Patterns of mosaicism for sequence and copy-number variants discovered through clinical deep sequencing of disease-related genes in one million individuals

Authors

Rebecca Truty, Susan Rojahn, Karen Ouyang, ..., Leslie Burnett, Robert L. Nussbaum, Swaroop Aradhya

Correspondence

swaroop.aradhya@invitae.com

Truty et al. describe mosaic sequence and copy number variants identified through genetic testing. Nearly 6,000 variants across >500 genes contributed to ~2% of molecular diagnoses. Mosaic variants were mostly in cancer-related genes, at higher levels in younger individuals, and appeared to correlate with later disease onset or milder phenotypes.

Truty et al., 2023, The American Journal of Human Genetics *110*, 551–564
April 6, 2023
https://doi.org/10.1016/j.ajhg.2023.02.013

CellPress

Patterns of mosaicism for sequence and copy-number variants discovered through clinical deep sequencing of disease-related genes in one million individuals

Rebecca Truty,[1] Susan Rojahn,[1] Karen Ouyang,[1] Curtis Kautzer,[1] Michael Kennemer,[1] Daniel Pineda-Alvarez,[1] Britt Johnson,[1] Amanda Stafford,[1] Lina Basel-Salmon,[2,3,4] Sulagna Saitta,[5] Anne Slavotinek,[6] Settara C. Chandrasekharappa,[7] Carlos Jose Suarez,[8] Leslie Burnett,[9] Robert L. Nussbaum,[1,10] and Swaroop Aradhya[1,8,*]

Summary

DNA variants that arise after conception can show mosaicism, varying in presence and extent among tissues. Mosaic variants have been reported in Mendelian diseases, but further investigation is necessary to broadly understand their incidence, transmission, and clinical impact. A mosaic pathogenic variant in a disease-related gene may cause an atypical phenotype in terms of severity, clinical features, or timing of disease onset. Using high-depth sequencing, we studied results from one million unrelated individuals referred for genetic testing for almost 1,900 disease-related genes. We observed 5,939 mosaic sequence or intragenic copy number variants distributed across 509 genes in nearly 5,700 individuals, constituting approximately 2% of molecular diagnoses in the cohort. Cancer-related genes had the most mosaic variants and showed age-specific enrichment, in part reflecting clonal hematopoiesis in older individuals. We also observed many mosaic variants in genes related to early-onset conditions. Additional mosaic variants were observed in genes analyzed for reproductive carrier screening or associated with dominant disorders with low penetrance, posing challenges for interpreting their clinical significance. When we controlled for the potential involvement of clonal hematopoiesis, most mosaic variants were enriched in younger individuals and were present at higher levels than in older individuals. Furthermore, individuals with mosaicism showed later disease onset or milder phenotypes than individuals with non-mosaic variants in the same genes. Collectively, the large compendium of variants, disease correlations, and age-specific results identified in this study expand our understanding of the implications of mosaic DNA variation for diagnosis and genetic counseling.

Introduction

A change in the genome that occurs after conception can be propagated through different cellular lineages, resulting in an individual who carries the change in only some tissues and at varying levels.[1] The developmental stage during which a post-zygotic variant arises and the cellular lineages in which it is propagated ultimately determines the tissues, and the proportion of cells within those tissues, that will harbor the mosaic variant. Most post-zygotic changes are benign, and somatic mosaicism for DNA variants is common in human genomes.[2] However, mosaic variants that occur in genes associated with Mendelian diseases may influence how those conditions arise, manifest, progress, and transmit.[1,3] For example, a mosaic disease-causing variant can result in later-onset or milder clinical signs than those resulting from a non-mosaic version of the same variant.[4,5] If a mosaic variant populates cell lineages that eventually contribute to gonadal tissue development (i.e., germline mosaicism), it may be transmitted to the next generation as a non-mosaic variant. Such an observation can be mistaken for a *de novo* event, thereby confounding the estimation of recurrence risk, if mosaicism in one of the affected individuals' parents is not experimentally confirmed.[6]

