long-read WGS) with much larger short-read WGS data sets may be a promising strategy to achieve sample sizes appropriate for robust genotype–phenotype correlation<sup>34,115,237,238</sup>.

### Saturated genome-wide maps of SV mutation rates and dosage sensitivity

Metrics of mutational constraint for protein-coding genes have revolutionized the analysis and interpretation of short variants in human genetic research and medical genomics. These maps are built on principled models of the rate at which new mutations arise in human DNA before natural selection removes deleterious alleles from the population<sup>239</sup>, which is important to quantify the expected number

of short variants for a given locus or gene based on its primary DNA sequence alone in the absence of selection<sup>136,240</sup>. However, the same breakthroughs have not yet been realized for SVs largely due to the absence of accurate neutral mutational models for SVs at sequence resolution. The field has painstakingly gained an understanding of the mechanisms of SV formation and double-stranded DNA break repair over the last several decades<sup>241</sup>, but this knowledge has not yet translated to a comprehensive sequence-level understanding of where new SVs are most likely to arise in the human germline. Over the next decade, a crucial research focus will be to characterize SV mutation rates at nucleotide resolution and then use these insights to build genome-wide maps of dosage sensitivity and SV constraint for all genes and non-coding loci. Central to these challenges will be the integration of disparate, massive genomic data sets and distributed computing to train statistical or machine learning models capable of identifying which combination of genomic features predispose to the generation of new SVs. Given that the empirically observed rate of true de novo SVs is <1 per generation<sup>103,201</sup>, it is unlikely that accurate SV mutation rate models will be able to be trained from de novo SVs alone. Thus, another major impediment to parameterizing SV mutation rate models will be establishing the set of criteria to isolate subsets of SVs not subjected to selection pressures (analogous to synonymous short variants)<sup>239</sup>.

Massive SV data sets will be required to saturate models of dosage sensitivity for all human protein-coding genes. For example, based on the published gnomAD SV reference data set<sup>28</sup>, our power analyses

## Glossary

### Aneuploidies

Deviations of the expected copy number of an entire chromosome, usually by means of trisomy (gain of an extra copy of a chromosome) on autosomes or deviations from common sex chromosome complements (XX or XY).

### Breakpoints

A novel adjacency between two sequences not originally colinear in a reference genome; it is used to define structural variants in terms of genomic coordinates.

### Chromosomal translocations

The reciprocal exchange of two chromosome arms between non-homologous chromosomes; can result in balanced (that is, copy number neutral) or unbalanced derivative products.

### Dosage sensitivity

A property of a specified locus, such as a gene, indicating intolerance to changes in copy number, which can be specific to decreases in copy number (that is, haploinsufficiency) or increases in copy number (that is, triplosensitivity).

### Exome sequencing

Sequencing-by-synthesis of targeted loci known to encode proteins — typically 1% of all genomic DNA in humans.

### Genomic disorders

Human diseases typically characterized by syndromic, multi-system phenotypes that are caused by the recurrent deletion or duplication of a certain chromosomal segment or locus.

### Intragenic exonic duplications

Tandem duplications encompassing one or more coding exons but whose breakpoints do not extend beyond the start or end of the corresponding gene transcript.

### Karyotyping

A technique in molecular genetics that involves staining metaphase chromosomes prior to microscopic visualization. Commonly used to identify gross chromosomal abnormalities such as translocations.

### Linkage disequilibrium

A property describing two or more genetic variants whose genotypes are correlated among individuals sampled from a population, causing these variants to seem 'linked'. This phenomenon is usually due to these variants appearing on the same haplotype.

### Microarray

A technique in molecular genetics to measure the relative abundance of thousands of pre-specified nucleotide sequences in parallel by hybridizing a sample of interest to an array of short oligonucleotide probes. This technique has been commonly used to detect large copy number variants (CNVs; among other uses).

### Mobile element insertions

Insertion of a segment of genomic DNA corresponding to one of a few known families of mobile elements that can be transposed or retrotransposed via an RNA intermediate.

### Multiallelic CNVs

(mCNVs). Sites of copy number variation showing a wider distribution of copy number alleles than expected for a diploid locus, frequently reaching four or more distinct copy numbers in a sampled population.

### Non-allelic homologous recombination

(NAHR). A recombination event that occurs between two segments of DNA that have high sequence similarity but do not localize to the same genomic coordinate (that is, are non-allelic); such events can produce CNVs and other structural variants depending on the orientation of the homologous sequences involved.

### Pangenome

An alternative representation for 'reference' genome of a given species wherein an immutable core genome sequence shared by all members of a species is supplemented by known variant sequences observed among members of that species. Typically, these variant sequences are only included if they surpass some frequency threshold in a sampled population (for example, if they comprise >1% of all alleles at a given locus). Pangenomes are often computationally encoded as directional graphs with nodes corresponding to DNA sequences and edges corresponding to linear connections between those DNA sequences that are known to exist in at least one member of the species; thus, any individual chromosome can be reconstructed as a single path through this pangenome graph.

### Population stratification

A property of outbred populations, including humans, wherein certain genetic variants will appear at different frequencies in different subsets of individuals within that population due to demographic history. For example, in humans, population stratification is often observed for variants in individuals from different continents due to genetic drift or selection causing the frequency of those variants to diverge over evolutionary time (often many tens or hundreds of generations).

### Segmental duplications

Genomic segments  $\geq 1$  kb that share  $\geq 90\%$  sequence identity with at least one other paralogous region elsewhere in the genome.

