frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 29 November 2022
DOI 10.3389/fgene.2022.1060898

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU), United States

REVIEWED BY
Zhong-duo Wang,
Guangdong Ocean University, China
Martijn Derks,
Wageningen University and Research, Netherlands

*CORRESPONDENCE
Željka Pezer,
zpezer@irb.hr

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 03 October 2022
ACCEPTED 15 November 2022
PUBLISHED 29 November 2022

CITATION
Pokrovac I and Pezer Ž (2022), Recent advances and current challenges in population genomics of structural variation in animals and plants. *Front. Genet.* 13:1060898. doi: 10.3389/fgene.2022.1060898

Recent advances and current challenges in population genomics of structural variation in animals and plants

Ivan Pokrovac and Željka Pezer*

Laboratory for Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute, Zagreb, Croatia

The field of population genomics has seen a surge of studies on genomic structural variation over the past two decades. These studies witnessed that structural variation is taxonomically ubiquitous and represent a dominant form of genetic variation within species. Recent advances in technology, especially the development of long-read sequencing platforms, have enabled the discovery of structural variants (SVs) in previously inaccessible genomic regions which unlocked additional structural variation for population studies and revealed that more SVs contribute to evolution than previously perceived. An increasing number of studies suggest that SVs of all types and sizes may have a large effect on phenotype and consequently major impact on rapid adaptation, population divergence, and speciation. However, the functional effect of the vast majority of SVs is unknown and the field generally lacks evidence on the phenotypic consequences of most SVs that are suggested to have adaptive potential. Non-human genomes are heavily under-represented in population-scale studies of SVs. We argue that more research on other species is needed to objectively estimate the contribution of SVs to evolution. We discuss technical challenges associated with SV detection and outline the most recent advances towards more representative reference genomes, which opens a new era in population-scale studies of structural variation.

KEYWORDS

genomic structural variation, structural variants, copy number variation, population genomics, evolution, adaptation, speciation

Changing definition and typical properties

Genetic variation is the major focus of population genetics as it provides raw material upon which evolutionary forces act to create phenotypic diversity. Over the past two decades, it has become evident that variation in the linear structure of the genome is taxonomically ubiquitous and that it affects a much larger portion of the genome than the variation in the sequence itself (Huddleston et al., 2017; Kosugi et al., 2019; Hämälä et al., 2021; Box 1). This form of genetic variation results in structural variants (SVs) that can affect orientation (inversions), position (translocations), or copy number. The latter are collectively termed copy number variants (CNVs) and include deletions, insertions, and

amplifications of a sequence. A specific group of CNVs termed presence-absence variations (PAVs) refers to sequences that exist in some genomes while completely missing in other genomes of the same species (Saxena et al., 2014). SVs were first defined as events of at least 1 kilobase pairs (kbp) in length (Feuk et al., 2006) but the definition has since expanded to encompass sizes down to 50 bp and larger (Alkan et al., 2011; Sudmant et al., 2015b). Our increasing understanding of the prevalence of SVs as major contributors to genetic variation has led to the inclusion of other genome rearrangements and elements in this definition, that were long before known to have a variable structure within population. The current definition based on SV size also includes interspersed elements (such as transposable elements; TEs), tandem repeats (including micro-, mini-, and macrosatellites) as well as aneusomy and aneuploidy (Pös et al., 2021).

Spontaneous, *de novo* SVs occur several hundred-fold less frequently than point mutations (Belyeu et al., 2021), although the mutation rate varies considerably by SV type (Collins et al., 2020). Recent large family-trio studies in humans and rhesus monkeys estimated that less than one *de novo* CNV is formed per genome per generation (Belyeu et al., 2021; Thomas et al., 2021). Interestingly, parental age does not affect the rate of these mutations in either species, in contrast to single nucleotide variants (SNVs), which accumulate with paternal age in both species (Kong et al., 2012; Wang et al., 2020). This difference between SNVs and CNVs was proposed to be due to the mechanism of their formation—CNVs are thought to form during meiosis which occurs only once per generation, whereas SNVs can arise as errors during replication in mitosis or unrepaired DNA damage—processes which occur frequently over a lifetime in the germline (Thomas et al., 2021).

Some genomic regions show an extraordinary propensity for structural variation such that they reach mutation rates hundreds and thousands of times higher than nucleotide substitutions, according to some estimates (Zhang et al., 2009). These are referred to as recurrent SVs. Their high mutability is attributable to the repetitive architecture of the genomic region in which they reside, which enables non-allelic homologous recombination (NAHR). Among all known mechanisms of SV formation, NAHR is thought to occur the most frequently, when two highly similar but non-allelic DNA sequence repeats align and crossover during meiosis, causing deletion, duplication, or inversion of the region between the repeats, depending on the orientation of the aligned sequences (Zhang et al., 2009). These mediators of NAHR are usually considered to be CNVs themselves as they exist in the genome in variable low or high copy numbers, such as segmental duplications, transposable elements, and tandem repeats. Other mechanisms of SV formation such as non-homologous end joining (NHEJ), microhomology-mediated break-induced replication (MMBIR), fork stalling and template switching (FoSTeS), and replication slippage are not dependent on high sequence similarity and create mainly non-recurrent SVs. These mechanisms and events are usually discussed in the context of genomic disorders (Hastings et al., 2009; Carvalho and Lupski, 2016), although they may contribute to natural polymorphism without seemingly negative effects.

Effect on gene expression and phenotypic variation

The high mutability of SVs is reflected in their high variability within population. For example, it is currently estimated that any human individual contains on average 16 Mb of structural variation (Ebert et al., 2021) or up to 27,000 SVs, including highly repetitive elements (Chaisson et al., 2019; see Box 1). According to the data from NCBI's database of human genomic structural variation (dbVar), almost 100,000 regions in the human genome are affected by SVs at population frequency ≥1% (Box 1). Given this abundance and high variability within population, SVs are expected to have a large impact on phenotypic variation. However, determining the functional effects of the majority of SVs is difficult, especially in natural populations which are not readily amenable to genetic manipulations (Lauer and Gresham 2019). The association of SVs with gene expression remains the most commonly used proxy for assigning phenotypic consequences. An ever-increasing number of population-scale studies have emerged to suggest that SVs of all types contribute to phenotypic variation on multiple layers of gene regulation. CNVs can alter gene dosage (Handsaker et al., 2015) and thus directly affect protein levels, as shown for the human salivary amylase gene (Perry et al., 2007). Structural variants can also modulate gene expression by re-organizing chromatin domains. Perturbations of topologically associated domains (TADs) can lead to the formation of novel regulatory modules, as shown in humans, apes, and mice (Spielmann et al., 2018; Fudenberg and Pollard, 2019; Gilbertson et al., 2022). CNVs can encompass regulatory elements, such as in the case of an enhancer that controls a gene *NDP* that is responsible for wing pigmentation in pigeons (Vickrey et al., 2018). Expression of this gene is positively correlated with both increased melanism and enhancer copy number. In crows, the same gene is associated with plumage variation but is controlled by a different SV type - an LTR retrotransposon insertion that causes reduced expression (Weissensteiner et al., 2020). SVs can affect whole regulatory networks by affecting single key transcription factors and thus have a large phenotypic effect. This was recently exemplified by a mutation in the *ENO* gene, which encodes a transcription factor that regulates floral meristem size in tomatoes - an 85-bp deletion in the promoter of *ENO* was shown to be responsible for the increase in fruit size during tomato domestication (Yuste-Lisbona et al., 2020). Copy number variation in introns causes variable gene length and is commonly found in healthy human populations. These CNVs reside inside genes with essential

functions and are proposed to be responsible for their differential regulation between individuals (Rigau et al., 2019). A recent genome-wide association study (GWAS) based on presence-absence variations in rapeseed identified PAVs among different ecotypes that altered the expression of genes responsible for flowering regulation (Song et al., 2020).