Mosaicism in clinically affected individuals has been reported in genes associated with both X-linked disorders (e.g., *PCDH19* [MIM: 300460], *IKBKG* [MIM: 300248], *COL4A5* [MIM: 303630]) and autosomal disorders (e.g., *VHL* [MIM: 608537], *PAX6* [MIM: 607108], *NF1* [MIM: 613113], *NIPBL* [MIM: 608667]), ranging from cancer syndromes and neurological disorders to syndromic developmental disorders. Mosaic variants have been most extensively investigated in hereditary cancer syndromes[7–9] and in certain pediatric overgrowth syndromes.[10,11] When a variant appears to be mosaic in a blood or saliva sample submitted from an individual with cancer, it can be difficult to determine if the DNA change originated during early development and is present in different tissues (constitutional mosaicism) or if it occurred later in life,

[1]Invitae, 1400 16th Street, San Francisco, CA 94103, USA; [2]Rabin Medical Center-Beilinson Hospital and Schneider Children's Medical Center of Israel, Petach Tikva, Israel; [3]Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel; [4]Felsenstein Medical Research Center, Petach Tikva, Israel; [5]Division of Clinical Genetics, Departments of Pediatrics and Obstetrics and Gynecology, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA; [6]Division of Human Genetics, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; [7]Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA; [8]Department of Pathology, Stanford University School of Medicine, Stanford, CA 94301, USA; [9]Invitae Australia, Sydney, NSW, Australia; [10]School of Medicine, University of California - San Francisco, San Francisco, CA, USA
*Correspondence: swaroop.aradhya@invitae.com
https://doi.org/10.1016/j.ajhg.2023.02.013.

either within malignant tissue itself or in a restricted cell lineage. This can be especially challenging to differentiate for variants in genes such as *TP53* (MIM: 191170), which is associated with both early- and late-onset malignancies.[9,12,13] Another confounding factor is clonal hematopoiesis (CH), in which a subpopulation of blood cells develop a growth advantage because they acquire and propagate new variants, particularly in cancer-related genes.[14,15] CH is observed as a natural phenomenon in adults, and it is often related to mosaic variants in *TP53*, particularly in older adults.[16,17]

Mosaicism in genes involved in hereditary disease is under-recognized, in part due to the limited ability of traditional sequencing methods to detect mosaic variants that are present in a limited population of cells.[18–21] As genetic testing has shifted toward next-generation sequencing (NGS), which affords high sensitivity for detecting variants, more cases of mosaicism are being discovered.[21–23] A few studies of mosaicism in clinical samples have been reported,[21,22,24] including a recent retrospective analysis of exome-sequencing data from nearly 12,000 individuals that identified more than 70 clinically significant mosaic variants.[24] This and other studies have focused mostly on single-nucleotide variants (SNVs), so the prevalence of mosaic intragenic copy-number variants (CNVs) in clinical samples has not been elucidated. Analysis of deep sequencing data from many disease-related genes and individuals is needed to construct a more accurate picture of clinically relevant mosaicism and its implications.

We evaluated high depth-of-coverage NGS data from a clinical cohort of one million individuals referred for genetic testing to determine the frequency and types of mosaicism in clinical samples, distribution of mosaicism across gene and disease types, transmission of mosaic variants among family members, clinical implications of mosaic variants, and technical complexities of detecting mosaicism. The results of this extensive analysis of mosaic sequence and intragenic copy-number variants provide insights that can better inform expectations for genetic testing and cascade testing of family members, help clinicians recognize correlations between mosaicism and phenotypic severity, and provide useful context to inform clinical management and genetic counseling for individuals with such variants.

Subjects and methods

Individuals referred for genetic testing

The cohort consisted of individuals referred for genetic testing at a clinical laboratory (Invitae). Multi-gene panels were ordered by physicians mainly for diagnostic purposes for clinically affected individuals; panel testing was also ordered for unaffected individuals who had a strong family history of cancer or for reproductive carrier screening purposes. Clinicians do not routinely and consistently provide information about the purpose of testing to testing laboratories, precluding stratification of affected and unaffected individuals. However, based on the pattern of referrals for gene panel testing, we estimate the proportion of clinically affected individuals at 75% of the entire cohort in this study. Peripheral blood, saliva, or genomic DNA samples accessioned between March 2015 and August 2020 from individuals aged 0–90 years were received for diagnostic NGS panel testing for a range of hereditary disorders or for reproductive carrier screening. The few individuals >90 years old were placed into a single group of 90-year-olds per a privacy requirement. Women represented 73% of the cohort since a majority of genetic tests were performed for hereditary cancer or for reproductive carrier screening, both of which are sought for women more often than for men. Informed consent was obtained by health care providers (HCPs), who also submitted demographic and clinical data for individuals in their care through a test order form or online portal. De-identified data were retrieved from internal databases with institutional review board approval (WCG IRB, #20160282). In some cases, a sample from a second tissue type (e.g., skin, buccal cells) was tested to evaluate whether a reported mosaic variant was a constitutional change. Only data from unrelated probands were analyzed in this study, except in examinations of parent-to-child transmissions of variants.