### Short tandem repeats

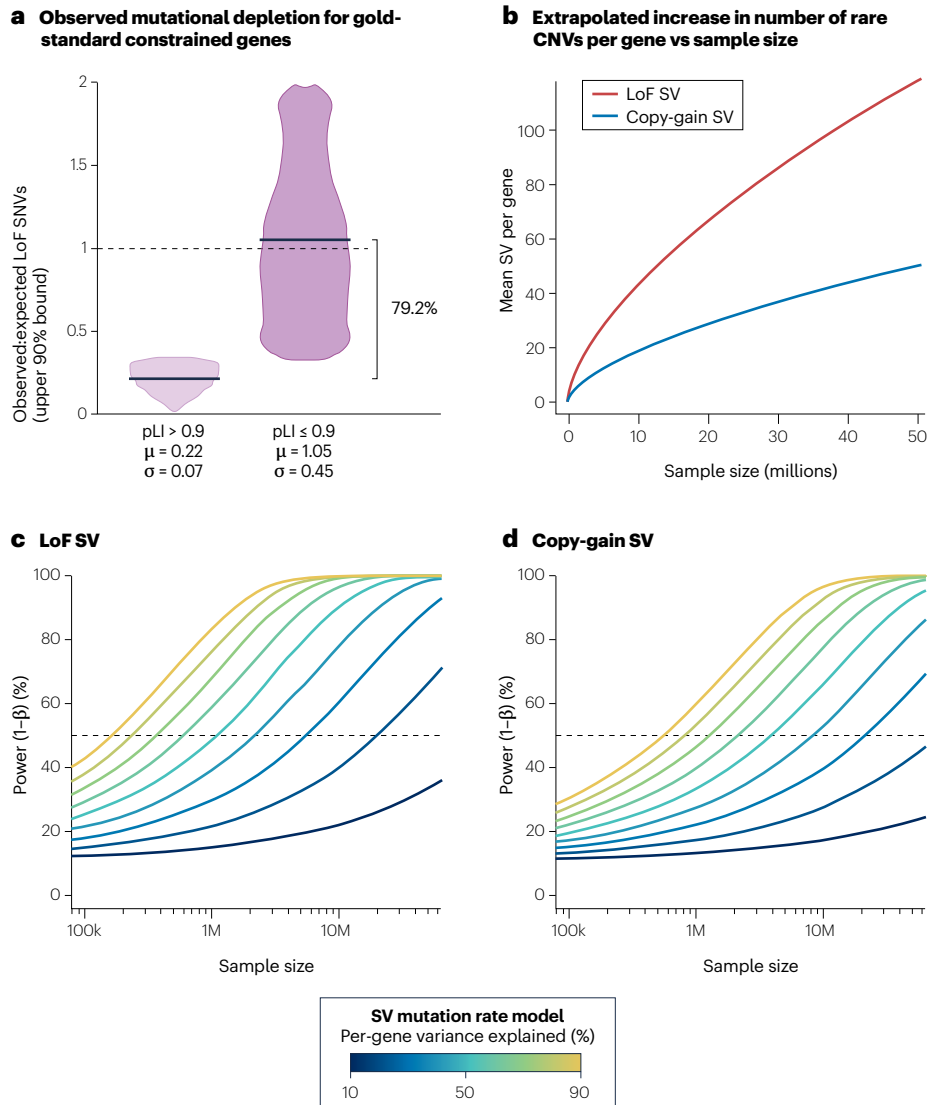
(STRs). Tracts of genomic DNA where a short (1–10 bp) sequence motif is repeated in tandem and thus predisposed to expansion or contraction due to DNA polymerase slippage. STRs can be considered a class of structural variants when their expanded or contracted allele differs by  $\geq 50$  bp as compared to a reference genome.

### Variable number of tandem repeats

Arrays of short (10–100 bp) tandem DNA repeats that vary in copy number between individuals due to DNA polymerase slippage. Variable number of tandem repeats can be considered a class of structural variants when their expanded or contracted allele differs by  $\geq 50$  bp as compared to a reference genome.

indicate that no fewer than ~100,000 samples will be required to construct metrics of haploinsufficiency from SV data with accuracy comparable to established short variant constraint maps, assuming SV mutation rate models capable of explaining >80% of the variance in observed LoF SV counts per gene (Fig. 5). Even less accurate mutational models explaining just 50% of gene-level variance in SV counts will still be powered to detect roughly half of all haploinsufficient genes at sample sizes of ~one million individuals, which should be a realistic target over the next 1–3 years based on data sets currently in

generation<sup>38,203,242,243</sup>. However, power analyses imply that the detection of triplosensitive genes from copy-gain duplication data will likely lag behind haploinsufficiency by roughly one order of magnitude in terms of the number of samples required for equal statistical power (Fig. 5). The reduced power for detecting triplosensitive genes is driven by the lower abundance of copy-gain duplications in the human genome owing to the vast size required for duplications to span the entirety of many protein-coding genes. This anticipated shortage of copy-gain duplication data can be surmounted with accurate mutation rate



**Fig. 5 | Projections for dosage sensitivity mapping in the human genome.**

**a–d.** A major goal of modern human genomics is to map and quantify loci constrained against the accumulation of variants in the general population due to negative phenotypic consequences (for example, severe disease) when disrupted by genetic variation. For structural variants (SVs), this phenomenon is usually termed dosage sensitivity, which describes the relative intolerance of a genomic locus to increased or decreased copy number caused by loss-of-function (LoF) or copy-gain SVs, respectively. Simulations were performed to project what sample sizes would be necessary to construct genome-wide maps of dosage sensitivity for all known protein-coding genes<sup>251</sup>. Estimating the magnitude of depletion of naturally occurring LoF single-nucleotide variants (SNVs) for a gold-standard set of genes constrained against LoF (defined as probability of LoF intolerance (pLI) > 0.9)<sup>136</sup> found that the average constrained gene harboured just 22% of the number of LoF SNVs expected under a neutral mutational model, and this was 79.2% less than the average unconstrained gene (**a**). The rate at which new gene-disruptive SVs will be identified for the average gene was extrapolated by empirically down-sampling SV data from the Genome Aggregation Database (gnomAD) and extrapolating

the gradual accrual of new sites as a function of increasing sample sizes following a power law (**b**). These calculations demonstrated that LoF SVs, predominantly comprised of small deletions overlapping coding exons, will continue to accrue approximately three times more rapidly than whole-gene duplications owing to the larger size of duplications required to span an entire gene locus. Finally, power analyses were performed based on the data from **a** and **b** to estimate the minimum sample size at which one would expect at least 50% power to detect the average gene sensitive to LoF SVs (haploinsufficient genes; **c**) or copy-gain SVs (triplosensitive genes; **d**). A key factor influencing the outcome of these power analyses was the relative accuracy of neutral mutational models for SVs to parameterize the number of SVs that should be expected per gene under the complete absence of selection. No such SV mutation rate models exist for humans; even if we assume a relatively accurate model could be developed to explain  $\geq 50\%$  of inter-gene variance in counts of LoF or copy-gain SVs, our power calculations estimate that sample sizes exceeding one million individuals will be required for the confident mapping of dosage-sensitive genes in the human genome. CNVs, copy number variants. Figure adapted with permission from ref. 251, Ryan Lewis Collins.

