Human Molecular Genetics, 2021, Vol. 30, No. 20 R174–R186

https://doi.org/10.1093/hmg/ddab209
Advance Access Publication Date: 22 July 2021
Invited Review Article

INVITED REVIEW ARTICLE

Discovery of genomic variation across a generation

Brett Trost[1,†], Livia O. Loureiro[1,‡] and Stephen W. Scherer[1,2,*,¶]

[1]The Centre for Applied Genomics and Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON M5G 0A4, Canada and [2]McLaughlin Centre and Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada

*To whom correspondence should be addressed. Tel: +1 4168137613; Fax: +1 4168138319; Email: stephen.scherer@sickkids.ca

Abstract

Over the past 30 years (the timespan of a generation), advances in genomics technologies have revealed tremendous and unexpected variation in the human genome and have provided increasingly accurate answers to long-standing questions of how much genetic variation exists in human populations and to what degree the DNA complement changes between parents and offspring. Tracking the characteristics of these inherited and spontaneous (or *de novo*) variations has been the basis of the study of human genetic disease. From genome-wide microarray and next-generation sequencing scans, we now know that each human genome contains over 3 million single nucleotide variants when compared with the ~3 billion base pairs in the human reference genome, along with roughly an order of magnitude more DNA—approximately 30 megabase pairs (Mb)—being 'structurally variable', mostly in the form of indels and copy number changes. Additional large-scale variations include balanced inversions (average of 18 Mb) and complex, difficult-to-resolve alterations. Collectively, ~1% of an individual's genome will differ from the human reference sequence. When comparing across a generation, fewer than 100 new genetic variants are typically detected in the euchromatic portion of a child's genome. Driven by increasingly higher-resolution and higher-throughput sequencing technologies, newer and more accurate databases of genetic variation (for instance, more comprehensive structural variation data and phasing of combinations of variants along chromosomes) of worldwide populations will emerge to underpin the next era of discovery in human molecular genetics.

Introduction

Perhaps the greatest paradigm shift for genetics research in recent years has been the move from analyzing just one gene at a time to being able to interrogate the entire genome at once—every gene, be it coding or non-coding, along with all the DNA in between (1–3). Driven by extraordinary innovations in laboratory technology and information sciences, this advance has led to the (re)-birth of the field of genomics (4), particularly as it impacts health care (5). We consider it a re-birth because, from the earliest studies of chromosomes 60–70 years ago, the first direct vantage point of genetics was the morphological anatomy of the genome, not the gene (6–8). As summarized in Figure 1, the classes of genetic variation being described at that time (e.g. aneuploidies; large translocations and deletions) were those that could be seen from cytogenetically stained chromosomes. Although higher-resolution banding eventually enabled the detection of subtler changes, all of these experiments were inextricably linked to the limits of microscopic observation (9).

†Brett Trost, http://orcid.org/0000-0003-4863-7273
‡Livia O. Loureiro, http://orcid.org/0000-0003-0098-7901
¶Stephen W. Scherer, http://orcid.org/0000-0002-8326-1999
Received: April 28, 2021. **Revised:** July 9, 2021. **Accepted:** July 19, 2021

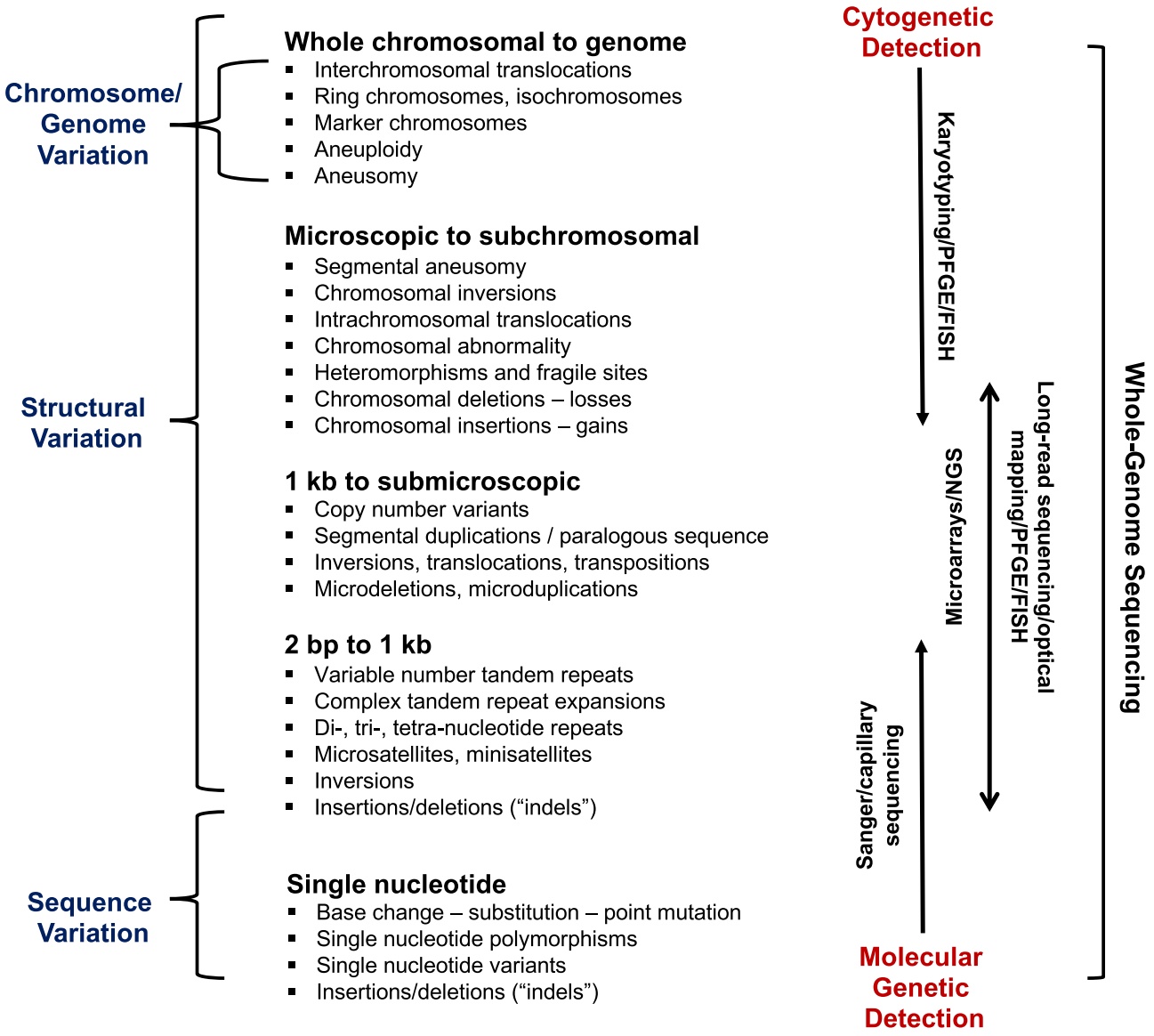

Figure 1. Types of variation found in the human genome and the primary technologies used to detect them (43). The types of variation, and various (sometimes synonymous) terms used to describe them, are grouped as ‘sequence variation’ and ‘structural variation’, the latter encompassing chromosomal/genome variation. The lower end-size of structural variation is typically defined to fall in the 50–1000 nt range, but definitions vary (9,172). FISH, fluorescence *in situ* hybridization (here also encompassing spectral karyotyping); PFGE, pulse field gel electrophoresis; NGS, next-generation sequencing (including both short-read and long-read technologies, the latter being particularly useful for identifying intermediate-size structural variation). There are many other important technologies used to discover and map genetic variation and we include those that have been most impactful for the original discoveries discussed in this review, including those that are still used by clinical diagnostic laboratories. Important references are provided in Tables 1 and 2 and the main text.