Next-generation sequencing and variant calling

Each DNA sample was tested on the NGS panel(s) requested by the HCP to identify sequence variants and intragenic CNVs (i.e., exon-level deletions or duplications), as described previously.[25,26] In the curated targeted gene panels (not partitions or slices from exome or genome sequencing), an optimized distribution of oligonucleotide baits (Agilent Technologies, Roche, IDT, Twist Bioscience) was designed to capture the coding exons, 10–20 bases of flanking intronic sequence, and certain non-coding regions of clinical interest. Across the entire clinical cohort, 1,892 unique genes were analyzed. All primary sequencing was performed on HiSeq or NovaSeq instruments (Illumina) with a minimum depth of coverage of 50× (mean, 350×).

SNVs, small insertions and deletions (indels), large indels, structural variants with breakpoints in target sequences, and CNVs were identified using a suite of bioinformatics tools.[26–28] Sequence reads were aligned using NovoAlign, and SNVs and small indels were called using a modified Genome Analysis Toolkit (GATK) HaplotypeCaller. Mosaic sequence variants were identified based on the observed allele balance (AB)—the number of reads containing a specific allele divided by the total number of reads aligning to the specific genomic locus—and on a gene-specific threshold that differentiated mosaic from non-mosaic heterozygous variants. Sequence variants with ABs ranging from 0.06 to 0.4 on the primary Illumina-based NGS assay were evaluated as possibly mosaic. Mosaic CNVs were determined by manual inspection. NGS and bioinformatics methods for detecting all mosaic variants were validated in a series of experiments described in the supplemental methods, Figures S1–S4, and Tables S1 and S2. Clinically significant mosaic variants that met established internal criteria, as well as some technically challenging non-mosaic variants, underwent orthogonal confirmation with PacBio (Pacific Biosciences) or Sanger sequencing, multiplex ligation-dependent probe amplification-based sequencing (MLPASeq), or exon-focused microarray-based comparative genomic hybridization (exon array CGH).[27] The PacBio sequencing was performed to a minimum of 50× depth of sequence coverage to qualitatively confirm the presence of the variant.

Clinical variant interpretation

Variants were classified according to their clinical significance using Sherloc, a validated variant classification system based on guidelines from the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology.[29,30] Sherloc employs a semi-quantitative framework for evaluating and combining clinical, functional, and computational evidence to determine classifications. Sherloc also incorporates phenotype evidence for cases in which an individual has clinical features that are highly predictive of a disorder.[31]

Our analyses included all mosaic variants classified as pathogenic (P), likely pathogenic (LP), or variant(s) of uncertain significance (VUSs), based on the tiered classification system prescribed by the ACMG. *FMR1* premutations and increased risk alleles (IRAs) were included with P/LP variants. The ClinGen resource defines IRAs as variants with very low penetrance such that their effects are incomplete and do not necessarily manifest in a Mendelian pattern of inheritance.[32] Molecular diagnoses were defined as one P/LP variant in a gene associated with an autosomal-dominant or X-linked disease or two P/LP variants in *trans* in the same gene associated with a rare autosomal-recessive disease. For the purpose of analyzing molecular diagnoses involving mosaic variants, we also considered one P/LP variant and one VUS in *trans* in a gene associated with a rare autosomal-recessive disease as a likely positive result due to the higher prior probability that such a variant combination explains disease etiology. A molecular diagnosis does not necessarily constitute a clinical diagnosis and can be made without details such as an individual's clinical phenotype.[29] A compendium of the variants described in this study is available in Table S3.