While these and other studies illustrate the contribution of individual SVs to phenotypic variation *via* gene regulation, they do not attest to the extent to which SVs explain overall variation in gene transcription within population. Several studies to date have attempted to ascertain the causality of SVs at expression quantitative trait loci (eQTLs). The most comprehensive study thus far, performed in humans and based on over 600 individuals and 48 tissues, found that SVs are causal at 2.66% of eQTLs which represents a tenfold enrichment relative to their abundance in the genome (Scott et al., 2021). This study revealed that, among all SV types, multiallelic CNVs, both coding and non-coding, have the highest association with eQTLs and that the contribution of transposable element insertions was small. Prior estimates based on a limited number of samples and tissues are in discordance with the study by Scott et al. (2021), as they found either a much larger or much smaller proportion of eQTLs to be caused by SVs. For example, a study based on 13 tissues from 147 individuals estimated up to 6.8% of eQTLs are driven by a causal SV (Chiang et al., 2017). An earlier study associated only 0.56% of eQTLs with SVs (Sudmant et al., 2015b), but it was based on a single cell line although the number of individuals was comparable to the study performed by Scott et al. (2021). This large disagreement in estimates between studies suggests that future efforts should employ a more exhaustive number of tissue types, and possibly target a variety of biological processes, to more precisely assess the contribution of SVs on gene expression in a tissue- and condition-specific manner. Indeed, genes with tissue-specific expression exhibit greater copy number variability than genes with widespread expression (Dopman and Hartl, 2007; Henrichsen et al., 2009; Keel et al., 2016), suggesting that SVs more often have roles in specialized rather than general processes. A recent study based on only two tissue types in three-spined sticklebacks found a strong positive correlation between gene copy number and expression in almost 40% of analyzed CNVs (Huang et al., 2019). Such high association becomes less surprising when one considers that gene-encompassing CNVs were previously found to be enriched for immune activity genes in sticklebacks and that the study focused on immune tissues where these genes are expected to be expressed. Another study identified thousands of tandemly repeated minisatellite sequences variable in copy number within population to be associated with local expression and DNA methylation levels (Garg et al., 2021). These CNVs were associated with genes that have been linked with human phenotypes through genome-wide association studies and were strongly enriched for regulatory elements such as enhancers and promoters, suggesting that these non-coding multiallelic CNVs may be causal for human phenotypes and have regulatory functions.

In summary, multiallelic CNVs seem to be a class of SVs that is the most strongly implicated in the contribution of SVs to variation in gene expression. However, the presented figures are likely underestimates. We can expect to approach more precise estimates with the addition of a more comprehensive set of tissues and by analyzing diverse biological conditions in future studies. Despite the large discrepancies in estimates, current knowledge collectively suggests that both coding and non-coding SVs may have a tremendous impact on gene expression, and thus affect phenotypes in the ways we are just beginning to understand. GWASs based on SNVs have not been able to completely identify the genetic components underpinning (human) traits and disorders; over the past decade, a growing body of evidence has accumulated to suggest SVs as a source of this "missing heritability" (Sudmant et al., 2015b; De Coster et al., 2021; Garg et al., 2022; Zhou et al., 2022).

Impact on evolution

Hundreds of CNVs can be found in the genomes of healthy individuals and they show strong signatures of population structure in numerous species (Sudmant et al., 2015a; Pezer et al., 2015; Xu et al., 2016). This has been used as an argument to propose that the majority of CNVs evolve under neutral evolutionary pressures, such that the patterns of copy number variation seen in populations are mainly shaped by demographic events, mutation rate, and genetic drift (Iskow et al., 2012). However, even such generalizations of the evolutionary implications of CNVs (and other SVs) should be considered in their functional contexts. Recent studies in humans and rhesus monkeys revealed that *de novo* gene deletions outnumber duplications by several times (Belyeu et al., 2021; Thomas et al., 2021), but this ratio becomes skewed over time, as illustrated by the proportion of fixed gene losses along the primate lineage, which becomes smaller (Fortna et al., 2004; Dumas et al., 2007; Sudmant et al., 2013; Thomas et al., 2021). This suggests that, over generations, purifying selection acts against deletions of complete genes. The vast majority of SVs seem to be depleted from functional regions of the genome and segregate at low frequencies, as shown by studies in different species (Pezer et al., 2015; Hämälä et al., 2021). Signals of pervasive selection against all types of SVs that overlap genes, except whole-gene duplications, have recently been discovered in a large analysis of thousands of human genomes (Collins et al., 2020). These studies collectively suggest that most SVs affecting genes are deleterious. A somewhat contrasting observation came from a recent study that suggested that SVs significantly contribute to non-neutral variation in humans (Saitou et al., 2022). Assuming that majority of SVs evolve neutrally, this study looked for SVs with unusual allele frequency distribution among

populations and came to a surprising number of over 500 putatively adaptive SVs in humans. A proportion of these included SVs that affect exons and were dominated by multiallelic CNVs.

Contribution to adaptation

Structural variants exist in extremely heterogeneous forms, in terms of type (insertion, deletion, duplication, inversion, and translocation), size, mutation rate, and genomic context. Consequently, even without technical difficulties in their discovery, they constitute a substantial challenge for evolutionary studies. While the current picture of the evolutionary effects of SVs remains incomplete, their contribution to adaptive evolution and diversification is becoming more evident (Radke and Lee, 2015; Saitou et al., 2022). An increasing number of studies suggest that SVs are involved in a variety of adaptations in a range of taxonomic groups, affecting different biological systems such as immunity, metabolism, and sensory perception. Instances of naturally occurring parallel ecological divergence provide an especially useful framework for detecting potentially adaptive SVs. The idea is that if the frequency of an SV is higher in a derived population of a certain ecotype compared to the ancestral population of a different ecotype, and this is observed repeatedly in multiple independent populations, that SV is likely contributing to the adaptive phenotype. Adaptation of marine fish to freshwater represents such a system. A study by Ishikawa et al. (2019) found that a gene involved in fatty acid desaturation was duplicated in freshwater lineages. Transgenic manipulation of this gene enabled marine lineages to produce fatty acids and survive in freshwater that lacks fatty acids. This suggested that differences in gene dosage contribute to differences in survival on fatty acid–deficient diets. In a follow-up study, additional gene duplications were identified to be associated with freshwater colonization, including genes involved in immune function and thyroid hormone metabolism (Ishikawa et al., 2022). In another study, two large chromosome inversions were identified to exhibit parallel association with freshwater adaptation (Zong et al., 2021). These inversions contained multiple genes involved in various processes such as metabolism, immunoregulation, growth, maturation, and osmoregulation, thus potentially affecting morphology, physiology and behavior. It was recently found that large inversions were common and widespread in natural populations of deer mice and several inversions with significant differences in allele frequency between forest and prairie ecotypes were identified, which likely contribute to local adaptation (Harringmeyer and Hoekstra, 2022). It has been proposed that among all SVs, chromosomal inversions are the most frequently linked to adaptive traits (reviewed in Wellenreuther and Bernatchez 2018). However, a wealth of studies suggests that CNVs may be comparable if not even dominant in this aspect. Since the initial discovery of copy number variation, more and more instances of CNVs with a putative role in local adaptation of human populations are emerging (Iskow et al., 2012; Hsieh et al., 2019; Quan et al., 2021; Saitou et al., 2022). Both deletions and duplications are implicated. For example, recurring exonic deletions in the haptoglobin gene were shown to contribute to human health by lowering cholesterol levels in the blood (Boettger et al., 2016). Copy number variations in genes *Ppd-B1* and *Vrn-A1* contribute to global adaptation of wheat to a wide range of environmental conditions (Würschum et al., 2015). These genes modulate the timing of flowering and their increase in copy number is associated with altered expression (Díaz et al., 2012). Furthermore, an increase in *EPSPS* gene copy number confers resistance to the herbicide glyphosate in different weed species (Gaines et al., 2010; Baek et al., 2021). Similarly, triplication of a gene associated with aluminum tolerance in some maize lines correlates with increased expression, which confers higher tolerance to aluminum in maize grown on acidic soils (Maron et al., 2013). Huang et al. (2019) studied the role of gene copy number in adaptation to distinct parasite environments between the lake and river habitats in sticklebacks. In some of these genes, copy number was differentiated between ecotypes and it positively correlated with transcript level, suggesting that gene dosage contributes to local adaptation by modulating expression. Similarly, specific SVs with signs of local adaptation were recently uncovered in chocolate tree, some of which are linked to genes that are also differentially expressed between populations (Hämälä et al., 2021). They were enriched for functions related to immunity, emphasizing the role of SVs in local adaptation to specific pathogens. In the fruit fly, hundreds of TEs were identified to be associated with expression variation of nearby genes, some of them bearing adaptive signatures (Rech et al., 2022). Gene loss can also produce adaptive phenotypes, as suggested for polar bear evolution, where a considerable number of genes encoding olfactory receptors have been lost, as well as the salivary amylase-encoding gene and genes involved in fatty acid metabolism (Rinker et al., 2019). These CNVs evolved rapidly over a short evolutionary period, driven by a dietary shift from omnivorous to carnivorous during polar bear evolution. Even some gene retrocopies show signatures of positive selection, as shown by recent studies in humans and mice (Schrider et al., 2013; Zhang and Tautz 2022).