models; for example, duplication mutation rate models exceeding prediction accuracy of  $R^2 > 70\%$  would enable  $>50\%$  power to detect the average triplosensitive gene at sample sizes of roughly one million individuals. Analyses of copy-gain duplications will also have greater power to detect small triplosensitive genes or other small triplosensitive elements due to the inverse relationship between CNV size and their abundance in the general population<sup>28</sup>.

## Integration of SVs and short variants in disease association studies and diagnostic testing

The influence of SVs in human disease and clinical genetic testing is profound in comparison to other variant classes, yet few previous studies have methodically integrated short variants and SVs into disease associations. Given the rapidly growing WGS-derived reference data bases of SVs<sup>28,29</sup>, it may be feasible to impute sequence-resolved SVs into existing GWAS data sets for common and complex traits by leveraging linkage disequilibrium between short variants and SVs<sup>26</sup>. Imputation may become a powerful technique to include common SVs in GWAS but it is unlikely to obviate the need for direct discovery and genotyping of duplications and mCNVs in GWAS data sets given their lower average observed linkage disequilibrium<sup>28,107,121</sup>. However, statistical methods combining exome sequencing and microarrays have recently been proposed and applied to large biobank data sets to impute tandem repeats in coding sequences<sup>170</sup>; such approaches may also improve the imputation accuracy for larger SVs that disrupt linkage disequilibrium such as mCNVs and common duplications. Imputation will also not be a viable approach for rare and de novo SVs in severe Mendelian phenotypes or clinical testing, which will require ab initio SV discovery in large WGS cohorts. Even when SVs can be imputed (or directly genotyped) in large research cohorts, improved statistical frameworks must be developed to jointly evaluate the combined effects of SNVs, indels and SVs in disease association studies. Several promising approaches have been proposed<sup>180,209</sup> but none have been broadly adopted yet by the biomedical research community nor extended to large WGS data sets. Specialized assays and algorithms will also be required to profile somatically arising SVs in the context of diseases other than cancer, which is further impeded by the requirement of sampling the specific somatic tissues relevant for each disease, although recent findings have given cause for optimism on this topic<sup>49,110,244</sup>.

Ultimately, the potential for SV research to benefit human health is through more accurate and sensitive genetic diagnostics in clinical practice. The current approaches to SV ascertainment and interpretation in diagnostic screens are heterogeneous and severely limited by myriad factors. Among these limitations is that almost all published estimates of diagnostic yields from SVs are derived from relatively low-resolution technologies such as karyotyping, chromosomal microarray and exome sequencing; by contrast, short-read and long-read WGS can capture the entire range of SV sizes and frequencies in a single assay and could represent a future replacement for existing platforms for diagnostic SV testing<sup>222</sup>. Notably, this future development presents an equity challenge in genomic testing as only a handful of global sites worldwide currently have the technical and analytic capacity to perform routine long-read WGS data processing and analyses of  $>4$  million short variants and  $\sim 25,000$  SVs per human genome. Moreover, the ability to detect SVs that are cryptic to conventional technologies will not by itself generate immediate clinical utility without major advances in rigorous, evidence-based clinical guidelines for interpreting such SVs – in particular, non-coding repeat expansions and SVs in highly repetitive and non-genic contexts where most SVs unique to long-read

assemblies are discovered. Such studies are ongoing across multiple consortia, and the interpretation of non-coding SVs and those events with molecular consequences other than LoF in a given gene presents a major next frontier for the field<sup>11,245</sup>. Such SVs rarely satisfy existing diagnostic criteria for pathogenicity despite the average genome carrying approximately 22 duplicated genes and another 11 genes interrupted by intragenic exonic duplications in addition to hundreds of balanced, intronic and near-coding SVs, an indeterminate subset of which likely influence gene expression or function through splicing, *cis*-regulation or other mechanisms<sup>28</sup>. Finally, as the field moves towards more comprehensive, unified representations of genome structure in the form of genome graphs and pangenome references, new clinical standards will need to be established to interpret the SVs implied by the genome graphs of individual patients relative to a pangenome reference.

In conclusion, the field of human genetics has made paradigm-shifting leaps forward in the discovery, representation and interpretation of variation that alters genome organization and its consequences on genome function. There are grand challenges that remain, most notably the prediction of the direct and regulatory functional impact of SVs, the adoption of best practices for data processing and SV discovery, a uniform approach to the representation and annotation of genome assemblies unique to each individual, and the adoption of robust standards that enable uniform and globally accessible clinical genetic testing for all individuals.

Published online: 21 January 2025

## References

- Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- Orita, M., Suzuki, Y., Sekiya, T. & Hayashi, K. Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics* **5**, 874–879 (1989).
- Wang, D. G. et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
- Altshuler, D., Donnelly, P. & The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
- Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
- Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
- Kosugi, S. et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20**, 117 (2019).
- Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
- Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
- This paper describes the phase 3 SV release of the 1000 Genomes Project, which provided an unprecedented level of insight into the diversity of SVs in the global human population and has stood as one of the gold-standard multi-ancestry data sets in the SV field over the subsequent decade.**
- Korbel, J. O. et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
- Riggs, E. R. et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.* **22**, 245–257 (2020).
- Jacobs, P. A., Baikie, A. G., Court Brown, W. M. & Strong, J. A. The somatic chromosomes in mongolism. *Lancet* **1**, 710 (1959).
- Tjio, J. H. & Levan, A. The chromosome number of man. *Hereditas* **42**, 1–6 (1956).
- Warburton, D. De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *Am. J. Hum. Genet.* **49**, 995–1013 (1991).
- Cohen, A. J. et al. Hereditary renal-cell carcinoma associated with a chromosomal translocation. *N. Engl. J. Med.* **301**, 592–595 (1979).
- Funderburk, S. J., Spence, M. A. & Sparkes, R. S. Mental retardation associated with “balanced” chromosome rearrangements. *Am. J. Hum. Genet.* **29**, 136–141 (1977).
- Hou, J.-W., Wang, T.-R. & Chuang, S.-M. An epidemiological and aetiological study of children with intellectual disability in Taiwan. *J. Intellect. Disabil. Res.* **42**, 137–143 (1998).