Modern genomics arguably began with the elucidation of the structure of DNA in the 1950s (10) and the determination of the genetic code and the modern concept of the gene in the 1960s (11,12). The next three decades saw the development of a plethora of revolutionary DNA sequencing and recombinant DNA cloning technologies that allowed the decoding of individual genes at the nucleotide level, leading to the identification of point mutations and more complex di-, tri- and tetra-nucleotide variants (13–15). Together, the new genomics technologies consolidated genetic (14,16,17) and physical linkage (18,19) strategies and provided the basis for generating the first holistic descriptions of chromosomes and the genome. The decade bridging the year 2000 brought forward chromosomal microarray analysis [CMA; (20–26)], which afforded truly global genotyping capability, including assessment of submicroscopic deletions and duplications in disease samples, as well as the discovery of a previously unrealized amount of DNA copy number variation (CNV) in all individuals (27–29). Moreover, the implementation of automated fluorescence-based DNA sequencing, including clone-end and full-clone ‘shotgun’ sequencing, led to the 2001 release of working draft assemblies of the human genome (1,2), with the first ‘full’ reference sequence, denoted GRCh35, published in 2004 (3). The availability of a high-quality reference assembly provided an entry point for concurrent personal genome sequencing and the generation of integrated maps of genetic variation (30). Recognition of the importance of accurate human genome sequencing at scale led to the (ultimately canceled) $10M ‘Archon Genomics X PRIZE’ to the first group able to

sequence haplotype-resolved genomes satisfying what turned out to be then (and still remain) unreachable criteria for cost and accuracy (31). Perhaps the single most important technology underpinning the current state of genomics is massively parallel DNA sequencing, which was first developed in the late 2000s (32–36). These 'next-generation sequencing' (NGS) technologies can be used to study the human genome at population scale with unprecedented resolution. Augmented by NGS, the latest release of the human reference genome, GRCh38, includes over 97 million more sequenced bases than GRCh35 (3,37–39).

In formulating this review, we aimed to examine two questions fundamental to our understanding of human genetics and its application to medicine—namely, how much variation exists in our diploid genome, and with this baseline, how does its nucleotide composition change from one generation to the next? At the inauguration of the important journal *Human Molecular Genetics* some 30 years ago, having (mostly) accurate answers to these vital questions would have seemed unattainable. *Circa* 2021, however, for the historically well-studied chromosomal- and sequence-level variation, this information is nearing perfection, at least in most euchromatic DNA. In contrast, data for intermediate-sized structural variation (9,40–46), the last broad class of variation to be characterized (Fig. 1), are now catching up as new technologies and algorithms are developed (47,48).

Genetic Variation at the Level of the Individual Human

In 2001, two separate groups, the International Human Genome Sequencing Consortium and Celera Genomics, published initial haploid drafts of the human genome. Both sequences were derived from composites of individuals, and they were generated using highly automated fluorescence-based Sanger DNA sequencing (49) from clone-based and random whole-genome sequencing (WGS), respectively (1,2). In 2007, the 'HuRef' genome—the first genome sequence of an individual human (Craig Venter)—was assembled (50), providing a pivotal starting point to query how much genetic variation exists within a 'diploid' human genome. For this once-in-a-generation project, which built upon Celera Genomics' original efforts (and cost ~$70M), ~1000 bp reads were generated from over 30 million random DNA fragments using Sanger sequencing. These reads were then assembled into 4528 scaffolds, with the assembly strategy enabling alternate alleles in the diploid genome to be defined.

Comparison of this accurate assembly to the reference genome of the time revealed 3 213 401 single nucleotide variants (SNVs) and 851 575 insertions/deletions (indels), which collectively encompassed 12.3 Mb of DNA (Table 1). The observation that non-SNV variants comprised 22% of events in HuRef but 74% of modified base pairs, implying a substantial contribution of larger genetic variants to overall variation, set the standard for how future personal genomes might be characterized, irrespective of the technology used. Further analysis of the HuRef assembly, combined with CMA (22,23), identified 12 178 structural variants (SVs); combined with the non-SNV alterations identified in the initial study, this yielded an estimated total of 39.5 Mb of non-SNV unbalanced variation, along with 90 inversions encompassing 9.3 Mb (51). Thus, the HuRef genome differs from the reference by only ~0.1% when considering SNVs alone, but by a far larger amount (~1.3%) when considering all forms of unbalanced variation. A compelling lesson from this and other early studies of the human genome was that no single sequencing (or other) technology could accurately reveal all of the classes of genetic variation shown in Figure 1 (52).

Additional early studies used direct (clones not required) massively parallel sequencing technologies to generate personal genome sequences for two other pioneers of genome research—James Watson and James Lupski, both of European ancestry (53,54). These million-dollar projects utilized 454 pyrosequencing (32,33) and massively parallel sequencing by ligation (35), yielding 3 322 093 and 3 420 306 SNVs, respectively, with only a few SVs being reported. Concurrently, using what would become a mainstay technology in genomics (Solexa, eventually becoming Illumina sequencing), Bentley *et al.* (34) analyzed the genome of a male Yoruban individual using massively parallel sequencing-by-synthesis. Their data revealed nearly 1 million more SNVs compared with the previously-mentioned genomes of individuals of European ancestry (50,53,54), as well as >400 000 indels and 5000 SVs, many of which were previously unknown. A separate analysis of African hunter-gatherers, the oldest lineages of modern humans, revealed a similar number of SNVs (~4 million) as reported by Bentley, with the trend being that more genetic variation tends to be found in 'older' populations [(55); Table 1].

Published in 2009, the sequencing of the first Korean genome (AK1) used an integrated approach with Illumina shotgun sequencing, bacterial artificial chromosome sequencing and CMA, reporting 3 453 653 SNVs, 170 202 indels and 1237 SVs (56). Interestingly, only 37% of the non-synonymous SNVs in AK1 were also found in both the previously-sequenced African (34) and Chinese (57) genomes. A *de novo* assembly of the AK1 genome with haplotype phasing was subsequently generated (58) using Pacific Biosciences (PacBio) long-read sequencing (59), Illumina short reads (34) and 10x Genomics linked-read technology (60–62). A similar number of SNVs were detected (3 472 576 versus 3 453 653) along with more refined SV data afforded by the long-read technology, including many sequences not found in the human reference genome. Other notable projects sequencing the genomes of individuals of Asian descent include a high-coverage phased Chinese genome [HX1; (63)] and a haploid Japanese genome reference assembled through the consensus among three donors (64) using high-coverage PacBio long reads (59) and Bionano Genomics optical mapping (65). Approximately 2.5 million new SNVs and over 14 000 SVs were reported in the composite Japanese genome, many of which were found to be common in the Japanese population (Table 1). The Japanese study also demonstrated that population-specific reference genomes may facilitate the identification of disease-associated variants compared with using the standard reference. Given that analysis pipelines often ignore sequence reads that do not map to the GRCh38 reference sequence, the construction of this and other population-specific reference genomes (66–69) will surely prove to be important in accurately capturing the full spectrum of DNA sequence (including complex and repetitive elements), as well as genetic variation, in diverse human populations. Additional strategies for improving the reference genome include adjusting all alleles to the major allele form (70).

De novo Mutation Across a Generation

Cataloguing the nature and extent of inherited genetic variation in human populations is important from an evolutionary perspective (71–74), and determining the presence of new variants (*de novo* mutations or DNMs) is critical in medical genomics

Table 1. Important WGS studies examining the extent of variation in a genome[a]

Study	Genome	Ancestry[c]	Sex	DNA	SNVs	Indels		CNVs/SVs			Non-SNV variation (Mb)[b]		Technologies[e]
						Ins	Del	Ins/Dup[d]	Del	Inv	Unbalanced	Balanced	
Levy *et al.* 2007 *PloS Biol.* (50)	HuRef (Venter)	EUR	M	Blood	3 213 401	275 512[f]	283 961[f]	30	32	90	–	–	CMA, SS
Pang *et al.* 2010 *Genome Biol.* (51)					–	412 304[g]	383 775[g]	9915[g]	13 867[g]	167	39.5	9.3	
Wheeler *et al.* 2008 *Nature* (53)	Watson	EUR	M	Blood	3 322 093	65 677	157 041	9	14	–	–	–	454, CMA
Lupski *et al.* 2010 *N. Eng. J. Med.* (54)	III-4 (Lupski)	EUR	M	Blood	3 420 306	–	–	123	111	–	–	–	CMA, SOLiD
Ebert *et al.* 2021 *Science* (124)	NA12878	EUR	F	LCL	3 643 864	367 945	372 590	13 954	8931	108	15.3	21.7	PB, S-seq
Bentley *et al.* 2008 *Nature* (34)	NA18507	AFR	M	Blood	4 139 196	176 221	228 195	2345	5704	–	–	–	IL
Schuster *et al.* 2010 *Nature* (55)	KB1	AFR	M	Blood	4 053 781	–	–	–	–	–	–	–	454, IL
Ebert *et al.* 2021 *Science* (124)	HG03125	AFR	F	LCL	4 470 531	438 853	449 021	16 355	10 775	120	17.5	22.2	PB, S-seq
Chaisson *et al.* 2019 *Nat. Commun.* (141)	NA19240	AFR	F	LCL	–	419 842[h]	370 245[h]	17 026	12 421	129[i]	39.8	19.6[i]	10X, BN, Hi-C, IL, ON, PB, S-seq
Kim *et al.* 2009 *Nature* (56)	AK1	EAS	M	Blood	3 453 653	75 141	95 061	581	656	–	–	–	CMA, IL
Seo *et al.* 2016 *Nature* (58)				LCL	3 472 576	169 314 total		10 077	7358	71	13.6[j]	13.5	10x, BN, IL, PB
Shi *et al.* 2016 *Nat. Commun.* (63)	HX1	EAS	M	Blood	3 518 309	16 625 690 total		10 284	9891	–	11.0[j]	–	BN, IL, PB
Ebert *et al.* 2021 *Science* (124)	HG00512	EAS	M	LCL	3 620 202	367 796	370 030	14 055	8937	122	15.5	21.0	PB, S-seq
Chaisson *et al.* 2019 *Nat. Commun.* (141)	HG00514	EAS	F	LCL	–	335 762[h]	297 565[h]	15 566	10 291	121[i]	39.3	14.1[i]	10X, BN, Hi-C, IL, ON, PB, S-seq
Takayama *et al.* 2021 *Nat. Commun.* (64)	JG1[k]	EAS	M	Blood	2 501 575	–	–	8697	6190	–	–	–	BN, IL, PB
Ebert *et al.* 2021 *Science* (124)	HG02492	SAS	M	LCL	3 565 097	372 637	347 792	13 993	8994	108	16.3	20.8	PB, S-seq
Chaisson *et al.* 2019 *Nat. Commun.* (141)	HG00733	AMR	F	LCL	–	343 950[h]	304 170[h]	16 566	10 607	128[i]	31.6	17.9[i]	10X, BN, Hi-C, IL, ON, PB, S-seq
Ebert *et al.* 2021 *Science* (124)	HG00731	AMR	M	LCL	3 693 860	379 989	379 972	14 009	8867	107	15.6	20.0	PB, S-seq