Contribution to speciation

Changes in genome structure can lead to incompatibilities between populations and thus enhance speciation. SVs can enable reproductive isolation through various mechanisms such as suppressed recombination, hybrid incompatibility, and intrinsic postzygotic or premating isolation (reviewed in Zhang

et al., 2021). Inversions, especially large ones, seem to be particularly implicated in suppressing recombination. In heterozygotes for such SV, inverted region fails to pair with non-inverted allele during meiosis, preventing them from crossover. This results in both variants independently accumulating mutations in their sequences over time, creating "genomic islands of divergence," which eventually leads to incompatibilities (Zong et al., 2021). Incompatibility can also be caused by CNVs, especially if affecting the whole gene, as exemplified by a duplication of a key photosynthetic gene in the yellow monkeyflower (Zuellig and Sweigart, 2018). This variant causes lethality in naturally occurring hybrids between two closely related species, presumably by misregulated transcription. Copy number variation can also play a role in assortative mate choice, as suggested for hundreds of CNVs found to be associated with reinforcement of sexual isolation between the two European subspecies of the house mouse (North et al., 2020). Premating isolation can even be mediated by TE, as shown for the 2.25-kb LTR retrotransposon insertion which affects plumage in birds, a trait associated with prezygotic isolation through social and sexual selection (Weissensteiner et al., 2020). This study nicely illustrates how even a single and small SV can change the evolutionary trajectory of a population and potentially lead to divergence and speciation. Translocations represent a special type of SVs, in that they are often associated with genome instability and negative outcome such as infertility and oncogenesis (Aplan, 2006; Mitelman et al., 2007). Rare instances are found as naturally occurring polymorphisms in healthy individuals. One of the well-studied examples is Robertsonian fusion in the house mouse subspecies *Mus musculus domesticus*, which refers to the translocation of the whole chromosome arm, *i.e.* the joining of two telocentric chromosomes to create a metacentric chromosome. Robertsonian fusions are more frequent in small and geographically isolated populations and they are proposed to contribute to reproductive isolation (Garagna et al., 2014). Similar to inversions, translocations have been associated with the suppression of recombination and have recently been implicated in genetic divergence between subspecies of bananas (Martin et al., 2020) and populations of spiny frogs (Xia et al., 2020).

Structural variants as loci of large effect

Some SVs are large enough to span many genes and regulatory regions. Consequently, they can simultaneously affect multiple traits, acting as supergenes of large effect. Such a role has often been assigned to large inversions, proposed to be associated with complex phenotypes (reviewed in Wellenreuther and Bernatchez 2018). An inversion that contains multiple advantageous alleles will be more strongly selected for than an inversion containing a single favorable gene variant. These alleles are also more likely to be coinherited due to suppressed recombination in heterokaryotypes, contributing further to their rapid increase in frequency in the population under selection. Consequently, large inversions are considered to have significant roles in rapid environmental adaptation and speciation. SVs can also cause dramatic changes in the regulation of multiple genes by disrupting chromatin domains and exposing certain promoters to certain enhancers for the first time. It was proposed that translocations and inversions perturbed TADs and thus created differences in promoter-enhancer connections between humans and mice that are responsible for differential regulation of genes involved in immune response between the two species (Gilbertson et al., 2022). SVs are otherwise strongly depleted from TAD boundaries and active chromatin states, suggesting that they are under negative selection (Fudenberg and Pollard, 2019). Single SVs often impact the expression of multiple genes, two on average in humans (Scott et al., 2021), suggesting that they frequently exert a pleiotropic effect on phenotypic diversity. Evidence of an SV with a strong and immediate effect on phenotype came from a recent experimental evolution study on nematode. Zhao et al. (2020) studied the genetic basis of adaptation to food sources in *Caenorhabditis elegans* and found a recombinant inbred line with increased fitness. They detected a complex SV as its genetic basis; this complex rearrangement caused duplication of a gene involved in exploration behavior and modified its expression. It was proposed that the SV occurred as a single genomic instability event and became fixed in a population because it provided a fitness advantage in a new environment. These findings highlight the potential of SVs in causing dramatic structural changes in the genome which can substantially and instantaneously affect phenotypes. The majority of such large events are expected to be deleterious. However, under specific circumstances, some variants may provide a strong selective advantage which would enable them to quickly rise in frequency within the population and even become fixed over a short evolutionary time.

Population-scale studies of SVs are strongly biased toward humans

Studying genetic variation in natural populations is crucial for understanding how genomes evolve. Assessing the degree of structural variation in various species and populations contributes to our general understanding of its role in evolutionary processes. Nevertheless, population-scale studies of SVs are heavily biased toward humans (Figure 1) and insights gained mainly from studies on human populations guide our general perception of structural variation (Box 1). However, modern humans have a specific population history that involved at least one severe bottleneck followed by rapid expansion and repeated founder effect (Watkins et al., 2001; Amos and Hoffman, 2010), which resulted in substantially lower