18. Knight, S. J. L. et al. Subtle chromosomal rearrangements in children with unexplained mental retardation. *Lancet* **354**, 1676–1681 (1999).
19. Iafrate, A. J. et al. Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
20. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
21. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
22. Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
23. McCarroll, S. A. et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
24. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurler, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
25. Mills, R. E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
26. Hehir-Kwa, J. Y. et al. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat. Commun.* **7**, 12989 (2016).
27. Sudmant, P. H. et al. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
28. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).  
**This paper describes the initial SV component of the gnomAD, which is a widely adopted reference resource for evaluating the frequencies and distributions of genetic variation in the human population.**
29. Abel, H. J. et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).
30. Almarri, M. A. et al. Population structure, stratification, and introgression of human structural variation. *Cell* **182**, 189–199 (2020).
31. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440 (2022).
32. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
33. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).  
**In this study, members of the HGSVC apply an exhaustive combination of genomic technologies to thoroughly characterize all SVs present in the genomes of three parent-child trios, which yields one of the most comprehensive SV data sets produced to date for an individual set of genomes.**
34. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675 (2019).
35. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
36. Porubsky, D. et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986–2005 (2022).
37. Hallast, P. et al. Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature* **621**, 355–364 (2023).
38. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
39. All of Us Research Program Genomics Investigators. Genomic data in the All of Us research program. *Nature* **627**, 340–346 (2024).
40. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).  
**This seminal publication from the T2T consortium describes the first complete (gapless) sequencing of a single human genome, which marks the beginning of the era of complete human genomes and pangenome graphs.**
41. Wagner, J. et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* **40**, 672–680 (2022).
42. Hurler, M. E., Dermitzakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends Genet.* **24**, 238–245 (2008).
43. Spielmann, M., Lupianez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
44. Beyter, D. et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).  
**This study is the largest published analysis of SVs based on long-read genome sequencing in a human population to date and demonstrates the value of long-read technologies in identifying and genotyping SVs — especially tandem repeats — associated with human traits and diseases.**
45. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
46. Mahmoud, M. et al. Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).
47. Chin, C. S. et al. A diploid assembly-benchmark for variants in the major histocompatibility complex. *Nat. Commun.* **11**, 4794 (2020).
48. Yi, K. & Ju, Y. S. Patterns and mechanisms of structural variations in human cancer. *Exp. Mol. Med.* **50**, 98 (2018).
49. Loh, P. R. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
50. McConnell, M. J. et al. Mosaic copy number variation in human neurons. *Science* **342**, 632–637 (2013).
51. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).  
**This study is a high-quality survey of mCNVs in the human population and includes the initial characterization of ‘runaway haplotypes’ in certain ancestry groups that have undergone a recent expansion in copy number.**
52. Mahmoud, M. et al. Utility of long-read sequencing for All of Us. *Nat. Commun.* **15**, 837 (2024).
53. Luning Prak, E. T. & Kazazian, H. H. Mobile elements and the human genome. *Nat. Rev. Genet.* **1**, 134–144 (2000).
54. Stewart, C. et al. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* **7**, e1002236 (2011).
55. Gardner, E. J. et al. The mobile element locator tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).
56. Sherman, R. M. et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
57. Kehr, B. et al. Diversity in non-repetitive human sequences not found in the reference genome. *Nat. Genet.* **49**, 588–593 (2017).
58. Wong, K. H. Y., Levy-Sakin, M. & Kwok, P.-Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat. Commun.* **9**, 3040 (2018).
59. Wei, W. et al. Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. *Nature* **611**, 105–114 (2022).
60. Fan, H. & Chu, J. Y. A brief review of short tandem repeat mutation. *Genomics Proteom. Bioinform.* **5**, 7–14 (2007).
61. Willems, T., Gymrek, M., Highnam, G., Mittelman, D. & Erlich, Y. The landscape of human STR variation. *Genome Res.* **24**, 1894–1904 (2014).
62. Talkowski, M. E. et al. Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *Am. J. Hum. Genet.* **88**, 469–481 (2011).
63. Talkowski, M. E. et al. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* **149**, 525–537 (2012).
64. Redin, C. et al. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.* **49**, 36–45 (2017).  
**This paper is the largest published analysis of balanced chromosomal abnormalities at nucleotide resolution in developmental disorders, which emphasizes the role of this unique class of SVs in the pathogenesis of severe pediatric disorders.**
65. Lowther, C. et al. Balanced chromosomal rearrangements offer insights into coding and noncoding genomic features associated with developmental disorders. Preprint at *medRxiv* <https://doi.org/10.1101/2022.02.15.22270795> (2022).
66. Chiang, C. et al. Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat. Genet.* **44**, 390–397, S391 (2012).
67. Abyzov, A. et al. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat. Commun.* **6**, 7256 (2015).
68. Brand, H. et al. Paired-duplication signatures mark cryptic inversions and other complex structural variation. *Am. J. Hum. Genet.* **97**, 170–176 (2015).
69. Borg, K. et al. Molecular analysis of a constitutional complex genome rearrangement with 11 breakpoints involving chromosomes 3, 11, 12, and 21 and a approximately 0.5-Mb submicroscopic deletion in a patient with mild mental retardation. *Hum. Genet.* **118**, 267–275 (2005).
70. Carvalho, C. M. et al. Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Hum. Mol. Genet.* **18**, 2188–2203 (2009).
71. Zhang, F. et al. The DNA replication FoStEs/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat. Genet.* **41**, 849–853 (2009).
72. Carvalho, C. M. et al. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.* **43**, 1074–1081 (2011).
73. Quinlan, A. R. & Hall, I. M. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet.* **28**, 43–53 (2012).
74. Sanders, A. D. et al. Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res.* **26**, 1575–1587 (2016).
75. Levy-Sakin, M. et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* **10**, 1025 (2019).
76. Hermetz, K. E. et al. Large inverted duplications in the human genome form via a fold-back mechanism. *PLoS Genet.* **10**, e1004139 (2014).
77. Collins, R. L. et al. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* **18**, 36 (2017).
78. Kloosterman, W. P. et al. Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum. Mol. Genet.* **20**, 1916–1924 (2011).
79. Chatron, N. et al. The enrichment of breakpoints in late-replicating chromatin provides novel insights into chromoanagenesis mechanisms. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.07.17.206771> (2020).
80. Weckselblatt, B., Hermetz, K. E. & Rudd, M. K. Unbalanced translocations arise from diverse mutational mechanisms including chromothripsis. *Genome Res.* **25**, 937–947 (2015).
81. Liu, P. et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* **146**, 889–903 (2011).

82. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).  
**This study reports the first observation of chromothripsis, which ignited entire sub-disciplines within human genetics and cancer genomics focused on identifying and characterizing extremely complex genomic rearrangements.**
83. Cortés-Ciriano, I. et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.* **52**, 331–341 (2020).
84. de Pagter, M. S. et al. Chromothripsis in healthy individuals affects multiple protein-coding genes and can result in severe congenital abnormalities in offspring. *Am. J. Hum. Genet.* **96**, 651–656 (2015).
85. Weckselblatt, B. & Rudd, M. K. Human structural variation: mechanisms of chromosome rearrangements. *Trends Genet.* **31**, 587–599 (2015).
86. Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
87. Gu, W., Zhang, F. & Lupski, J. R. Mechanisms for human genomic rearrangements. *Pathogenetics* **1**, 4 (2008).
88. Lieber, M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.* **79**, 181–211 (2010).
89. Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5**, e1000327 (2009).
90. Ottaviani, D., LeCain, M. & Sheer, D. The role of microhomology in genomic structural variation. *Trends Genet.* **30**, 85–94 (2014).
91. Balachandran, P. et al. Transposable element-mediated rearrangements are prevalent in human genomes. *Nat. Commun.* **13**, 7115 (2022).
92. Startek, M. et al. Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. *Nucleic Acids Res.* **43**, 2188–2198 (2015).
93. Zhang, C. Z. et al. Chromothripsis from DNA damage in micronuclei. *Nature* **522**, 179–184 (2015).
94. Logsdon, G. A. et al. The variation and evolution of complete human centromeres. *Nature* **629**, 136–145 (2024).
95. Wang, T. et al. The human pangenome project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
96. Itsara, A. et al. De novo rates and selection of large copy number variation. *Genome Res.* **20**, 1469–1481 (2010).
97. Sanders, S. J. et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).
98. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
99. Lappalainen, T., Scott, A. J., Brandt, M. & Hall, I. M. Genomic analysis in the age of human genome sequencing. *Cell* **177**, 70–84 (2019).
100. Brandler, W. M. et al. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
101. Kloosterman, W. P. et al. Characteristics of de novo structural changes in the human genome. *Genome Res.* **25**, 792–801 (2015).
102. Feusier, J. et al. Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res.* **29**, 1567–1577 (2019).
103. Belyeu, J. R. et al. De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am. J. Hum. Genet.* **108**, 597–607 (2021).  
**This study performed SV analyses from WGS of several thousand parent–child trios, which enabled the most accurate empirical estimates of SV mutation rates in humans to date.**
104. Halman, A. & Oshlack, A. Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. *F1000Research* **9**, 200 (2020).
105. Mitra, I. et al. Patterns of de novo tandem repeat mutations and their role in autism. *Nature* **589**, 246–250 (2021).
106. Fu, W., Zhang, F., Wang, Y., Gu, X. & Jin, L. Identification of copy number variation hotspots in human populations. *Am. J. Hum. Genet.* **87**, 494–504 (2010).
107. Conrad, D. F. & Hurler, M. E. The population genetics of structural variation. *Nat. Genet.* **39**, S30–S36 (2007).
108. Solís-Moruno, M., Batlle-Masó, L., Bonet, N., Aróstegui, J. I. & Casals, F. Somatic genetic variation in healthy tissue and non-cancer diseases. *Eur. J. Hum. Genet.* **31**, 48–54 (2023).
109. Yu, X. et al. Digital microfluidics-based digital counting of single-cell copy number variation (dd-scCNV Seq). *Proc. Natl Acad. Sci. USA* **120**, e2221934120 (2023).
110. Gao, T. et al. A pan-tissue survey of mosaic chromosomal alterations in 948 individuals. *Nat. Genet.* **55**, 1901–1911 (2023).
111. Li, S., Carss, K. J., Halldórsson, B. V. & Cortes, A. Whole-genome sequencing of half-a-million UK Biobank participants. Preprint at medRxiv <https://doi.org/10.1101/2023.12.06.23299426> (2023).
112. Jun, G. et al. Structural variation across 138,134 samples in the TOPMed consortium. Preprint at bioRxiv <https://doi.org/10.1101/2023.01.25.525428> (2023).
113. Logsdon, G. A. et al. Complex genetic variation in nearly complete human genomes. Preprint at bioRxiv <https://doi.org/10.1101/2024.09.24.614721> (2024).
114. Zhao, X. et al. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Hum. Genet.* **108**, 919–928 (2021).
115. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
116. Ziaei Jam, H. et al. A deep population reference panel of tandem repeat variation. *Nat. Commun.* **14**, 6711 (2023).
117. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
118. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
119. Itsara, A. et al. Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84**, 148–161 (2009).
120. Ruderfer, D. M. et al. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat. Genet.* **48**, 1107–1111 (2016).  
**This study from the Exome Aggregation Consortium represents one of the first well-powered attempts to quantify both haploinsufficiency and triplosensitivity for all human protein-coding genes.**
121. Jakubosky, D. et al. Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. *Nat. Commun.* **11**, 2928 (2020).
122. Ohno, S. *Evolution by Gene Duplication* (Springer-Verlag, 1970).
123. Dumas, L. et al. Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res.* **17**, 1266–1277 (2007).
124. Dennis, M. Y. & Eichler, E. E. Human adaptation and evolution by segmental duplication. *Curr. Opin. Genet. Dev.* **41**, 44–52 (2016).
125. Fiddes, I. T. et al. Human-specific NOTCH2NL genes affect notch signaling and cortical neurogenesis. *Cell* **173**, 1356–1369 (2018).
126. Giannuzzi, G. et al. The human-specific BOLA2 duplication modifies iron homeostasis and anemia predisposition in chromosome 16p11.2 autism individuals. *Am. J. Hum. Genet.* **105**, 947–958 (2019).
127. Boettger, L. M. et al. Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat. Genet.* **48**, 359–366 (2016).
128. Perry, G. H. et al. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).  
**This study describes the initial observation that amylase gene copy number differs among human populations, correlates with salivary amylase protein abundance, and mirrors social transitions to agrarianism, collectively comprising one of the most famous examples of human adaptation due to positively selected SVs.**
129. Bolognini, D. et al. Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature* **634**, 617–625 (2024).
130. Stefánsson, H. et al. A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
131. Boettger, L. M., Handsaker, R. E., Zody, M. C. & McCarroll, S. A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* **44**, 881–885 (2012).
132. Marshall, C. R. et al. Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).  
**This study reported one of the first systematic analyses of balanced and unbalanced chromosomal abnormalities in individuals with autism spectrum disorder, revealing that large SVs are a major contributor to abnormal neurodevelopment.**
133. Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
134. Coe, B. P. et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014).
135. Douard, E. et al. Effect sizes of deletions and duplications on autism risk across the genome. *Am. J. Psychiatry* **178**, 87–98 (2021).
136. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
137. Han, L. et al. Functional annotation of rare structural variation in the human brain. *Nat. Commun.* **11**, 2990 (2020).
138. Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
139. Ferraro, N. M. et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369**, eaaz5900 (2020).
140. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
141. Gunning, A. C. et al. Recurrent de novo NAHR reciprocal duplications in the ATAD3 gene cluster cause a neurogenetic trait with perturbed cholesterol and mitochondrial metabolism. *Am. J. Hum. Genet.* **106**, 272–279 (2020).
142. Usdin, K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res.* **18**, 1011–1019 (2008).
143. Vorechovsky, I. Transposable elements in disease-associated cryptic exons. *Hum. Genet.* **127**, 135–154 (2010).
144. Sundaram, V. et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* **24**, 1963–1976 (2014).
145. Cao, X. et al. Polymorphic mobile element insertions contribute to gene expression and alternative splicing in human tissues. *Genome Biol.* **21**, 185 (2020).
146. Liang, L. et al. Complementary Alu sequences mediate enhancer-promoter selectivity. *Nature* **619**, 868–875 (2023).
147. Middelkamp, S. et al. Molecular dissection of germline chromothripsis in a developmental context using patient-derived iPSCs. *Genome Med.* **9**, 9 (2017).
148. van Heesch, S. et al. Genomic and functional overlap between somatic and germline chromosomal rearrangements. *Cell Rep.* **9**, 2001–2010 (2014).
149. Moore, J. E. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