[a]We selected studies spanning the start of personal genome sequencing in 2007 until 2021, including those from diverse populations analyzed using different technologies. The size definitions used to categorize indels (insertions and deletions) and CNVs (insertions/duplications and deletions) varied between studies, leading to significant differences in numbers presented. The Levy *et al.* study (HuRef/Venter genome) provides a composite analysis, demonstrating that relative to the reference genome, ~1.3% of nucleotides were affected by indels and CNVs compared with 0.1% by SNVs. More recent studies further support the idea that non-SNV variation affects several times more nucleotides than SNVs (58,124,141). Where reported, balanced SVs (inversions in most studies) encompass between 9.3 and 22.2 Mb (average 18 Mb). Data from these studies are typically also accessible in public repositories (159,173–175).

[b]The total number of base pairs affected by non-SNV sequence changes. Unbalanced changes include insertions and deletions of all sizes, whereas balanced changes include inversions.

[c]Abbreviations: AFR, African; AMR, Admixed American; EAS, East Asian; EUR, European; SAS, South Asian.

[d]Additions of genetic material are typically described as insertions when detected by comparisons between assembled genomes and as duplications when detected using chromosomal microarray analysis.

[e]The technologies used for sequencing, assembly and variant detection. Abbreviations: 10x, 10x Genomics linked reads (60–62); 454, 454 Life Sciences pyrosequencing (32,33); BN, Bionano Genomics optical mapping (65); CMA, chromosomal microarray analysis (20–26); IL, Illumina (Solexa) sequencing (34); ON, Oxford Nanopore Technologies sequencing (137–140); PB, Pacific Biosciences sequencing (59); SOLiD, sequencing by oligonucleotide ligation and detection (35); SS, Sanger sequencing (49); S-seq, strand-seq (135,136).

[f]Values represent homozygous indels; 292 102 heterozygous indels (not stratified by insertions and deletions in the paper) were also detected.

[g]Insertions and deletions detected using assembly comparison are listed under indels, whereas those detected using other methods are listed under CNVs/SVs.

[h]Detected by Illumina sequencing.

[i]Reflects simple inversions as tabulated in Supplementary Table 9 of Chaisson *et al.* (141)

[j]Excludes indels.

[k]Composite of three different Japanese males.

LCL, lymphoblastoid cell line.

(75–77). Early estimates of mutation rates were made using cross-species comparisons (78), small numbers of human genetic loci (79) or—in a seminal paper in *Human Molecular Genetics*—specific tandem repeat loci (14). However, the direct measurement of genome-wide mutation rates requires WGS of biological parent–child trios, which has only become feasible at scale, with increasing completeness and accuracy, over the last 10 years. Therefore, the first such studies included small numbers of trios (80,81), with more recent studies involving orders of magnitude more families, often as part of disease studies (Table 2). For reasons of cost (a 30x coverage genome today at ~$1000) and accuracy (at least for SNVs), the sequencing method of choice has been Illumina short-read technology, so accordingly, most of the DNM data presented are limited to SNVs. As discussed in Table 1, comprehensive and accurate detection of larger variants is challenging with short-read data alone, so until recently, much of the information for *de novo* CNVs has come from CMA (Table 2).

Considering SNVs alone, studies have revealed 35–82 DNMs per generation within the mappable genome [(80–94); Table 2]. Although reasonably consistent, these estimates are not perfectly comparable across studies due to differences in the proportion of the genome assessed. After adjustment, studies consistently report a mutation rate of $\sim 1.2 \times 10^{-8}$ per nucleotide per generation (83,84,88,92–94). Interestingly, mutation rate estimates from trios are highly concordant with earlier estimates (78,79). By comparing DNMs in monozygotic twins, it has been estimated that ~97% are germline in origin, whereas 3% are somatic (87). Although some studies in Table 2 include individuals ascertained for specific diseases, little difference has been observed in the total number of constitutional *de novo* SNVs compared with healthy individuals (95).

Many DNM studies have examined the parental age effect—the number of additional DNMs per year of parental age. This effect is greater in fathers, with estimates ranging from 0.64 to 2.0 additional DNMs per additional year of age versus 0.24–0.42 for mothers (Table 2). As a result, fathers contribute more DNMs per generation than mothers; paternal/maternal ratios of 3–5 have been reported (83,84,88,92), an observation increasingly made in studies of autism (90,91,96,97). Although DNMs in general are more likely to be of paternal origin, some genomic regions exhibit a significant bias toward maternally-derived DNMs (89).

Although most DNM studies have examined homogeneous population groups [e.g. Dutch, Icelandic or Danish citizens; (87,92,93)] or have not investigated the effect of ancestry, one study found that mutation rates were generally consistent across populations, but were ~7% lower in Amish individuals (94). The same study found that the contribution of additive genetic effects to mutation rate is non-existent (94); thus, variation in mutation rate not explained by parental age is likely due to some combination of non-additive genetic effects and environmental factors. In the case of the Amish, it seems plausible that the observed difference could be partially accounted for by some combination of consanguinity and lifestyle factors, such as reduced exposure to mutagens.

Interestingly, WGS studies have revealed no clear impact of extreme environmental exposure on DNM rates, including in children of parents exposed to dioxin (98) or to radiation from the atomic bombings of Hiroshima and Nagasaki (99) or the Chernobyl nuclear accident (100).

DNMs do not occur with equal probability throughout the genome; rather, their frequency is influenced by sequence context. Trio studies have shown that ~2/3 of DNMs are transitions and that these events occur 20x more frequently at CpG sites (83). DNMs from younger fathers are more likely to occur in late-replicating genomic regions, whereas no such effect has been observed in mothers or older fathers (87). Because early-replicating regions are more gene-rich (101), this bias may further increase the probability of a deleterious DNM originating from an older father. Representing ~2% of all DNMs, DNM clusters have been observed, typically within 20 kb windows, and appear to have distinct mutational signatures compared with non-clustered DNMs (87,89). The number of DNM clusters increases with parental age at an approximately equal rate for mothers and fathers; this suggests that they arise from a different mutational mechanism (compared with non-clustered DNMs) that is common between mothers and fathers (89), although some differences in paternally- versus maternally-derived clusters have been observed (92). Studies of autism have also observed clustered DNMs (82,90), which are mainly maternally-derived and are often found adjacent to *de novo* CNVs (90). A comprehensive review of mutational patterns, as well as the disease implications of *de novo* variants, is published (102).

Recent studies have estimated that 4–13 *de novo* indels occur per generation (90–93,95). Deletions were found to be more common than insertions, and even-sized indels were more common than odd-sized indels (93). Specialized algorithms for identifying *de novo* indels within tandem repeat loci have detected ~55 events per genome in healthy individuals (103), along with a paternal origin bias and age effect. The corresponding tandem repeat *de novo* rate, estimated at 5.6×10^{-5} per generation per locus, is far lower than much earlier estimates for tandem repeats based on a few loci and PCR-based tests (14), reflecting changes in accuracy afforded by better technology and genome-wide genotyping ability. However, that so many *de novo* indels were detected in tandem repeat regions over and above those detected in non-repetitive regions suggests that the total degree of *de novo* variation has been underestimated—not only for indels, but also for other classes of variation shown in Figure 1. As new technologies and algorithms improve our ability to interrogate repetitive and difficult-to-map regions of the genome, measured *de novo* rates for all types of variation will rise.