Analyses

Clinical areas were designated as hereditary cancer, pediatrics and rare disease, neurology, cardiology, and reproductive carrier screening. Pediatrics and rare disease encompassed disorders related to pediatric congenital anomalies, epilepsy, and neurodevelopmental disorders as well as those related to metabolic disorders, immunology, ophthalmology, hematology, and dermatology. Since genes were sometimes included in multiple panels associated with different clinical specialties (e.g., hereditary cancer and cardiology), the results were categorized by both (1) the gene-clinical area, which was an internal classification based on gene-disease associations and inclusion of genes in Invitae's panels and (2) the test-clinical area, which was the primary clinical specialty to which the HCP-ordered panel(s) belonged. To examine the effect of mosaicism on phenotype, we compared phenotype-related Sherloc evidence codes (based on clinical information submitted by the ordering HCP or in the literature) applied during variant interpretation to estimate the difference in clinical presentations in individuals with and without mosaic variants in genes associated with distinctive disorders, as described previously.[31]

Relationships between mosaicism and age were explored at both the variant level and at the individual person level. Descriptive statistics were used to compare the mean age at testing between individuals with mosaic variants and those with non-mosaic variants. A t test was used to assess the significance of the level of mosaicism (as measured by AB) between individuals with mosaic versus non-mosaic variants and between individuals <18 years of age and those ≥ 18 years of age; significance was set at $p < 0.05$.

Results

Prevalence of mosaicism in clinical samples

Mosaic variant detection methods were applied to NGS sequencing results from 1,034,580 unrelated individuals referred for clinical genetic testing, representing an equivalent of 68,360,003 single-gene tests. Various characteristics of the cohort are shown in Table S4; two-thirds of these individuals were referred for hereditary cancer testing, and the rest were referred for pediatrics and rare disease testing, reproductive carrier screening, cardiology testing, or neurology testing.

We observed 5,939 mosaic variants in 5,695 individuals, representing 0.6% of individuals in the overall cohort. These variants were detected in 509 (26.9%) of the 1,892 genes sequenced. We did not observe any difference in the rate of mosaic variants between peripheral blood and saliva samples. The vast majority of individuals with mosaic variants had just one, but a small proportion (4%) had two or more (Table S5). Nearly 90% of individuals with multiple somatic mosaic variants were at least 50 years of age, and the variants were predominantly in genes related to hereditary cancer (mostly *TP53* and *ATM* [MIM: 607585] and, to a lesser extent, *CHEK2* [MIM: 604373] and *NF1*). Half of the individuals with multiple mosaic variants had two or three within a single gene, often clustered within *TP53* and *ATM*.

Of the 5,939 mosaic variants detected, 5,046 (85%) were SNVs, 690 (11.6%) were indels, and 203 (3.4%) were CNVs. Most mosaic variants were missense changes, present in genes associated with autosomal-dominant disorders (Figure 1), and unique to one individual. The mean sequencing depth and the mean AB at the positions of the mosaic SNVs and indels was 625× (range, 33×–4,715×) and 0.16 (range, 0.06–0.94), respectively. The 203 mosaic CNVs included 149 deletions and 54 duplications; two-thirds of these CNVs included the complete coding sequence and the remaining were partial-gene events (with 12 single-exon and 6 promoter CNVs). Half of the mosaic CNVs were in cancer-related genes and most of the remaining were in genes linked to pediatrics and rare disease.

Tissue distribution

Although most HCPs did not pursue the laboratory's standard offer to test an additional tissue type when a mosaic variant was identified, we were able to compare findings from two tissue types in 74 individuals with mosaic variants. Seventeen of these individuals harbored the variant in a mosaic state in the second tissue type, indicating constitutional mosaicism. The AB range did not vary markedly between the two tested tissues when a mosaic variant was present in both. In the remaining 53 individuals, the originally identified mosaic variant was absent in the second tissue type. The AB range also did not vary between cases in which a mosaic variant