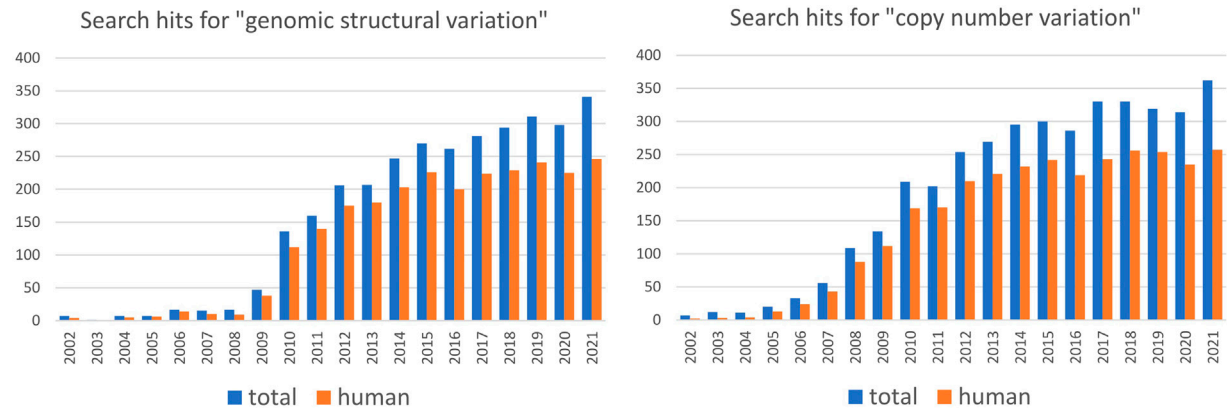

FIGURE 1
Number of articles per year containing indicated search term in PubMed in the period from 2002 to 2021. Search is performed for the combination of indicated term AND "population" (data in blue) or for indicated term AND "population" AND "human" (data in red). The data suggest that the majority of population-scale SV and CNV studies are based on humans.

genetic diversity compared to many other species. Moreover, genetic boundaries between human populations are often blurry, reflecting the frequent population movement and admixture. Humans are also characterized by a small effective population size (Tenesa et al., 2007; Park 2011), which is known to reduce the efficacy of natural selection and increase the influence of genetic drift. Thus, human populations by no means embody a "typical" evolutionary trajectory and more studies of SVs in non-human populations are needed to disentangle the roles that SVs play in evolution and ecological specialization. Based on growing evidence, SVs may be the key players of rapid adaptation to changing environments and naturally occurring examples of parallel evolution represent excellent opportunities to study the genetic architecture of rapid adaptation, such as adaptation to freshwater discussed above. The independence of studied populations that converged adaptive traits is desirable: the stronger the evidence that they independently evolved similar traits under the same selective pressure, the stronger the association with the underlying variant. Population studies in non-human species may provide more instances of such independent, parallel evolution as a framework for studying the role of SVs in adaptation and speciation. For instance, adaptation to the subterranean environment has been documented for many taxa, yet the impact of structural variation in this context is still unexplored. Numerous examples of parallel evolution can also be found in domesticated species, and evidence of SVs playing a part in trait evolution during domestication in plants and animals is emerging. For example, white coat color was independently selected for in sheep and goats - in both species, this trait is associated with duplication of the agouti signaling protein (*ASIP*) gene (Norris and Whan, 2008; Fontanesi et al., 2009). In plants, the loss of seed shattering was repeatedly selected for during domestication and is often associated with a deletion in gene *Sh1* in different cereal species (Lin et al., 2012; Choi et al., 2019). From an evolutionary point of view, domestication is a very specific process that usually involves a population bottleneck that substantially decreases genetic diversity and increases the frequency of domestication alleles (Gaut et al., 2018). It has been proposed that, at least in plants, deletions underly some of the crucial domestication traits, whereas during later stages of domestication (*i.e.* during diversification) various SV types facilitate local adaptation (Gaut et al., 2018; Lye and Purugganan, 2019). Hence, although they may provide some interesting examples of parallel evolution, domesticated species may not represent a general model for studying the role of SVs in evolution.

Challenges in the detection of structural variants

There is no doubt that sequencing technology based on short reads has tremendously advanced our knowledge of the prevalence of structural variation in populations and its impact on health and evolution over the past two decades. Numerous algorithms and approaches have been designed and employed to detect structural variants from short-read sequencing (SRS) data (reviewed in Kosugi et al., 2019). While many of them represent an improvement in some specific aspect, they all suffer from three basic problems, associated with technical limitations inherent to short reads. First, no single algorithm can detect SVs of all types and sizes. As shown by an exhaustive study that compared the performance of 69 existing algorithms for SV detection from WGS data, most algorithms

perform best for particular SV types and, in some cases, for particular size ranges (Kosugi et al., 2019). Even when the same approach and the same algorithm is used, substantial differences may exist between samples that are not due to biological differences. For example, in the read-depth approach, lower coverage will lead to fewer SVs being identified, the power to detect smaller events will be compromised and neighboring SVs may collapse into single calls due to diminished resolution (Pezer et al., 2015). These problems make comparisons between studies difficult, as each approach applied to the same biological sample will result in a different set of SVs.

Second, the true positive rate of SRS-based methods is generally low while the false positive rates can be as high as 90%; again, both are heavily dependent on the size and type of SVs (Mahmoud et al., 2019). Differences in the processing of samples and data before SV calling can also strongly affect the accuracy of the final call set. For example, Khayat et al. (2021) found that sequencing centers and especially read mapping methods contribute significantly to variability between call sets. In particular, their results suggest that one-fifth of all calls represent false positives that are solely contributed by the mapper. These problems have major consequences on reproducibility and can greatly affect the interpretation between studies.

Third, methods based on SRS are unable to (accurately and reliably) identify SVs in repetitive genomic regions, stemming from the uncertainty of the true origin of reads that can be equally well mapped to multiple genomic positions. As a consequence, these problematic, repetitive regions are often omitted in genomic analyses. However, recent analyses based on long-read sequencing (LRS) technologies suggest that these regions may be the greatest source of variation. In human genomes, up to 90% of SVs (mostly smaller than 1 kbp) detected from LRS data were unknown from previous SRS-based analyses (Chaisson et al., 2015; Huddleston et al., 2017; Audano et al., 2019; Ebert et al., 2021; Quan et al., 2021). This means that SRS-based methods are blind to the vast majority of variation. This problem is particularly relevant in analyses of genomes with high repetitive DNA content such as in many plant species.

Despite its power to detect variation that is inaccessible to SRS, LRS has several drawbacks which directly limit its use in large population studies: it is more expensive, requires more input DNA, and has lower sample throughput than SRS (Ho et al., 2020). Consequently, not many population genomic studies based on long reads have emerged so far (Audano et al., 2019; Weissensteiner et al., 2020; Beyter et al., 2021; Quan et al., 2021; Yan et al., 2021; Rech et al., 2022). Majority of these studies employ a hybrid strategy which involves sequencing a smaller number of genomes by using long reads while the remaining samples are sequenced with short-read technology (Ho et al., 2020; De Coster et al., 2021; Quan et al., 2022). Structural variants identified by LRS in representative genomes can then be genotyped from SRS data in all other samples. This approach combines the advantages of both read-sequencing technologies: the power of LRS to discover multiple types and a wider size range of SVs (Quan et al., 2022), and the generally high genotyping precision of SRS-based algorithms (Kosugi et al., 2019). Even so, not all SVs that are detected from long reads can be accurately genotyped from short-read data, and as much as half remain invisible to it (Huddleston et al., 2017; Chakraborty et al., 2019; Ebert et al., 2021). Furthermore, LRS produces reads that are still insufficiently long to resolve all SVs. For instance, detection algorithms based on long reads that consider information on soft clipped reads and intra-read discordance are much worse at discovering CNVs larger than >100 kbp than are algorithms based on the read-depth approach from short reads (Kosugi et al., 2019). Hence, despite the advances made related to improved identification of smaller SVs by long reads, much of the most complex genomic regions remains inaccessible. Optical mapping is a technology of choice for resolving such regions as it generates molecules that can be over 1 Mb long and can therefore bridge larger repetitive regions (Ho et al., 2020). It has been successfully applied in some population studies which resolved previously undetected large SVs and identified novel genome content not found in the reference genome sequence (Levy-Sakin et al., 2019; Weissensteiner et al., 2020). However, optical mapping has several weak points, such as a high error rate, a lack of information on the actual sequence underlying the molecules, and the inherent inability to determine precise SV breakpoints. The widespread use of optical mapping is further hindered by lower throughput and the lack of alternative and publicly available tools for SV detection (see Li et al., 2017; Raeisi Dehkordi et al., 2021). Another promising technology that has the potential to detect large SVs and those in repetitive regions is high-throughput chromosome conformation capture (Hi-C). Hi-C is typically used for studying 3D genome interactions, and although several tools have been developed for SV discovery from Hi-C data, these are specifically designed for human genomes and are limited to the detection of SVs larger than 1 Mbp. Most recently, a framework named EagleC was developed that has the power to detect events down to 1 kb in any species genome, providing sufficient coverage (Wang et al., 2022b). This tool illustrates the potential of Hi-C application in SV discovery from large sample sets, and further developments in this direction will enable widespread and more comprehensive population-scale studies of SVs by use of Hi-C technology.