Compared with SNVs and indels, *de novo* rates for CNVs and SVs have been less well-characterized. CMA has revealed that CNV mutation rates differ depending on CNV size and that large *de novo* CNVs are substantially more frequent in individuals with autism compared with unaffected individuals (104–107), some of which are recurrent and clinically relevant (108). Another autism study estimated the rate of *de novo* CNVs > 10 kb at 0.05 per generation (90). Recently, Collins *et al.* (109) used WGS to estimate mutation rates for SVs > 50 bp, with each generation averaging 0.15 *de novo* deletions, 0.1 insertions, 0.04 duplications and 0.001 inversions. Yet another recent study found ~0.16 *de novo* SVs per healthy individual, along with a significantly higher rate (0.21) in individuals with autism (110). Interestingly, the latter study found that most *de novo* SVs originated from the father but did not find statistical evidence for a parental age effect on *de novo* SV rate, which is in contrast to the well-established parental age effect for *de novo* SNVs (82,83,87–89,92,94).

Redefining Genomic Variation Using Short- and Long-Read WGS

As affordable WGS has become commonplace, the ability to comprehensively detect the many classes of genetic variation in large, diverse sets of individuals (111–117) has improved

Table 2. Important genome-wide studies examining *de novo* variation across a generation[a]

Study	Families	Phenotype[b]	Technology[c]	DNM rate (events/generation)[d]	Paternal age effect[e]	Maternal age effect[e]
Sebat *et al.* 2007 *Science* (105)	264	ASD	CMA	0.01 CNVs[f]	–	–
Itsara *et al.* 2010 *Genome Res.* (107)	2197	ASD	CMA	Varies by size[g]	–	–
Roach *et al.* 2010 *Science* (81)	1	See note[h]	WGS	70 SNVs	–	–
Conrad *et al.* 2011 *Nat. Genet.* (80)	2	NA	WGS	42 SNVs	–	–
Michaelson *et al.* 2012 *Cell* (82)	10	ASD	WGS	58 SNVs	1.0 SNVs	–
Kong *et al.* 2012 *Nature* (83)	78	ASD, SCZ	WGS	63 SNVs	2.0 SNVs	–
Campbell *et al.* 2012 *Nat. Genet.* (84)	5	NA	WGS	35 SNVs[i]	–	–
Gilissen *et al.* 2014 *Nature* (85)	50	ID	WGS	82 SNVs, 0.16 CNVs	–	–
Francioli *et al.* 2014, 2015 *Nat. Genet.* (86,87)	250	NA	WGS	43 SNVs	1.1 SNVs	–
Wong *et al.* 2016 *Nat. Commun.* (88)	693	PTB	WGS	39 SNVs	0.64 SNVs	0.35 SNVs
Goldmann *et al.* 2016 *Nat. Genet.* (89)	816	PTB	WGS	45 SNVs	0.91 SNVs	0.24 SNVs
Yuen *et al.* 2016 *NPJ Genom. Med.* (90)	200	ASD	WGS	51 SNVs, 4 indels, 0.05 CNVs[j]	–	–
Yuen *et al.* 2017 *Nat. Neurosci.* (91)	1239	ASD	WGS	74 SNVs, 13 indels	–	–
Jónsson *et al.* 2017 *Nature* (92)	1548	Various	WGS	65 SNVs, 5 indels	1.51 SNVs+indels	0.37 SNVs+indels
Maretty *et al.* 2017 *Nature* (93)	50	NA	WGS	64 SNVs, 6 indels	–	–
An *et al.* 2018 *Science* (95)	1902	ASD	WGS	62 SNVs, 6 indels	–	–
Kessler *et al.* 2020 *Proc. Natl. Acad. Sci.* (94)	1465	Various	WGS	64 SNVs	1.35 SNVs	0.42 SNVs
Collins *et al.* 2020 *Nature* (109)	970	Various	WGS	0.29 SVs[k]	–	–
Belyeu *et al.* 2021 *Am. J. Hum. Genet.* (110)	2396	ASD	WGS	0.16 SVs[l]	Not significant	Not significant
Mitra *et al.* 2021 *Nature* (103)	1637	ASD	WGS	53 tandem repeat indels[m]	Significant[n]	–

[a]We selected studies that tested for genome-wide *de novo* mutation events from population control or disease datasets. Each study has strengths and weaknesses in design, data capture and experimental validation. Four comprehensive studies (90–93) report an average of 64 SNV, 7 indel and 0.05 CNV events per generation.
[b]The phenotype or disease of participants in the study. 'NA' means that only healthy controls were used or that no disease phenotype was indicated. ASD, autism spectrum disorder; ID, intellectual disability; PTB, preterm birth; SCZ, schizophrenia.
[c]The technology used for variant detection. CMA, chromosomal microarray analysis; WGS, whole-genome sequencing.
[d]DNM rates are reported in terms of events per generation because this measure is generalizable across variant types (i.e. also including indels and SVs). As mentioned in the text, after adjusting for the proportion of the genome assessed, estimates of per-nucleotide mutation rates for *de novo* SNVs are consistently reported as $\sim 1.2 \times 10^{-8}$ per generation.
[e]The estimated number of additional *de novo* variants per year of parental age.
[f]CNVs > 99 kb in unaffected individuals only.
[g]CNVs > 30 kb: 0.012; CNVs > 500 kb: 0.0065.
[h]The two siblings in this study each had two recessive disorders.
[i]This study also estimated mutation rates based on heterozygous positions within autozygous segments, giving a per-nucleotide mutation rate of 1.2×10^{-8} per generation.
[j]CNVs > 10 kb.
[k]Includes 0.15 deletions, 0.1 insertions, 0.04 duplications and 0.001 inversions.
[l]Value is for healthy individuals; DNM rate was significantly higher in ASD-affected individuals (0.21 SVs/generation).
[m]Value is for healthy individuals; DNM rate was slightly but significantly higher in ASD-affected individuals (55 tandem repeat indels/generation).
[n]Paternal age effect was statistically significant, but no slope given.

considerably, aided by the development of variant benchmarking resources (118–120). These studies have, in turn, enabled the study of disease (109,121,122), human migration and adaptation patterns (123) and evolution (124). As genetic variation becomes better defined across different ancestry groups (93,125–128), including in archaic genomes (Denisova, Neanderthal) (129,130), an increasing amount of genetic variation is being found among lineages. Personal genome sequencing of diverse populations with different technologies is also revealing novel DNA sequences (and therefore genetic variation) not currently present in the human reference genome and corresponding databases (55,58,67,131). In perhaps the most astounding example of the power of sequencing technology to map variants across a generation, an 'F1' offspring of a *Homo sapiens neanderthalensis* and *Homo sapiens denisova* was discerned (132). Most of the aforementioned studies concentrate on SNVs, since they are the easiest to discover from the current industry-standard short-read sequencing technology.

Recently, papers describing 'end-to-end' chromosome assemblies have been published, focusing on using long-read sequencing technologies to enable SV discovery and mapping [Table 1; (133,134)]. In a *tour de force* effort, PacBio long-read (59) and strand-specific (135,136) sequencing technologies were used to generate haplotype-resolved *de novo* assemblies of 32 diverse individuals at an estimated cost per genome of ~$20 000 (124). With this approach, 107 590 SVs were found, representing an average of 16 Mb of structural variation per individual, of which 68% were not discovered using standard short-read sequencing. In a parallel effort using a multi-platform approach [PacBio (59) and Oxford Nanopore (137–140) long-read sequencing, Illumina short-read sequencing (34), 10x Genomics linked reads (60–62) and Bionano Genomics optical mapping (65)], three trios of Han Chinese, Puerto Rican and Yoruban ancestry were sequenced, yielding SV sets 3–7x larger than most other standards (141). As shown in Table 1, the unbalanced SVs impacted 31.6, 39.3 and 39.8 Mb in admixed American, East Asian and African ancestries, respectively, all closer to what was found using the integrated approach in the HuRef/Venter project (50,51). The impact of balanced inversions is also shown in Table 1. Although giving near chromosome-level resolution, these long-read sequencing studies emphasize limitations in assembly and discrimination, particularly at gene-rich regions harboring complex structural variation. Given the current error rate of these technologies, accurately detecting SNVs still requires 'filling in' using short-read sequence data, highlighted by the fact that some trio studies do not overtly report DNM rates or SNV quality (124,141). In studies using cell line-derived DNA, the transforming viral integration process and culturing can cause modest but detectable changes in the genome (142,143), which may also be a confounder.