150. Fudenberg, G. & Pollard, K. S. Chromatin features constrain structural variation across evolutionary timescales. *Proc. Natl Acad. Sci. USA* **116**, 2175–2180 (2019).
151. Oz-Levi, D. et al. Noncoding deletions reveal a gene that is critical for intestinal function. *Nature* **571**, 107–111 (2019).
152. Franke, M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).
153. Lupianez, D. G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).  
**This study was among the first to demonstrate that SVs can cause Mendelian developmental diseases by altering the three-dimensional chromatin architecture of the genome rather than by direct disruption of coding genes themselves.**
154. Monlong, J. et al. Global characterization of copy number variants in epilepsy patients from whole genome sequencing. *PLoS Genet.* **14**, e1007285 (2018).
155. D’Haene, E. & Vergult, S. Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. *Genet. Med.* **23**, 34–46 (2021).
156. Jakubosky, D. et al. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat. Commun.* **11**, 2927 (2020).
157. Scott, A. J., Chiang, C. & Hall, I. M. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res.* **31**, 2249–2257 (2021).  
**This study reports the most recent SV analyses from the GTEx project, which is the largest and best-powered quantification of the effects of SVs on gene expression in humans available to date.**
158. Rice, A. M. & McLysaght, A. Dosage sensitivity is a major determinant of human copy number variant pathogenicity. *Nat. Commun.* **8**, 14366 (2017).
159. Aguirre, M., Rivas, M. A. & Priest, J. Phenome-wide burden of copy-number variation in the UK biobank. *Am. J. Hum. Genet.* **105**, 373–383 (2019).
160. Collins, R. L. et al. A cross-disorder dosage sensitivity map of the human genome. *Cell* <https://doi.org/10.1016/j.cell.2022.06.036> (2022).  
**This study reports the aggregation and systematic analysis of rare CNVs in nearly one million people, enabling genome-wide association scans of rare deletions and duplications for 54 diseases and the construction of well-calibrated haploinsufficiency and triplosensitivity metrics for all autosomal protein-coding genes.**
161. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, e1001154 (2010).
162. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010).
163. McCarroll, S. A. Extending genome-wide association studies to copy-number variation. *Hum. Mol. Genet.* **17**, R135–R142 (2008).
164. Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
165. Craddock, N. et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720 (2010).
166. Mace, A. et al. CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. *Nat. Commun.* **8**, 744 (2017).
167. Li, Y. R. et al. Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. *Nat. Commun.* **11**, 255 (2020).
168. Glessner, J. T. et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459**, 569–573 (2009).
169. Hujoel, M. L. A. et al. Influences of rare copy-number variation on human complex traits. *Cell* **185**, 4233–4248 (2022).
170. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P. R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat. Genet.* **53**, 1260–1269 (2021).
171. Payer, L. M. et al. Structural variants caused by *Alu* insertions are associated with risks for many human diseases. *Proc. Natl Acad. Sci. USA* **114**, E3984–E3992 (2017).
172. Kamitaki, N. et al. Complement genes contribute sex-biased vulnerability in diverse disorders. *Nature* **582**, 577–581 (2020).
173. Zablotsky, B. et al. Prevalence and trends of developmental disabilities among children in the United States: 2009–2017. *Pediatrics* **144**, e20190811 (2019).
174. Bell, J. et al. A total population study of diagnosed chromosome abnormalities in Queensland, Australia. *Clin. Genet.* **22**, 49–56 (1982).
175. van Karnebeek, C. D., Jansweijer, M. C., Leenders, A. G., Offringa, M. & Hennekam, R. C. Diagnostic investigations in individuals with mental retardation: a systematic literature review of their usefulness. *Eur. J. Hum. Genet.* **13**, 6–25 (2005).
176. Verma, R. S. & Dosik, H. Incidence of major chromosomal abnormalities in a referred population for suspected chromosomal aberrations: a report of 357 cases. *Clin. Genet.* **17**, 305–308 (1980).
177. Wright, C. F. et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).  
**This study describes the analysis and clinical interpretation of CNVs detected by chromosomal microarray in the Deciphering Developmental Disorders project, which provided one of the most comprehensive estimates of diagnostic yield for CNV testing in paediatric developmental disorders.**
178. Wapner, R. J. et al. Chromosomal microarray versus karyotyping for prenatal diagnosis. *N. Engl. J. Med.* **367**, 2175–2184 (2012).