Towards more representative reference genome

In population studies, structural variants are most commonly detected from sequencing data by aligning reads to the reference genome sequence and identifying patterns of discordance in alignment. If the reference genome is contiguous, an average

Box 1 SVs in numbers
- 92,934 common structural variant regions in human populations; according to the NCBI Curated Common Structural Variants dataset (dbVar study accession nstd186; Lappalainen et al., 2013)
- 27,662 SVs detected per person, including STRs and other highly repetitive elements (Chaisson et al., 2019)
- 16 Mbp—The average amount of structural variation per person (Ebert et al., 2021)
- 3–15X—More base pairs are affected by SVs than by SNVs (Pang et al., 2010; Huddleston et al., 2017; Hämälä et al., 2021)
- 3–10X—Higher inter-individual genomic difference at SVs than at SNVs (Pang et al., 2010; Sudmant et al., 2015b)
- 4.8%–9.5% of the human genome is affected by CNVs (Zarrei et al., 2015)
- 0.29—Number of *de novo* SVs per generation (in regions of the genome accessible to short-read sequencing) or one new SV every two to eight live births (Collins et al., 2020; Belyeu et al., 2021)
- 6.8%—the largest estimated proportion of eQTLs caused by SVs (Chiang et al., 2017)

read depth of 10x is considered sufficient for population-scale comparisons (Collins et al., 2020). However, reference genomes assembled at the chromosome level are rarely available, which hampers studies in the majority of species. Even human genomes seem to contain large regions not present in the reference genome, as shown by studies based on optical mapping and long reads (Audano et al., 2019; Levy-Sakin et al., 2019; Ebert et al., 2021). They are not merely repetitive and non-functional, but also encompass genes and regulatory elements. These studies question the completeness and the representativeness of the human reference genome. The latest version of the human reference assembly, T2T-CHM13, succeeded in closing all gaps found in the previous GRCh38 assembly and indeed represents the first completely sequenced genome (Nurk et al., 2022). However, similar to the GRCh38, in which the majority of sequence originates from a single individual (Ballouz et al., 2019), T2T-CHM13 represents only one haplotype, and while it improves analysis of human genetic variation to some extent (Aganezov et al., 2022), it cannot fully capture the genetic diversity among populations. Approaches to remove reference bias have started to emerge, to improve accuracy in population-scale SV analyses. In 2019, Sherman et al. (2019), sequenced 910 individuals of African descent and used all unaligned reads to assemble contigs *de novo*. These collectively constituted 300 million base pairs of sequences that were missing from the reference genome and illustrated that a single reference genome is suboptimal for population-based studies. Instead, the creation of a comprehensive pan-genome was proposed, based on all distinct human populations that would much better capture all the DNA present in humans. In 2019, the Human Pangenome Project was initiated, funded by the US National Human Genome Research Institute (NHGRI), with a goal to provide a more accurate and diverse representation of global genomic variation through the creation of a more sophisticated human reference genome (Wang. et al., 2022a; Khamsi, 2022).

Pangenomes are superior to single reference genomes because they combine genomes from multiple individuals and thus better incorporate genomic polymorphism within a population, and they are becoming increasingly used for SV studies in humans and other species (Beyter et al., 2021; Ebert et al., 2021; Qin et al., 2021; Yan et al., 2021; Zhou et al., 2022; for a list of studies based on plant pan-genomes see Yuan et al., 2021). Instead of being represented as a linear sequence, pangenomes are constructed as graphs to which sequencing reads are aligned (De Coster et al., 2021; Quan et al., 2022), enabling reliable genotyping of SVs by short reads in thousands of samples, which facilitates large population studies. However, approaches for graph-based genotyping are in their infancy, and tools for more efficient construction of complex graphs and alignment of reads to graphs are still under development (Quan et al., 2022).

The power of haplotype-resolved genomes

One of the major obstacles to a deeper understanding of SVs is the inability to accurately determine discrete SV alleles as it hinders evolutionary and population genetic studies of SVs, including analyses of allele frequency, estimations of the rates of recurrent mutation and incorporation of SVs in genome-wide association studies (Ebert et al., 2021; Saitou et al., 2022). This limitation can be overcome by resolving haplotypes. Studies that analyze haplotype-resolved genomes readily identify a substantial number of previously undetected SVs and additional genomic content not present in the reference genome (Huddleston et al., 2017; Wong et al., 2018; Chaisson et al., 2019; Levy-Sakin et al., 2019; Almarri et al., 2020; Ebert et al., 2021; Hämälä et al., 2021). Huddleston et al. (2017) sequenced genomes of two hydatiform moles - which are haploid; when they merged the two haploid genomes *in silico* to create an artificial diploid genome, over half of the heterozygous SVs were no longer detected from long-read sequencing data. This showed that the majority of SVs are not detectable unless the haplotype structure of the genomes is known and illustrates the importance of haploid resolution for the sensitivity of SV detection. However, determining the physical haplotype structure of genomes is yet not widely affordable and the haplotype-phasing methods are

still immature, preventing their wider application in population-scale studies.

Variant interpretation

Over the past two decades, a sheer abundance of studies has demonstrated that structural variants are by far the most dominant form of genetic variation. While our ability to detect SVs has increased tremendously, for the largest part we are still unable to explain the functional consequences (Ho et al., 2020; Yan et al., 2021). For example, in the majority of population-scale studies, evidence on the adaptive role of SVs is inferred from associations between SV frequency and environmental/behavioral traits, however, rare studies provide evidence based on phenotypic assays, such as gene expression and protein level, or fitness (Perry et al., 2007; Maron et al., 2013; Ishikawa et al., 2019; Zhao et al., 2020). Experimental evolution provides a powerful means to study adaptation, yet it is limited to species with short generation times such as single-cell organisms. In more complex, multicellular organisms, it was proposed that integrating SVs with layered biological data is crucial for a more complete understanding of the impact of SVs (Ho et al., 2020). These may include but are not limited to analyses of transcriptome, epigenome, proteome, and 3D chromatin structure.