Many studies, including one describing the use of Oxford Nanopore long-read technology to study the Icelandic population (117,144–148), reaffirm the need to consider large-scale copy number and structural variation in disease study design. In our own recent research, developing novel computational and statistical methods to analyze existing short-read sequence data for expanded tandem repeats led to the discovery of specific loci associated with autism (149), an intriguing finding given that most known disorders associated with tandem repeat expansions are monogenic (150). The same study also discovered extensive polymorphism in repeat motif size and sequence, often correlated with cytogenetic 'fragile site' variation along chromosomes (149). Moreover, 158 991 ultra-rare SVs were recently found through the study of 17 795 population controls, with 2% of individuals carrying megabase-scale SVs (117). The same study found reciprocal translocations at a rate of 1 in 1000 individuals, a number similar to that found using classical cytogenetics (151,152).

There are two fundamental steps to identifying associations between genotypes and health: variant detection and variant interpretation. With the combination of long-read technology and other sequencing methods now enabling the 'complete' sequencing of chromosomes (133,134,153), making further improvements for variant detection essentially represents an engineering problem. Although significant challenges remain, including cost reductions in long-read sequencing, accurate phasing of diploid genomes and scaling the end-to-end assembly process to entire populations, it seems plausible that variant detection will eventually become a *fait accompli*. To the contrary, variant interpretation is still in its early days, perhaps even reminiscent of examining chromosome banding in the 1960s (154–158). Although our ability to interpret the impact of copy number changes and loss-of-function sequence-level variants is somewhat mature, understanding the effects of most other alterations, such as missense variants and variants impacting regulatory elements, remains largely unresolved. The rapidly increasing pace by which sequencing data are now generated, along with the move to examining populations at scale and the use of multi-omics technologies, ultimately promise to reduce the time from data generation to data interpretation (159–167).

Conclusions

The current assembly of the human genome (GRCh38) comprises 3 099 706 404 bp. Comparing any other genome to it yields ~3–4 million SNVs and (with comprehensive multi-technology testing) ~10 times as many nucleotides impacted by unbalanced structural variations, most notably indels and CNVs (Table 1). Notwithstanding the many complexities in whole-genome analysis, it can be conservatively stated that ~1% variation exists between each of our DNA when compared with the reference, with those genomes arising from African and other ancestral populations exhibiting more genetic variation than those arising more recently in human history. A consistent message from the literature is that no single technology or method can detect all genetic variation, and knowledge of how the data (and databases housing it) were derived is essential to correctly interpreting it. The number of DNMs found in the mappable euchromatic DNA in a single individual is modest (fewer than 100), but this value may increase as more complex sequences are considered in tallies of genetic variation—noting, however, that nomenclature and reporting of SVs, in particular in repetitive regions, is challenging (159,168–170). Newer WGS technologies (e.g. long-read sequencing) that facilitate the discovery and genotyping of complex variants will have a growing impact in disease studies and population sequencing as their costs begin to compete with the more prevalent short-read technologies. When analyzing larger sample sizes for their genomic architecture, cost considerations mean that short-read sequencing studies will prevail, likely for a while, even when considering structural variation. Drawing from the fundamental genomic data presented in Tables 1 and 2, we calculate that from 4 billion births (171) and ~71 *de novo* SNVs/indels/CNVs per individual, >284 billion DNMs have arisen over the past 30 years of human history. Such a wellspring of genetic variation, once characterized, will power the next generation of studies in human molecular genetics.

Acknowledgements

We thank Drs. Si Lok, Richard Wintle and Ryan Yuen for editorial comments and Dr. Charles Lee and his team for providing information assisting in the creation of Table 1. Estimated project costs, which are based on many sources including media reports and the memory of S.W.S., are given to help inform the reader how they may have impacted study design.

Conflict of Interest statement. None declared.

Funding

B.T.'s postdoctoral fellowship has been supported by the Canadian Institutes of Health Research (CIHR) Banting Postdoctoral Fellowship and the Canadian Open Neuroscience Platform Research Scholar Award. L.O.L. holds the Lap-Chee Tsui Fellowship for Research Excellence. S.W.S. holds the Northbridge Chair in Paediatric Research at The Hospital for Sick Children and University of Toronto.

References

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
3. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
4. Green, E.D., Watson, J.D. and Collins, F.S. (2015) Human Genome Project: twenty-five years of big biology. *Nature*, **526**, 29–31.
5. Turro, E., Astle, W.J., Megy, K., Gräf, S., Greene, D., Shamardina, O., Allen, H.L., Sanchis-Juan, A., Frontini, M., Thys, C. *et al.* (2020) Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*, **583**, 96–102.
6. Jacobs, P.A., Baikie, A.G., Court Brown, W.M. and Strong, J.A. (1959) The somatic chromosomes in mongolism. *Lancet*, **273**, 710.
7. Edwards, J.H., Harnden, D.G., Cameron, A.H., Crosse, V.M. and Wolff, O.H. (1960) A new trisomic syndrome. *Lancet*, **275**, 787–790.
8. Patau, K., Smith, D.W., Therman, E., Inhorn, S.L. and Wagner, H.P. (1960) Multiple congenital anomaly caused by an extra autosome. *Lancet*, **275**, 790–793.
9. Feuk, L., Carson, A.R. and Scherer, S.W. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.
10. Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–738.
11. Crick, F.H., Barnett, L., Brenner, S. and Watts-Tobin, R.J. (1961) General nature of the genetic code for proteins. *Nature*, **192**, 1227–1232.
12. Portin, P. and Wilkins, A. (2017) The evolving definition of the term "gene". *Genetics*, **205**, 1353–1364.
13. Weber, J.L. and May, P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.*, **44**, 388–396.
14. Weber, J.L. and Wong, C. (1993) Mutation of human short tandem repeats. *Hum. Mol. Genet.*, **2**, 1123–1128.
15. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W. (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.
16. Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G. and Lathrop, M. (1992) A second-generation linkage map of the human genome. *Nature*, **359**, 794–801.
17. Murray, J.C., Buetow, K.H., Weber, J.L., Ludwigsen, S., Scherpbier-Heddema, T., Manion, F., Quillen, J., Sheffield, V.C., Sunden, S. and Duyk, G.M. (1994) A comprehensive human linkage map with centimorgan density. *Science*, **265**, 2049–2054.
18. Cohen, D., Chumakov, I. and Weissenbach, J. (1993) A first-generation physical map of the human genome. *Nature*, **366**, 698–701.
19. Hudson, T.J., Stein, L.D., Gerety, S.S., Ma, J., Castle, A.B., Silva, J., Slonim, D.K., Baptista, R., Kruglyak, L., Xu, S.H. *et al.* (1995) An STS-based map of the human genome. *Science*, **270**, 1945–1954.
20. Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P. and Fodor, S.P. (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 5022–5026.
21. Lipshutz, R.J., Morris, D., Chee, M., Hubbell, E., Kozal, M.J., Shah, N., Shen, N., Yang, R. and Fodor, S.P. (1995) Using oligonucleotide probe arrays to access genetic diversity. *BioTechniques*, **19**, 442–447.
22. Brown, P.O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genet.*, **21**, 33–37.
23. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. and Lockhart, D.J. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.*, **21**, 20–24.
24. Fan, J.B., Oliphant, A., Shen, R., Kermani, B.G., Garcia, F., Gunderson, K.L., Hansen, M., Steemers, F., Butler, S.L., Deloukas, P. *et al.* (2003) Highly parallel SNP genotyping. *Cold Spring Harb. Symp. Quant. Biol.*, **68**, 69–78.
25. Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H. *et al.* (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*, **1**, 109–111.
26. Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A.C., Thiruvahindrapuram, B., MacDonald, J.R., Mills, R. *et al.* (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.*, **29**, 512–520.
27. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
28. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Manér, S., Massa, H., Walker, M., Chi, M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
29. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
30. Snyder, M., Du, J. and Gerstein, M. (2010) Personal genome sequencing: current approaches and challenges. *Genes Dev.*, **24**, 423–431.
31. Kedes, L. and Liu, E.T. (2010) The Archon Genomics X PRIZE for whole human genome sequencing. *Nat. Genet.*, **42**, 917–918.