179. Sajjan, S. A. et al. Both rare and de novo copy number variants are prevalent in agenesis of the corpus callosum but not in cerebellar hypoplasia or polymicrogyria. *PLoS Genet.* **9**, e1003823 (2013).
180. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
181. Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
182. Magenis, R. E., Brown, M. G., Lacy, D. A., Budden, S. & LaFranchi, S. Is Angelman syndrome an alternate result of del(15)(q11q13)? *Am. J. Med. Genet.* **28**, 829–838 (1987).
183. Driscoll, D., Budarf, M. & Emanuel, B. A genetic etiology for DiGeorge syndrome: consistent deletions and microdeletions of 22q11. *Am. J. Hum. Genet.* **50**, 924 (1992).
184. Harel, T. & Lupski, J. R. Genomic disorders 20 years on—mechanisms for clinical manifestations. *Clin. Genet.* **93**, 439–449 (2018).
185. Weiss, L. A. et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).  
**This study reports the initial discovery of reciprocal CNVs at the 16p11.2 chromosomal locus and autism spectrum disorder; this 16p11.2 CNV is now recognized as one of the single most common genetic causes of abnormal human neurodevelopment.**
186. McCarthy, S. E. et al. Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.* **41**, 1223–1227 (2009).
187. Nuttle, X. et al. Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536**, 205–209 (2016).
188. Smolen, C. & Girirajan, S. The gene dose makes the disease. *Cell* **185**, 2850–2852 (2022).
189. Kurotaki, N. et al. Haploinsufficiency of NSD1 causes Sotos syndrome. *Nat. Genet.* **30**, 365–366 (2002).
190. Wilson, H. L. et al. Molecular characterisation of the 22q13 deletion syndrome supports the role of haploinsufficiency of SHANK3/PROSAP2 in the major neurological symptoms. *J. Med. Genet.* **40**, 575–584 (2003).
191. Lopez-Rivera, E. et al. Genetic drivers of kidney defects in the DiGeorge syndrome. *N. Engl. J. Med.* **376**, 742–754 (2017).
192. Lindsay, E. A. et al. Tbx1 haploinsufficiency in the DiGeorge syndrome region causes aortic arch defects in mice. *Nature* **410**, 97–101 (2001).  
**This study was one of the first to demonstrate that loss of a single gene (TBX1) within a larger genomic disorder CNV locus (22q11.2 deletion) is individually associated with one of the constituent phenotypes commonly observed in CNV carrier patients, which provided important empirical evidence for an oligogenic basis of genomic disorders.**
193. Iyer, J. et al. Pervasive genetic interactions modulate neurodevelopmental defects of the autism-associated 16p11.2 deletion in *Drosophila melanogaster*. *Nat. Commun.* **9**, 2548 (2018).
194. Singh, M. D. et al. NCBP2 modulates neurodevelopmental defects of the 3q29 deletion in *Drosophila* and *Xenopus laevis* models. *PLoS Genet.* **16**, e1008590 (2020).
195. Pizzo, L. et al. Functional assessment of the “two-hit” model for neurodevelopmental defects in *Drosophila* and *X. laevis*. *PLoS Genet.* **17**, e1009112 (2021).
196. Girirajan, S. et al. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *N. Engl. J. Med.* **367**, 1321–1331 (2012).
197. Albers, C. A. et al. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. *Nat. Genet.* **44**, 435–439 (2012).
198. Davies, R. W. et al. Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. *Nat. Med.* **26**, 1912–1918 (2020).
199. Owen, D. et al. Effects of pathogenic CNVs on physical traits in participants of the UK Biobank. *BMC Genomics* **19**, 867 (2018).
200. Gillissen, C. et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511**, 344–347 (2014).
201. Werling, D. M. et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018).
202. Turner, T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722 (2017).
203. Turro, E. et al. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**, 96–102 (2020).
204. Poultney, C. S. et al. Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. *Am. J. Hum. Genet.* **93**, 607–619 (2013).
205. Kazazian, H. H. Jr. et al. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166 (1988).
206. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584 (2020).
207. Coe, B. P. et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
208. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
209. Fu, J. M. et al. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat. Genet.* **54**, 1320–1331 (2022).
210. Shanta, O., Noor, A. & Sebat, J. The effects of common structural variants on 3D chromatin structure. *BMC Genomics* **21**, 95 (2020).
211. Allou, L. et al. Non-coding deletions identify Maenli lncRNA as a limb-specific En1 regulator. *Nature* **592**, 93–98 (2021).
212. Anechiky, T. et al. Dissecting the causal mechanism of X-linked dystonia-parkinsonism by integrating genome and transcriptome assembly. *Cell* **172**, 897–909 (2018).
213. Maia, N., Nabais Sá, M. J., Melo-Pires, M., de Brouwer, A. P. M. & Jorge, P. Intellectual disability genomics: current state, pitfalls and future challenges. *BMC Genomics* **22**, 909 (2021).