Concluding remarks

Recent years have shown that genome plasticity is even larger than it was anticipated more than 10 years ago. Long-read sequencing technologies have enabled the discovery of a wealth of structural variation in previously inaccessible genomic regions and continuous efforts provide increasing evidence that SVs play important roles in population divergence, local adaptation, and speciation. However, there is currently no approach that would allow simultaneous detection of all SVs, and even methods based on long reads fail in complex genomic regions such as long tandemly repetitive sequences and segmental duplications. Therefore, systemic assessments of SVs' contribution to evolution are primarily hindered by the high cost of analyzing a large number of individuals to enable population-scale studies, and by the necessity to employ multiple available technologies, in order to capture all types of SVs and achieve greater resolution of SV detection. Without a such comprehensive approach, the investigations are limited to SVs of a particular size range or types. Pangenome assemblies provide a route to avoid costly sequencing by long reads and enable genotyping from short reads mapped on a reference genome that is derived from several individuals, representative of multiple populations. Investment in efforts to construct pangenomes in a multitude of species will enable more reliable and comprehensive SV detection and genotyping on a larger scale. The number of detected SVs is expected to increase further with the improvement of haplotype-phasing methods, and the wider application of such methods is expected to greatly advance our understanding of the impact of SVs on evolution.

Author contributions

ZP conceptualized the manuscript. IP and ZP wrote the manuscript.

Funding

This work was supported by the Croatian Science Foundation under Grant UIP-2019-04-7898.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Aganezov, S., Yan, S. M., Soto, D. C., Kirsche, M., Zarate, S., Avdeyev, P., et al. (2022). A complete reference genome improves analysis of human genetic variation. *Science* 376, eabl3533. doi:10.1126/science.abl3533

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12, 363–376. doi:10.1038/nrg2958

Almarri, M. A., Bergström, A., Prado-Martinez, J., Yang, F., Fu, B., Dunham, A. S., et al. (2020). Population structure, stratification, and introgression of human structural variation. *Cell* 182, 189–199. doi:10.1016/j.cell.2020.05.024

Amos, W., and Hoffman, J. I. (2010). Evidence that two main bottleneck events shaped modern human genetic diversity. *Proc. Biol. Sci.* 277, 131–137. doi:10.1098/rspb.2009.1473

Aplan, P. D. (2006). Causes of oncogenic chromosomal translocation. *Trends Genet.* 22, 46–55. doi:10.1016/j.tig.2005.10.002

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., et al. (2019). Characterizing the major structural variant alleles of the human genome. *Cell* 176, 663–675. e19. doi:10.1016/j.cell.2018.12.019

Baek, Y., Bobadilla, L. K., Giacomini, D. A., Montgomery, J. S., Murphy, B. P., and Tranel, P. J. (2021). Evolution of glyphosate-resistant weeds. *Rev. Environ. Contam. Toxicol.* 255, 93–128. doi:10.1007/398_2020_55

Ballouz, S., Dobin, A., and Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biol.* 20, 159. doi:10.1186/s13059-019-1774-4

Belyeu, J. R., Brand, H., Wang, H., Zhao, X., Pedersen, B. S., Feusier, J., et al. (2021). 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. doi:10.1016/j.ajhg.2021.02.012

Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., et al. (2021). 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. doi:10.1038/s41588-021-00865-4

Boettger, L. M., Salem, R. M., Handsaker, R. E., Peloso, G. M., Kathiresan, S., Hirschhorn, J. N., et al. (2016). Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. *Nat. Genet.* 48, 359–366. doi:10.1038/ng.3510

Carvalho, C. M., and Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* 17, 224–238. doi:10.1038/nrg.2015.25

Chaisson, M. J., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 517, 608–611. doi:10.1038/nature13907

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* 10, 1784. doi:10.1038/s41467-018-08148-z

Chakraborty, M., Emerson, J. J., Macdonald, S. J., and Long, A. D. (2019). Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* 10, 4872. doi:10.1038/s41467-019-12884-1

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., et al. (2017). The impact of structural variation on human gene expression. *Nat. Genet.* 49, 692–699. doi:10.1038/ng.3834

Choi, J. Y., Zaidem, M., Gutaker, R., Dorph, K., Singh, R. K., and Purugganan, M. D. (2019). The complex geography of domestication of the African rice Oryza glaberrima. *PLoS Genet.* 15, e1007414. doi:10.1371/journal.pgen.1007414

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., et al. (2020). A structural variation reference for medical and population genetics. *Nature* 581, 444–451. doi:10.1038/s41586-020-2287-8

De Coster, W., Weissensteiner, M. H., and Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. *Nat. Rev. Genet.* 22, 572–587. doi:10.1038/s41576-021-00367-3

Díaz, A., Zikhali, M., Turner, A. S., Isaac, P., and Laurie, D. A. (2012). Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). *PLoS One* 7, e33234. doi:10.1371/journal.pone.0033234

Dopman, E. B., and Hartl, D. L. (2007). A portrait of copy-number polymorphism in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19920–19925. doi:10.1073/pnas.0709888104

Dumas, L., Kim, Y. H., Karimpour-Fard, A., Cox, M., Hopkins, J., Pollack, J. R., et al. (2007). Gene copy number variation spanning 60 million years of human and primate evolution. *Genome Res.* 17, 1266–1277. doi:10.1101/gr.6557307

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372, eabf7117. doi:10.1126/science.abf7117

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nat. Rev. Genet.* 7, 85–97. doi:10.1038/nrg1767

Fontanesi, L., Beretti, F., Riggio, V., GómezGonzález, E., Dall'Olio, S., Davoli, R., et al. (2009). Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors. *Cytogenet. Genome Res.* 126, 333–347. doi:10.1159/000268089

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., et al. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* 2, E207. doi:10.1371/journal.pbio.0020207

Fudenberg, G., and Pollard, K. S. (2019). Chromatin features constrain structural variation across evolutionary timescales. *Proc. Natl. Acad. Sci. U. S. A.* 116, 2175–2180. doi:10.1073/pnas.1808631116

Gaines, T. A., Zhang, W., Wang, D., Bukun, B., Chisholm, S. T., Shaner, D. L., et al. (2010). Gene amplification confers glyphosate resistance in Amaranthus palmeri. *Proc. Natl. Acad. Sci. U. S. A.* 107, 1029–1034. doi:10.1073/pnas.0906649107

Garagna, S., Page, J., Fernandez-Donoso, R., Zuccotti, M., and Searle, J. B. (2014). The robertsonian phenomenon in the house mouse: Mutation, meiosis and speciation. *Chromosoma* 123, 529–544. doi:10.1007/s00412-014-0477-6

Garg, P., Jadhav, B., Lee, W., Rodriguez, O. L., Martin-Trujillo, A., and Sharp, A. J. (2022). A phenome-wide association study identifies effects of copy-number variation of VNTRs and multicopy genes on multiple human traits. *Am. J. Hum. Genet.* 109, 1065–1076. doi:10.1016/j.ajhg.2022.04.016

Garg, P., Martin-Trujillo, A., Rodriguez, O. L., Gies, S. J., Hadelia, E., Jadhav, B., et al. (2021). Pervasive cis effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression. *Am. J. Hum. Genet.* 108, 809–824. doi:10.1016/j.ajhg.2021.03.016

Gaut, B. S., Seymour, D. K., Liu, Q., and Zhou, Y. (2018). Demography and its effects on genomic variation in crop domestication. *Nat. Plants* 4, 512–520. doi:10.1038/s41477-018-0210-1

Gilbertson, S. E., Walter, H. C., Gardner, K., Wren, S. N., Vahedi, G., and Weinmann, A. S. (2022). Topologically associating domains are disrupted by evolutionary genome rearrangements forming species-specific enhancer connections in mice and humans. *Cell Rep.* 39, 110769. doi:10.1016/j.celrep.2022.110769