32. Leamon, J.H., Lee, W.L., Tartaro, K.R., Lanza, J.R., Sarkis, G.J., deWinter, A.D., Berka, J., Weiner, M., Rothberg, J.M. and Lohman, K.L. (2003) A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis*, **24**, 3769–3777.
33. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
34. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
35. McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, **19**, 1527–1541.
36. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G. *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78–81.
37. Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D. *et al.* (2017) Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, **27**, 849–864.
38. University of California Santa Cruz Genome Browser (2021) *UCSC Genome Browser: Statistics*. https://genome.ucsc.edu/goldenPath/stats.html (accessed date July 29, 2021).
39. National Center for Biotechnology Information (2021) *Genome Reference Consortium Human Build 38 Patch Release 13*. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39 (accessed date July 29, 2021).
40. Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D. *et al.* (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.
41. Feuk, L., MacDonald, J.R., Tang, T., Carson, A.R., Li, M., Rao, G., Khaja, R. and Scherer, S.W. (2005) Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.*, **1**, e56.
42. Feuk, L., Marshall, C.R., Wintle, R.F. and Scherer, S.W. (2006) Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum. Mol. Genet.*, **15**, R57–R66.
43. Scherer, S.W., Lee, C., Birney, E., Altshuler, D.M., Eichler, E.E., Carter, N.P., Hurles, M.E. and Feuk, L. (2007) Challenges and standards in integrating surveys of structural variation. *Nat. Genet.*, **39**, S7–S15.
44. Khaja, R., Zhang, J., MacDonald, J.R., He, Y., Joseph-George, A.M., Wei, J., Rafiq, M.A., Qian, C., Shago, M., Pantano, L. *et al.* (2006) Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet.*, **38**, 1413–1418.
45. Sharp, A.J., Cheng, Z. and Eichler, E.E. (2006) Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 407–442.
46. Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J.O., Baker, C., Malig, M., Mutlu, O. *et al.* (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.*, **41**, 1061–1067.
47. Schadt, E.E., Turner, S. and Kasarskis, A. (2010) A window into third-generation sequencing. *Hum. Mol. Genet.*, **19**, R227–R240.
48. Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A. and Waterston, R.H. (2017) DNA sequencing at 40: past, present and future. *Nature*, **550**, 345–353.
49. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, **74**, 5463–5467.
50. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
51. Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M.A., Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C. *et al.* (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.*, **11**, R52.
52. Pang, A.W.C., Macdonald, J.R., Yuen, R.K.C., Hayes, V.M. and Scherer, S.W. (2014) Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. *G3*, **4**, 63–65.
53. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
54. Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C.Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A. *et al.* (2010) Whole-genome sequencing in a patient with Charcot–Marie–tooth neuropathy. *N. Engl. J. Med.*, **362**, 1181–1191.
55. Schuster, S.C., Miller, W., Ratan, A., Tomsho, L.P., Giardine, B., Kasson, L.R., Harris, R.S., Petersen, D.C., Zhao, F., Qi, J. *et al.* (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature*, **463**, 943–947.
56. Kim, J.I., Ju, Y.S., Park, H., Kim, S., Lee, S., Yi, J.H., Mudge, J., Miller, N.A., Hong, D., Bell, C.J. *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature*, **460**, 1011–1015.
57. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
58. Seo, J.-S., Rhie, A., Kim, J., Lee, S., Sohn, M.-H., Kim, C.-U., Hastie, A., Cao, H., Yun, J.-Y., Kim, J. *et al.* (2016) *De novo* assembly and phasing of a Korean human genome. *Nature*, **538**, 243–247.
59. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.
60. Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M. and Jaffe, D.B. (2017) Direct determination of diploid genome sequences. *Genome Res.*, **27**, 757–767.
61. Elyanow, R., Wu, H.-T. and Raphael, B.J. (2018) Identifying structural variants using linked-read sequencing data. *Bioinformatics*, **34**, 353–360.
62. Marks, P., Garcia, S., Barrio, A.M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A. *et al.* (2019) Resolving the full spectrum of human genome variation using linked-reads. *Genome Res.*, **29**, 635–645.

63. Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S. *et al.* (2016) Long-read sequencing and *de novo* assembly of a Chinese genome. *Nat. Commun.*, **7**, 12065.
64. Takayama, J., Tadaka, S., Yano, K., Katsuoka, F., Gocho, C., Funayama, T., Makino, S., Okamura, Y., Kikuchi, A., Sugimoto, S. *et al.* (2021) Construction and integration of three *de novo* Japanese human genome assemblies toward a population-specific reference. *Nat. Commun.*, **12**, 226.
65. Pendleton, M., Sebra, R., Pang, A.W.C., Ummat, A., Franzen, O., Rausch, T., Stütz, A.M., Stedman, W., Anantharaman, T., Hastie, A. *et al.* (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods*, **12**, 780–786.
66. Fakhro, K.A., Staudt, M.R., Ramstetter, M.D., Robay, A., Malek, J.A., Badii, R., Al-Marri, A.A.-N., Abi Khalil, C., Al-Shakaki, A., Chidiac, O. *et al.* (2016) The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Hum. Genome Var.*, **3**, 16016.
67. Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M.P., Chavan, S., Vergara, C., Ortega, V.E. *et al.* (2019) Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.*, **51**, 30–35.
68. Shumate, A., Zimin, A.V., Sherman, R.M., Puiu, D., Wagner, J.M., Olson, N.D., Pertea, M., Salit, M.L., Zook, J.M. and Salzberg, S.L. (2020) Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol.*, **21**, 129.
69. Daw Elbait, G., Henschel, A., Tay, G.K. and Al Safar, H.S. (2021) A population-specific major allele reference genome from the United Arab Emirates population. *Front. Genet.*, **12**, 660428.
70. Ballouz, S., Dobin, A. and Gillis, J.A. (2019) Is it time to change the reference genome? *Genome Biol.*, **20**, 159.
71. Pritchard, J.K. and Przeworski, M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, **69**, 1–14.
72. Sunyaev, S., Ramensky, V., Koch, I., Lathe, W. 3rd, Kondrashov, A.S. and Bork, P. (2001) Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–597.
73. Pritchard, J.K. and Cox, N.J. (2002) The allelic architecture of human disease genes: common disease-common variant…or not? *Hum. Mol. Genet.*, **11**, 2417–2423.
74. Buchanan, J.A. and Scherer, S.W. (2008) Contemplating effects of genomic structural variation. *Genet. Med.*, **10**, 639–647.
75. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
76. Veltman, J.A. and Brunner, H.G. (2012) *De novo* mutations in human genetic disease. *Nat. Rev. Genet.*, **13**, 565–575.
77. Ronemus, M., Iossifov, I., Levy, D. and Wigler, M. (2014) The role of *de novo* mutations in the genetics of autism spectrum disorders. *Nat. Rev. Genet.*, **15**, 133–141.
78. Nachman, M.W. and Crowell, S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.
79. Kondrashov, A.S. (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.*, **21**, 12–27.
80. Conrad, D.F., Keebler, J.E.M., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V. *et al.* (2011) Variation in genome-wide mutation rates within and between human families. *Nat. Genet.*, **43**, 712–714.
81. Roach, J.C., Glusman, G., Smit, A.F.A., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M. *et al.* (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, **328**, 636–639.
82. Michaelson, J.J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., Jian, M., Liu, G., Greer, D., Bhandari, A. *et al.* (2012) Whole-genome sequencing in autism identifies hot spots for *de novo* germline mutation. *Cell*, **151**, 1431–1442.
83. Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A. *et al.* (2012) Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature*, **488**, 471–475.
84. Campbell, C.D., Chong, J.X., Malig, M., Ko, A., Dumont, B.L., Han, L., Vives, L., O'Roak, B.J., Sudmant, P.H., Shendure, J. *et al.* (2012) Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.*, **44**, 1277–1281.
85. Gilissen, C., Hehir-Kwa, J.Y., Thung, D.T., van de Vorst, M., van Bon, B.W.M., Willemsen, M.H., Kwint, M., Janssen, I.M., Hoischen, A., Schenck, A. *et al.* (2014) Genome sequencing identifies major causes of severe intellectual disability. *Nature*, **511**, 344–347.
86. Francioli, L.C., Menelaou, A., Pulit, S.L., van Dijk, F., Palamara, P.F., Elbers, C.C., Neerincx, P.B.T., Ye, K., Guryev, V., Kloosterman, W.P. *et al.* (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.*, **46**, 818–825.
87. Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., van Duijn, C.M., Swertz, M., Wijmenga, C., van Ommen, G. *et al.* (2015) Genome-wide patterns and properties of *de novo* mutations in humans. *Nat. Genet.*, **47**, 822–826.
88. Wong, W.S.W., Solomon, B.D., Bodian, D.L., Kothiyal, P., Eley, G., Huddleston, K.C., Baker, R., Thach, D.C., Iyer, R.K., Vockley, J.G. *et al.* (2016) New observations on maternal age effect on germline *de novo* mutations. *Nat. Commun.*, **7**, 10486.
89. Goldmann, J.M., Wong, W.S.W., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A.B., Glusman, G., Vissers, L.E.L.M., Hoischen, A., Roach, J.C. *et al.* (2016) Parent-of-origin-specific signatures of *de novo* mutations. *Nat. Genet.*, **48**, 935–939.
90. Yuen, R.K.C., Merico, D., Cao, H., Pellecchia, G., Alipanahi, B., Thiruvahindrapuram, B., Tong, X., Sun, Y., Cao, D., Zhang, T. *et al.* (2016) Genome-wide characteristics of *de novo* mutations in autism. *NPJ Genom. Med.*, **1**, 16027.
91. Yuen, R.K.C., Merico, D., Bookman, M., L Howe, J., Thiruvahindrapuram, B., Patel, R.V., Whitney, J., Deflaux, N., Bingham, J., Wang, Z. *et al.* (2017) Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neuroscience*, **20**, 602–611.
92. Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M.T., Hjorleifsson, K.E., Eggertsson, H.P., Gudjonsson, S.A. *et al.* (2017) Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland. *Nature*, **549**, 519–522.
93. Maretty, L., Jensen, J.M., Petersen, B., Sibbesen, J.A., Liu, S., Villesen, P., Skov, L., Belling, K., Theil Have, C., Izarzugaza, J.M.G. *et al.* (2017) Sequencing and *de novo* assembly of 150 genomes from Denmark as a population reference. *Nature*, **548**, 87–91.