214. de Ligt, J. et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
215. Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
216. Miller, D. T. et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am. J. Hum. Genet.* **86**, 749–764 (2010).
217. Schaefer, G. B. & Mendelsohn, N. J. Clinical genetics evaluation in identifying the etiology of autism spectrum disorders: 2013 guideline revisions. *Genet. Med.* **15**, 399–407 (2013).
218. Gardner, E. J. et al. Contribution of retrotransposition to developmental disorders. *Nat. Commun.* **10**, 4630 (2019).
219. Torene, R. I. et al. Mobile element insertion detection in 89,874 clinical exomes. *Genet. Med.* **22**, 974–978 (2020).
220. Pfundt, R. et al. Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. *Genet. Med.* **19**, 667–675 (2017).
221. Vissers, L. et al. A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. *Genet. Med.* **19**, 1055–1063 (2017).
222. Lowther, C. et al. Systematic evaluation of genome sequencing as a first-tier diagnostic test for prenatal and pediatric disorders. *Am. J. Hum. Genet.* **110**, 1454–1469 (2023).
223. Lord, J. et al. Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. *Lancet* **393**, 747–757 (2019).
224. Talkowski, M. E. et al. Clinical diagnosis by whole-genome sequencing of a prenatal sample. *N. Engl. J. Med.* **367**, 2226–2232 (2012).
225. Sanchis-Juan, A. Complex structural variants resolved by short-read and long-read whole genome sequencing in Mendelian disorders. *Genome Med.* **10**, 95 (2018).
226. Wahlster, L. et al. Familial thrombocytopenia due to a complex structural variant resulting in a WAC-ANKRD26 fusion transcript. *J. Exp. Med.* **218**, e20210444 (2021).
227. Witt, D. et al. Genome sequencing identifies complex structural MLH1 variant in undivided Lynch syndrome. *Mol. Genet. Genom. Med.* **11**, e2151 (2023).
228. Lilleväli, H. et al. Genome sequencing identifies a homozygous inversion disrupting QDPR as a cause for dihydropteridine reductase deficiency. *Mol. Genet. Genom. Med.* **8**, e1154 (2020).
229. Pagnamenta, A. T. et al. The impact of inversions across 33,924 families with rare disease from a national genome sequencing project. *Am. J. Hum. Genet.* **111**, 1140–1164 (2024).
230. Höps, W. et al. Impact and characterization of serial structural variations across humans and great apes. *Nat. Commun.* **15**, 8007 (2024).
231. Khera, A. V. et al. Whole genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation* **139**, 1593–1602 (2019).
232. Depienne, C. & Mandel, J.-L. 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* **108**, 764–785 (2021).
233. Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *Bioinformatics* **39**, btac743 (2023).
234. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
235. Morales, J. et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).
236. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
237. Eggertsson, H. P. et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).
238. Chen, S. et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
239. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
240. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
241. Scully, R., Panday, A., Elango, R. & Willis, N. A. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat. Rev. Mol. Cell Biol.* **20**, 698–714 (2019).
242. Venner, E. et al. Whole-genome sequencing as an investigational device for return of hereditary disease risk and pharmacogenomic results as part of the All of Us Research Program. *Genome Med.* **14**, 34 (2022).
243. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
244. Sherman, M. A. et al. Large mosaic copy number variations confer autism risk. *Nat. Neurosci.* **24**, 197–203 (2021).
245. Riggs, E. R. et al. Towards an evidence-based process for the clinical interpretation of copy number variation. *Clin. Genet.* **81**, 403–412 (2012).
246. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).
247. Uddin, M. et al. A high-resolution copy-number variation resource for clinical and population genetics. *Genet. Med.* **17**, 747–752 (2015).
248. Zarrei, M. et al. Gene copy number variation and pediatric mental health/neurodevelopment in a general population. *Hum. Mol. Genet.* **32**, 2411–2421 (2023).
249. Maxwell, E. K. et al. Profiling copy number variation and disease associations from 50,726 DiscovEHR Study exomes. Preprint at *bioRxiv* <https://doi.org/10.1101/119461> (2017).
250. Babadi, M. et al. GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nat. Genet.* **55**, 1589–1597 (2023).
251. Collins, R. L. *The Landscape and Consequences of Structural Variation in the Human Genome*. Thesis, Harvard University (2022).

## Acknowledgements

We thank Stephanie Hao, M.S., for her assistance and consultation with figure generation. R.L.C. and M.E.T. are supported by the National Institutes of Health (HD081256, HD096326, HD099547, HG008895, MH106826, NS093200, P01GM061354, P50HD028138, R01HD081256, R01HD091797, R01HD096326, R01MH11776, R01MH115957, U01MH105669, and K99CA286805), the Simons Foundation for Autism Research Initiative (SFARI #573206), the National Science Foundation (GRFP #2017240332), the Board of Trustees of the Dana-Farber Cancer Institute, and the KBF Canada Foundation.

## Author contributions

The authors contributed equally to all aspects of the article.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41576-024-00808-9>.

**Peer review information** *Nature Reviews Genetics* thanks Andrew Sharp; Fritz Sedlazeck; and Ryan Mills, who co-reviewed with Weichen Zhou, for their contribution to the peer review of this work.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2025