Hämälä, T., Wafula, E. K., Guiltinan, M. J., Ralph, P. E., de Pamphilis, C. W., and Tiffin, P. (2021). Genomic structural variants constrain and facilitate adaptation in natural populations of Theobroma cacao, the chocolate tree. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2102914118. doi:10.1073/pnas.2102914118

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M., et al. (2015). Large multiallelic copy number variations in humans. *Nat. Genet.* 47, 296–303. doi:10.1038/ng.3200

Harringmeyer, O. S., and Hoekstra, H. E. (2022). Chromosomal inversion polymorphisms shape the genomic landscape of deer mice. *Nat. Ecol. Evol.* doi:10.1038/s41559-022-01890-0

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., and Ira, G. (2009). Mechanisms of change in gene copy number. *Nat. Rev. Genet.* 10, 551–564. doi:10.1038/nrg2593

Henrichsen, C. N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F., et al. (2009). Segmental copy number variation shapes tissue transcriptomes. *Nat. Genet.* 41, 424–429. doi:10.1038/ng.345

Ho, S. S., Urban, A. E., and Mills, R. E. (2020). Structural variation in the sequencing era. *Nat. Rev. Genet.* 21, 171–189. doi:10.1038/s41576-019-0180-9

Hsieh, P., Vollger, M. R., Dang, V., Porubsky, D., Baker, C., Cantsilieris, S., et al. (2019). Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* 366, eaax2083. doi:10.1126/science.aax2083

Huang, Y., Feulner, P. G. D., Eizaguirre, C., Lenz, T. L., Bornberg-Bauer, E., Milinski, M., et al. (2019). Genome-wide genotype-expression relationships reveal both copy number and single nucleotide differentiation contribute to differential gene expression between stickleback ecotypes. *Genome Biol. Evol.* 11, 2344–2359. doi:10.1093/gbe/evz148

Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., et al. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* 27, 677–685. doi:10.1101/gr.214007.116

Ishikawa, A., Kabeya, N., Ikeya, K., Kakioka, R., Cech, J. N., Osada, N., et al. (2019). A key metabolic gene for recurrent freshwater colonization and radiation in fishes. *Science* 364, 886–889. doi:10.1126/science.aau5656

Ishikawa, A., Yamanouchi, S., Iwasaki, W., and Kitano, J. (2022). Convergent copy number increase of genes associated with freshwater colonization in fishes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 377, 20200509. doi:10.1098/rstb.2020.0509

Iskow, R. C., Gokcumen, O., and Lee, C. (2012). Exploring the role of copy number variants in human adaptation. *Trends Genet.* 28, 245–257. doi:10.1016/j.tig.2012.03.002

Keel, B. N., Lindholm-Perry, A. K., and Snelling, W. M. (2016). Evolutionary and functional features of copy number variation in the cattle genome. *Front. Genet.* 7, 207. doi:10.3389/fgene.2016.00207

Khamsi, R. (2022). A more-inclusive genome project aims to capture all of human diversity. *Nature* 603, 378–381. doi:10.1038/d41586-022-00726-y

Khayat, M. M., Sahraeian, S. M. E., Zarate, S., Carroll, A., Hong, H., Pan, B., et al. (2021). Hidden biases in germline structural variant detection. *Genome Biol.* 22, 347. doi:10.1186/s13059-021-02558-x

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488, 471–475. doi:10.1038/nature11396

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* 20, 117. doi:10.1186/s13059-019-1720-5

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., et al. (2013). DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Res.* 41, D936–D941. doi:10.1093/nar/gks1213

Lauer, S., and Gresham, D. (2019). An evolving view of copy number variants. *Curr. Genet.* 65, 1287–1295. doi:10.1007/s00294-019-00980-0

Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A. K. Y., McCaffrey, J., et al. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* 10, 1025. doi:10.1038/s41467-019-08992-7

Li, L., Leung, A. K., Kwok, T. P., Lai, Y. Y. Y., Pang, I. K., Chung, G. T., et al. (2017). OMSV enables accurate and comprehensive identification of large structural variations from nanochannel-based single-molecule optical maps. *Genome Biol.* 18, 230. doi:10.1186/s13059-017-1356-2

Lin, Z., Li, X., Shannon, L. M., Yeh, C. T., Wang, M. L., Bai, G., et al. (2012). Parallel domestication of the Shattering1 genes in cereals. *Nat. Genet.* 44, 720–724. doi:10.1038/ng.2281

Lye, Z. N., and Purugganan, M. D. (2019). Copy number variation in domestication. *Trends Plant Sci.* 24, 352–365. doi:10.1016/j.tplants.2019.01.003

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biol.* 20, 246. doi:10.1186/s13059-019-1828-7

Maron, L. G., Guimarães, C. T., Kirst, M., Albert, P. S., Birchler, J. A., Bradbury, P. J., et al. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl. Acad. Sci. U. S. A.* 110, 5241–5246. doi:10.1073/pnas.1220766110

Martin, G., Baurens, F. C., Hervouet, C., Salmon, F., Delos, J. M., Labadie, K., et al. (2020). Chromosome reciprocal translocations have accompanied subspecies evolution in bananas. *Plant J.* 104, 1698–1711. doi:10.1111/tpj.15031

Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer* 7, 233–245. doi:10.1038/nrc2091

Norris, B. J., and Whan, V. A. (2008). A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. *Genome Res.* 18, 1282–1293. doi:10.1101/gr.072090.107

North, H. L., Caminade, P., Severac, D., Belkhir, K., and Smadja, C. M. (2020). The role of copy-number variation in the reinforcement of sexual isolation between the two European subspecies of the house mouse. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375, 20190540. doi:10.1098/rstb.2019.0540

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., et al. (2022). The complete sequence of a human genome. *Science* 376, 44–53. doi:10.1126/science.abj6987

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* 11, R52. doi:10.1186/gb-2010-11-5-r52

Park, L. (2011). Effective population size of current human population. *Genet. Res.* 93, 105–114. doi:10.1017/S0016672310000558

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* 39, 1256–1260. doi:10.1038/ng2123

Pezer, Ž., Harr, B., Teschke, M., Babiker, H., and Tautz, D. (2015). Divergence patterns of genic copy number variation in natural populations of the house mouse (*Mus musculus* domesticus) reveal three conserved genes with major population-specific expansions. *Genome Res.* 25, 1114–1124. doi:10.1101/gr.187187.114

Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B., et al. (2021). DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomed. J.* 44, 548–559. doi:10.1016/j.bj.2021.02.003

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184, 3542–3558.e16. doi:10.1016/j.cell.2021.04.046

Quan, C., Li, Y., Liu, X., Wang, Y., Ping, J., Lu, Y., et al. (2021). Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. *Genome Biol.* 22, 159. doi:10.1186/s13059-021-02382-3

Quan, C., Lu, H., Lu, Y., and Zhou, G. (2022). Population-scale genotyping of structural variation in the era of long-read sequencing. *Comput. Struct. Biotechnol. J.* 20, 2639–2647. doi:10.1016/j.csbj.2022.05.047

Radke, D. W., and Lee, C. (2015). Adaptive potential of genomic structural variation in human and mammalian evolution. *Brief. Funct. Genomics* 14, 358–368. doi:10.1093/bfgp/elv019

Raeisi Dehkordi, S., Luebeck, J., and Bafna, V. (2021). FaNDOM: Fast nested distance-based seeding of optical maps. *Patterns* 2, 100248. doi:10.1016/j.patter.2021.100248