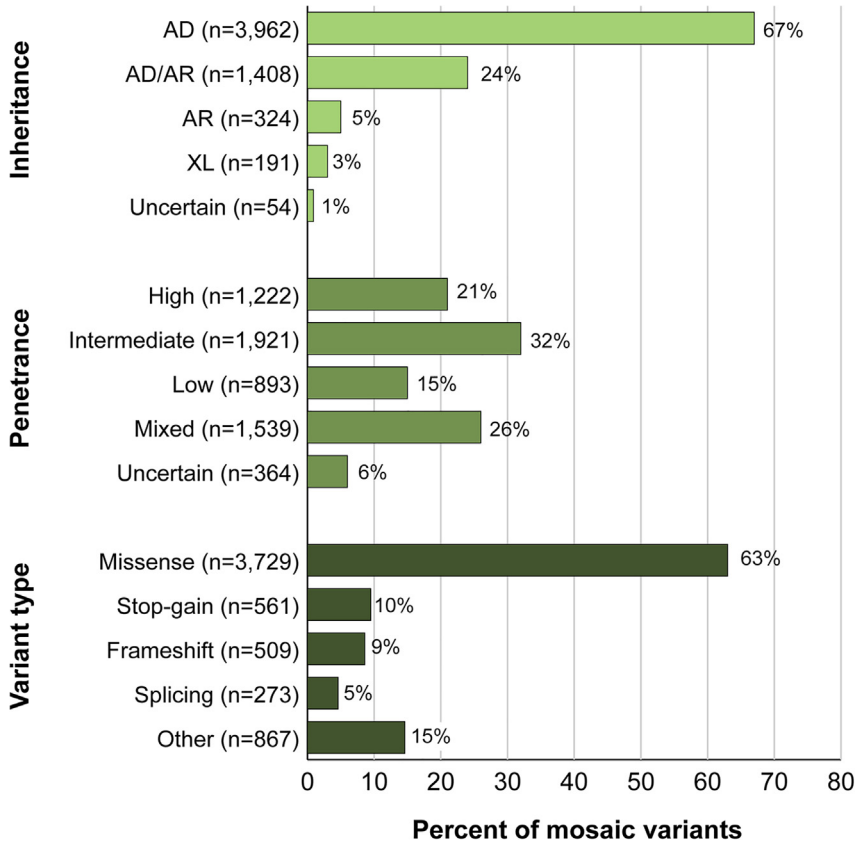


Figure 1. Distribution of mosaic variants shown by disease inheritance, penetrance, and variant type

n = 5,939. AD, autosomal dominant; AR, autosomal recessive; XL, X-linked.

was present in two tested tissues and cases in which only the first sample had a mosaic variant. The four remaining cases had the originally observed mosaic variant in an apparently non-mosaic state in the second tissue type, although in three of these four cases the AB for the variant was lower than typical for a non-mosaic variant, suggesting that these were also possibly constitutional changes.

Clinical interpretation of mosaic variants

Out of 5,939 mosaic SNVs or CNVs identified, 2,323 (39%) were classified as P/LP and the remaining as VUSs. These 5,939 mosaic variants represented 0.7% of all P/LP/VUS results in the clinical cohort, and mosaic P/LP variants in particular contributed to 1% of all P/LP results. Among the 203 mosaic CNVs observed, 145 were classified as P/LP and the remaining as VUSs; almost all of the mosaic CNVs classified as VUSs were duplications.

Of the 2,323 mosaic P/LP variants, 76% were in genes associated with autosomal-dominant inheritance, mostly reflecting a bias in the types of genes offered for genetic testing. Genes associated with autosomal-recessive inheritance or X-linked inheritance each had roughly 5% or less of the mosaic P/LP variants. Less than 1% of mosaic P/LP variants were in genes associated with autosomal-recessive inheritance and in combination with a non-mosaic P/LP variant, and mosaic P/LP variants in X-linked genes were found in a comparable number of male and female individuals.

The correlation between a mosaic variant and disease can be difficult to establish, particularly when additional variants are detected in the same gene or in different genes in a single individual. Among all mosaic variants, 88% were the sole P/LP/VUS result in an individual, but the remaining were found alongside one or more non-mosaic variants (P/LP or VUS) in the same gene or a different gene or, in a few instances, alongside other mosaic variants.