94. Kessler, M.D., Loesch, D.P., Perry, J.A., Heard-Costa, N.L., Taliun, D., Cade, B.E., Wang, H., Daya, M., Ziniti, J., Datta, S. *et al.* (2020) *De novo* mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proc. Natl. Acad. Sci. U. S. A.*, **117**, 2560–2569.
95. An, J.-Y., Lin, K., Zhu, L., Werling, D.M., Dong, S., Brand, H., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L. *et al.* (2018) Genome-wide *de novo* risk score implicates promoter variation in autism spectrum disorder. *Science*, **362**, eaat6576.
96. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E. *et al.* (2014) The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature*, **515**, 216–221.
97. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L. *et al.* (2020) Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*, **180**, 568–584.e23.
98. Ton, N.D., Nakagawa, H., Ha, N.H., Duong, N.T., Nhung, V.P., Hien, L.T.T., Hue, H.T.T., Hoang, N.H., Wong, J.H., Nakano, K. *et al.* (2018) Whole genome sequencing and mutation rate analysis of trios with paternal dioxin exposure. *Hum. Mutat.*, **39**, 1384–1392.
99. Horai, M., Mishima, H., Hayashida, C., Kinoshita, A., Nakane, Y., Matsuo, T., Tsuruda, K., Yanagihara, K., Sato, S., Imanishi, D. *et al.* (2018) Detection of *de novo* single nucleotide variants in offspring of atomic-bomb survivors close to the hypocenter by whole-genome sequencing. *J. Hum. Genet.*, **63**, 357–363.
100. Yeager, M., Machiela, M.J., Kothiyal, P., Dean, M., Bodelon, C., Suman, S., Wang, M., Mirabello, L., Nelson, C.W., Zhou, W. *et al.* (2021) Lack of transgenerational effects of ionizing radiation exposure from the Chernobyl accident. *Science*, **372**, 725–729.
101. Koren, A. (2014) DNA replication timing: coordinating genome stability with genome regulation on the X chromosome and beyond: prospects & overviews. *BioEssays*, **36**, 997–1004.
102. Acuna-Hidalgo, R., Veltman, J.A. and Hoischen, A. (2016) New insights into the generation and role of *de novo* mutations in health and disease. *Genome Biol.*, **17**, 241.
103. Mitra, I., Huang, B., Mousavi, N., Ma, N., Lamkin, M., Yanicky, R., Shleizer-Burko, S., Lohmueller, K.E. and Gymrek, M. (2021) Patterns of *de novo* tandem repeat mutations and their role in autism. *Nature*, **589**, 246–250.
104. van Ommen, G.-J.B. (2005) Frequency of new copy number variation in humans. *Nat. Genet.*, **37**, 333–334.
105. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. *et al.* (2007) Strong association of *de novo* copy number mutations with autism. *Science*, **316**, 445–449.
106. The Autism Genome Project Consortium (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat. Genet.*, **39**, 319–328.
107. Itsara, A., Wu, H., Smith, J.D., Nickerson, D.A., Romieu, I., London, S.J. and Eichler, E.E. (2010) *De novo* rates and selection of large copy number variation. *Genome Res.*, **20**, 1469–1481.
108. Kumar, R.A., KaraMohamed, S., Sudi, J., Conrad, D.F., Brune, C., Badner, J.A., Gilliam, T.C., Nowak, N.J., Cook, E.H., Dobyns, W.B. *et al.* (2007) Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.*, **17**, 628–638.
109. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H. *et al.* (2020) A structural variation reference for medical and population genetics. *Nature*, **581**, 444–451.
110. Belyeu, J.R., Brand, H., Wang, H., Zhao, X., Pedersen, B.S., Feusier, J., Gupta, M., Nicholas, T.J., Brown, J., Baird, L. *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.
111. 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
112. Mao, Q., Ciotlos, S., Zhang, R.Y., Ball, M.P., Chin, R., Carnevali, P., Barua, N., Nguyen, S., Agarwal, M.R., Clegg, T. *et al.* (2016) The whole genome sequences and experimentally phased haplotypes of over 100 personal genomes. *Gigascience*, **5**, 42.
113. Telenti, A., Pierce, L.C.T., Biggs, W.H., di Iulio, J., Wong, E.H.M., Fabani, M.M., Kirkness, E.F., Moustafa, A., Shah, N., Xie, C. *et al.* (2016) Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 11901–11906.
114. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A. *et al.* (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, **538**, 201–206.
115. Reuter, M.S., Walker, S., Thiruvahindrapuram, B., Whitney, J., Cohn, I., Sondheimer, N., Yuen, R.K.C., Trost, B., Paton, T.A., Pereira, S.L. *et al.* (2018) The Personal Genome Project Canada: findings from whole genome sequences of the inaugural 56 participants. *CMAJ*, **190**, E126–E136.
116. Wall, J.D., Stawiski, E.W., Ratan, A., Kim, H.L., Kim, C., Gupta, R., Suryamohan, K., Gusareva, E.S., Purbojati, R.W., Bhangale, T. *et al.* (2019) The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*, **576**, 106–111.
117. Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C. *et al.* (2020) Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, **583**, 83–89.
118. Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W. and Salit, M. (2014) Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.*, **32**, 246–251.
119. Eberle, M.A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B.L., Bekritsky, M.A., Iqbal, Z., Chuang, H.-Y., Humphray, S.J., Halpern, A.L. *et al.* (2017) A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.*, **27**, 157–164.
120. Zook, J.M., Hansen, N.F., Olson, N.D., Chapman, L., Mullikin, J.C., Xiao, C., Sherry, S., Koren, S., Phillippy, A.M., Boutros, P.C. *et al.* (2020) A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.*, **38**, 1347–1355.
121. Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E., Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K. *et al.* (2019) Characterizing the major structural variant alleles of the human genome. *Cell*, **176**, 663–675.e19.
122. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P. *et al.* (2020) The mutational constraint

spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.
123. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J. *et al.* (2020) Insights into human genetic variation and population history from 929 diverse genomes. *Science*, **367**, eaay5012.
124. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R. *et al.* (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, **372**, eabf7117.
125. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.-C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R. *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451**, 998–1003.
126. Gudbjartsson, D.F., Helgason, H., Gudjonsson, S.A., Zink, F., Oddson, A., Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B.V., Hjartarson, E. *et al.* (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.*, **47**, 435–444.
127. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
128. Jeon, S., Bhak, Y., Choi, Y., Jeon, Y., Kim, S., Jang, J., Jang, J., Blazyte, A., Kim, C., Kim, Y. *et al.* (2020) Korean Genome Project: 1094 Korean personal genomes with clinical information. *Sci. Adv.*, **6**, eaaz7835.
129. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C. *et al.* (2012) A high-coverage genome sequence from an archaic Denisovan individual. *Science*, **338**, 222–226.
130. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C. *et al.* (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, **505**, 43–49.
131. Sherman, R.M. and Salzberg, S.L. (2020) Pan-genomics in the human genome era. *Nat. Rev. Genet.*, **21**, 243–254.
132. Slon, V., Mafessoni, F., Vernot, B., de Filippo, C., Grote, S., Viola, B., Hajdinjak, M., Peyrégne, S., Nagel, S., Brown, S. *et al.* (2018) The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature*, **561**, 113–116.
133. Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A. *et al.* (2020) Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, **585**, 79–84.
134. Logsdon, G.A., Vollger, M.R., Hsieh, P., Mao, Y., Liskovykh, M.A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P.C., Rhie, A. *et al.* (2021) The structure, function and evolution of a complete human chromosome 8. *Nature*, **593**, 101–107.
135. Sanders, A.D., Hills, M., Porubský, D., Guryev, V., Falconer, E. and Lansdorp, P.M. (2016) Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res.*, **26**, 1575–1587.
136. Sanders, A.D., Falconer, E., Hills, M., Spierings, D.C.J. and Lansdorp, P.M. (2017) Single-cell template strand sequencing by strand-seq enables the characterization of individual homologs. *Nat. Protoc.*, **12**, 1151–1176.
137. Kasianowicz, J.J., Brandin, E., Branton, D. and Deamer, D.W. (1996) Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.*, **93**, 13770–13773.
138. Howorka, S., Cheley, S. and Bayley, H. (2001) Sequence-specific detection of individual DNA strands using engineered nanopores. *Nat. Biotechnol.*, **19**, 636–639.
139. Stoddart, D., Heron, A.J., Mikhailova, E., Maglia, G. and Bayley, H. (2009) Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 7702–7707.
140. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**, 338–345.
141. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L. *et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, **10**, 1784.
142. Nickles, D., Madireddy, L., Yang, S., Khankhanian, P., Lincoln, S., Hauser, S.L., Oksenberg, J.R. and Baranzini, S.E. (2012) In depth comparison of an individual's DNA and its lymphoblastoid cell line using whole genome sequencing. *BMC Genomics*, **13**, 477.
143. McCarthy, N.S., Allan, S.M., Chandler, D., Jablensky, A. and Morar, B. (2016) Integrity of genome-wide genotype data from low passage lymphoblastoid cell lines. *Genomics Data*, **9**, 18–21.
144. Marshall, C.R., Howrigan, D.P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D.S., Antaki, D., Shetty, A., Holmans, P.A., Pinto, D. *et al.* (2017) Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.*, **49**, 27–35.
145. 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.
146. Li, Y.R., Glessner, J.T., Coe, B.P., Li, J., Mohebnasab, M., Chang, X., Connolly, J., Kao, C., Wei, Z., Bradfield, J. *et al.* (2020) Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. *Nat. Commun.*, **11**, 255.
147. Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T. *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.
148. Zhao, X., Collins, R.L., Lee, W.-P., Weber, A.M., Jun, Y., Zhu, Q., Weisburd, B., Huang, Y., Audano, P.A., Wang, H. *et al.* (2021) 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.
149. Trost, B., Engchuan, W., Nguyen, C.M., Thiruvahindrapuram, B., Dolzhenko, E., Backstrom, I., Mirceta, M., Mojarad, B.A., Yin, Y., Dov, A. *et al.* (2020) Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature*, **586**, 80–86.
150. Depienne, C. and Mandel, J.-L. (2021) 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am. J. Hum. Genet.*, **108**, 764–785.
151. Forabosco, A., Percesepe, A. and Santucci, S. (2009) Incidence of non-age-dependent chromosomal abnormalities: a population-based study on 88965 amniocenteses. *Eur. J. Hum. Genet.*, **17**, 897–903.

152. Jacobs, P.A. (2014) An opportune life: 50 years in human cytogenetics. *Annu. Rev. Genomics Hum. Genet.*, **15**, 29–46.
153. Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. and Gouil, Q. (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.*, **21**, 30.
154. Birney, E., Bateman, A., Clamp, M.E. and Hubbard, T.J. (2001) Mining the draft human genome. *Nature*, **409**, 827–828.
155. Butler, D. (2010) Human genome at ten: science after the sequence. *Nature*, **465**, 1000–1001.
156. Lander, E.S. (2011) Initial impact of the sequencing of the human genome. *Nature*, **470**, 187–197.
157. Rehm, H.L. and Fowler, D.M. (2019) Keeping up with the genomes: scaling genomic variant interpretation. *Genome Med.*, **12**, 5.
158. Amendola, L.M., Muenzen, K., Biesecker, L.G., Bowling, K.M., Cooper, G.M., Dorschner, M.O., Driscoll, C., Foreman, A.K.M., Golden-Grant, K., Greally, J.M. *et al.* (2020) Variant classification concordance using the ACMG-AMP variant interpretation guidelines across nine genomic implementation research studies. *Am. J. Hum. Genet.*, **107**, 932–941.
159. MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L. and Scherer, S.W. (2013) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, D986–D992.
160. Miller, N.A., Farrow, E.G., Gibson, M., Willig, L.K., Twist, G., Yoo, B., Marrs, T., Corder, S., Krivohlavek, L., Walter, A. *et al.* (2015) A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Med.*, **7**, 100.
161. Hasin, Y., Seldin, M. and Lusis, A. (2017) Multi-omics approaches to disease. *Genome Biol.*, **18**, 83.
162. Basel-Salmon, L., Orenstein, N., Markus-Bustani, K., Ruhrman-Shahar, N., Kilim, Y., Magal, N., Hubshman, M.W. and Bazak, L. (2019) Improved diagnostics by exome sequencing following raw data reevaluation by clinical geneticists involved in the medical care of the individuals tested. *Genet. Med. Off. J. Am. Coll. Med. Genet.*, **21**, 1443–1451.
163. Rockowitz, S., LeCompte, N., Carmack, M., Quitadamo, A., Wang, L., Park, M., Knight, D., Sexton, E., Smith, L., Sheidley, B. *et al.* (2020) Children's rare disease cohorts: an integrative research and clinical genomics initiative. *NPJ Genom. Med.*, **5**, 29.
164. Quaio, C.R.D.C., Moreira, C.M., Novo-Filho, G.M., Sacramento-Bobotis, P.R., Groenner Penna, M., Perazzio, S.F., Dutra, A.P., da Silva, R.A., Santos, M.N.P., de Arruda, V.Y.N. *et al.* (2020) Diagnostic power and clinical impact of exome sequencing in a cohort of 500 patients with rare diseases. *Am. J. Med. Genet. C Semin. Med. Genet.*, **184**, 955–964.
165. Eddy, S., Mariani, L.H. and Kretzler, M. (2020) Integrated multi-omics approaches to improve classification of chronic kidney disease. *Nat. Rev. Nephrol.*, **16**, 657–668.
166. Montaner, J., Ramiro, L., Simats, A., Tiedt, S., Makris, K., Jickling, G.C., Debette, S., Sanchez, J.-C. and Bustamante, A. (2020) Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke. *Nat. Rev. Neurol.*, **16**, 247–264.
167. Nam, A.S., Chaligne, R. and Landau, D.A. (2021) Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat. Rev. Genet.*, **22**, 3–18.
168. Harrison, S.M., Riggs, E.R., Maglott, D.R., Lee, J.M., Azzariti, D.R., Niehaus, A., Ramos, E.M., Martin, C.L., Landrum, M.J. and Rehm, H.L. (2016) Using ClinVar as a resource to support variant interpretation. *Curr. Protoc. Hum. Genet.*, **89**, 8.16.1–8.16.23.
169. Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D. and Cooper, D.N. (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.*, **136**, 665–677.
170. Higgins, J., Dalgleish, R., den Dunnen, J.T., Barsh, G., Freeman, P.J., Cooper, D.N., Cullinan, S., Davies, K.E., Dorkins, H., Gong, L. *et al.* (2021) Verifying nomenclature of DNA variants in submitted manuscripts: guidance for journals. *Hum. Mutat.*, **42**, 3–7.
171. Roser, M., Ritchie, H. and Ortiz-Ospina, E. (2019) World population growth. In *Our World in Data*. https://ourworldindata.org/grapher/births-and-deaths-projected-to-2100.
172. Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E. *et al.* (2006) Copy number variation: new insights in genome diversity. *Genome Res.*, **16**, 949–961.
173. Church, D.M., Lappalainen, I., Sneddon, T.P., Hinton, J., Maguire, M., Lopez, J., Garner, J., Paschall, J., DiCuccio, M., Yaschenko, E. *et al.* (2010) Public data archives for genomic structural variation. *Nat. Genet.*, **42**, 813–814.
174. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
175. Rigden, D.J. and Fernández, X.M. (2020) The 27th annual Nucleic Acids Research database issue and molecular biology database collection. *Nucleic Acids Res.*, **48**, D1–D8.