Rech, G. E., Radío, S., Guirao-Rico, S., Aguilera, L., Horvath, V., Green, L., et al. (2022). Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in Drosophila. *Nat. Commun.* 13, 1948. doi:10.1038/s41467-022-29518-8

Rigau, M., Juan, D., Valencia, A., and Rico, D. (2019). Intronic CNVs and gene expression variation in human populations. *PLoS Genet.* 15, e1007902. doi:10.1371/journal.pgen.1007902

Rinker, D. C., Specian, N. K., Zhao, S., and Gibbons, J. G. (2019). Polar bear evolution is marked by rapid changes in gene copy number in response to dietary shift. *Proc. Natl. Acad. Sci. U. S. A.* 116, 13446–13451. doi:10.1073/pnas.1901093116

Saitou, M., Masuda, N., and Gokcumen, O. (2022). Similarity-based analysis of allele frequency distribution among multiple populations identifies adaptive genomic structural variants. *Mol. Biol. Evol.* 39, msab313. doi:10.1093/molbev/msab313

Saxena, R. K., Edwards, D., and Varshney, R. K. (2014). Structural variations in plant genomes. *Brief. Funct. Genomics* 13, 296–307. doi:10.1093/bfgp/elu016

Schrider, D. R., Navarro, F. C., Galante, P. A., Parmigiani, R. B., Camargo, A. A., Hahn, M. W., et al. (2013). Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet.* 9, e1003242. doi:10.1371/journal.pgen.1003242

Scott, A. J., Chiang, C., and Hall, I. M. (2021). Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res.* 31, 2249–2257. doi:10.1101/gr.275488.121

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* 51, 30–35. doi:10.1038/s41588-018-0273-y

Song, J. M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., et al. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nat. Plants* 6, 34–45. doi:10.1038/s41477-019-0577-7

Spielmann, M., Lupiáñez, D. G., and Mundlos, S. (2018). Structural variation in the 3D genome. *Nat. Rev. Genet.* 19, 453–467. doi:10.1038/s41576-018-0007-0

Sudmant, P. H., Huddleston, J., Catacchio, C. R., Malig, M., Hillier, L. W., Baker, C., et al. (2013). Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* 23, 1373–1382. doi:10.1101/gr.158543.113

Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., et al. (2015a). Global diversity, population stratification, and selection of human copy-number variation. *Science.* 349. aab3761. doi:10.1126/science.aab3761

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015b). An integrated map of structural variation in 2, 504 human genomes. *Nature* 526, 75–81. doi:10.1038/nature15394

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E., et al. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* 17, 520–526. doi:10.1101/gr.6023607

Thomas, G. W. C., Wang, R. J., Nguyen, J., AlanHarris, R., Raveendran, M., Rogers, J., et al. (2021). Origins and long-term patterns of copy-number variation in rhesus macaques. *Mol. Biol. Evol.* 38, 1460–1471. doi:10.1093/molbev/msaa303

Vickrey, A. I., Bruders, R., Kronenberg, Z., Mackey, E., Bohlender, R. J., Maclary, E. T., et al. (2018). Introgression of regulatory alleles and a missense coding mutation drive plumage pattern diversity in the rock pigeon. *Elife* 7, e34803. doi:10.7554/eLife.34803

Wang, R. J., Thomas, G. W. C., Raveendran, M., Harris, R. A., Doddapaneni, H., Muzny, D. M., et al. (2020). Paternal age in rhesus macaques is positively associated with germline mutation accumulation but not with measures of offspring sociability. *Genome Res.* 30, 826–834. doi:10.1101/gr.255174.119

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., et al. (2022a). The human pangenome project: A global resource to map genomic diversity. *Nature* 604, 437–446. doi:10.1038/s41586-022-04601-8

Wang, X., Luan, Y., and Yue, F. (2022b). EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. *Sci. Adv.* 8, eabn9215. doi:10.1126/sciadv.abn9215

Watkins, W. S., Ricker, C. E., Bamshad, M. J., Carroll, M. L., Nguyen, S. V., Batzer, M. A., et al. (2001). Patterns of ancestral human diversity: An analysis of alu-insertion and restriction-site polymorphisms. *Am. J. Hum. Genet.* 68, 738–752. doi:10.1086/318793

Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K. J., Knief, U., Heim, W., et al. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nat. Commun.* 11, 3403. doi:10.1038/s41467-020-17195-4

Wellenreuther, M., and Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* 33, 427–440. doi:10.1016/j.tree.2018.04.002

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

Würschum, T., Boeven, P. H., Langer, S. M., Longin, C. F., and Leiser, W. L. (2015). Multiply to conquer: Copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. *BMC Genet.* 16, 96. doi:10.1186/s12863-015-0258-0

Xia, Y., Yuan, X., Luo, W., Yuan, S., and Zeng, X. (2020). The origin and evolution of chromosomal reciprocal translocation in quasipaa boulengeri (Anura, dicroglossidae). *Front. Genet.* 10, 1364. doi:10.3389/fgene.2019.01364

Xu, L., Hou, Y., Bickhart, D. M., Zhou, Y., Hayel, H. A., Song, J., et al. (2016). Population-genetic properties of differentiated copy number variations in cattle. *Sci. Rep.* 6, 23161. doi:10.1038/srep23161

Yan, S. M., Sherman, R. M., Taylor, D. J., Nair, D. R., Bortvin, A. N., Schatz, M. C., et al. (2021). Local adaptation and archaic introgression shape global diversity at human structural variant loci. *Elife* 10, e67615. doi:10.7554/eLife.67615

Yuan, Y., Bayer, P. E., Batley, J., and Edwards, D. (2021). Current status of structural variation studies in plants. *Plant Biotechnol. J.* 19, 2153–2163. doi:10.1111/pbi.13646

Yuste-Lisbona, F. J., Fernández-Lozano, A., Pineda, B., Bretones, S., Ortíz-Atienza, A., García-Sogo, B., et al. (2020). ENO regulates tomato fruit size through the floral meristem development network. *Proc. Natl. Acad. Sci. U. S. A.* 117, 8187–8195. doi:10.1073/pnas.1913688117

Zarrei, M., MacDonald, J. R., Merico, D., and Scherer, S. W. (2015). A copy number variation map of the human genome. *Nat. Rev. Genet.* 16, 172–183. doi:10.1038/nrg3871

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10, 451–481. doi:10.1146/annurev.genom.9.081307.164217

Zhang, L., Reifová, R., Halenková, Z., and Gompert, Z. (2021). How important are structural variants for speciation? *Genes* 12, 1084. doi:10.3390/genes12071084

Zhang, W., and Tautz, D. (2022). Tracing the origin and evolutionary fate of recent gene retrocopies in natural populations of the house mouse. *Mol. Biol. Evol.* 39, msab360. doi:10.1093/molbev/msab360

Zhao, Y., Long, L., Wan, J., Biliya, S., Brady, S. C., Lee, D., et al. (2020). A spontaneous complex structural variant in rcan-1 increases exploratory behavior and laboratory fitness of *Caenorhabditis elegans*. *PLoS Genet.* 16, e1008606. doi:10.1371/journal.pgen.1008606

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* 606, 527–534. doi:10.1038/s41586-022-04808-9

Zong, S. B., Li, Y. L., and Liu, J. X. (2021). Genomic architecture of rapid parallel adaptation to fresh water in a wild fish. *Mol. Biol. Evol.* 38, 1317–1329. doi:10.1093/molbev/msaa290

Zuellig, M. P., and Sweigart, A. L. (2018). Gene duplicates cause hybrid lethality between sympatric species of Mimulus. *PLoS Genet.* 14, e1007130. doi:10.1371/journal.pgen.1007130