A small proportion (4.8%) of the mosaic variants were observed alongside a non-mosaic variant within the same gene (Table S6). In 83 instances, this combination involved a mosaic VUS and a non-mosaic P/LP in a gene associated with autosomal-dominant inheritance; therefore, the mosaic variant likely was not disease causing. In another 48 instances, a mosaic P/LP variant was observed with a non-mosaic VUS in the same gene associated with an autosomal-dominant disorder; therefore, the mosaic variant was a probable explanation for disease. In genes associated with autosomal-recessive disorders, there were six instances of a mosaic P/LP variant and a non-mosaic P/LP variant (two instances involving *ACADM*); in these cases, the mosaic variant was a likely explanation for disease. Finally, we identified two instances of a mosaic P/LP variant in combination with a non-mosaic P/LP variant in an X-linked gene (*GATA1* [MIM: 305371] and *NEXMIF* [MIM: 300524]), both in female individuals. In the first case, the proband had trisomy 21 and transient myeloproliferative disorder; it has been reported that somatic mutations in *GATA1* can drive the development of myeloid leukemogenesis in Down syndrome.[33] In the second case, pathogenic variants in *NEXMIF* have been reported to cause disease in female individuals with unfavorable X-inactivation, but in this case the presence of two truncating variants (one mosaic and one non-mosaic) in *trans* likely explained the clinical diagnosis of seizures and developmental delay.[34]

In the context of these results, mosaic variants appeared to contribute to a molecular diagnosis in 2,182 probands, representing 1.8% of all 121,710 molecular diagnoses in this clinical cohort (separately, mosaic variants in 38 individuals accounted for <0.1% of all positive carrier screening results).

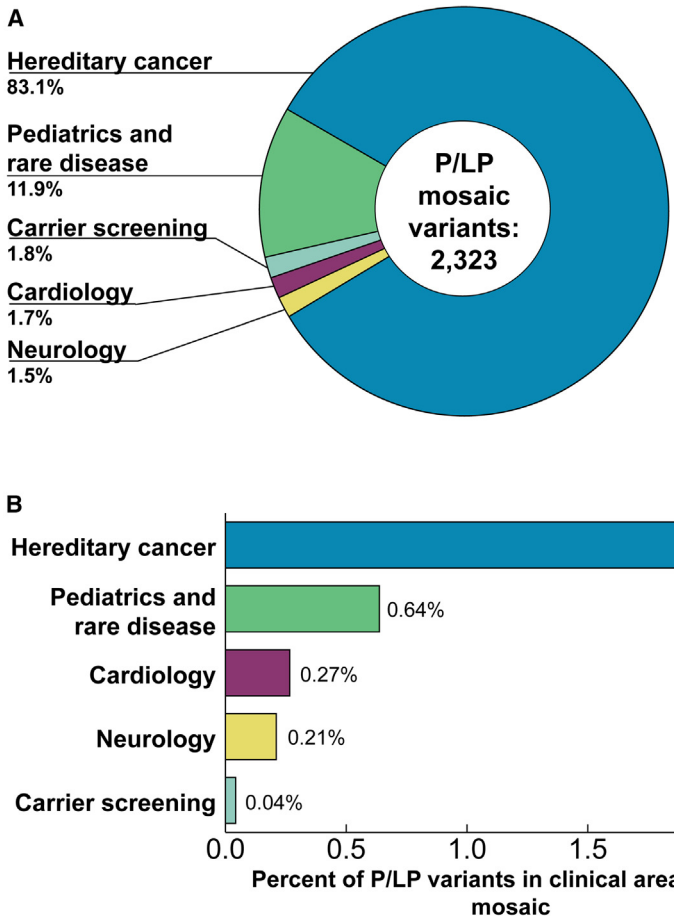


Figure 2. Prevalence of mosaic P/LP variants by clinical area

(A) Distribution of all observed mosaic P/LP variants (including premutation and increased-risk alleles) across test clinical areas.

(B) Percentage of P/LP variants within a test clinical area that were mosaic.

Note that in (A), the distribution of mosaic variants is influenced by the number of individuals tested in each clinical area, while in (B) the distribution of mosaic variants is normalized by the number of individuals tested in each clinical area. The number of individuals in each test clinical area is shown in Table S4. Peds, pediatrics; P/LP, pathogenic/likely pathogenic.

Mosaicism across clinical areas

The majority of mosaic P/LP variants were in individuals referred for hereditary cancer testing (Figures 2A and 3). Roughly 2% of all types of P/LP variants in cancer-related cases were mosaic (Figure 2B). Most of these variants were in *TP53, NF1*, *ATM*, or *CHEK2* and were found among individuals with a median age of 68 years (Figure 3), likely indicating age-related somatic variation. We also observed more than 10 mosaic P/LP variants each in *APC* (MIM: 611731), *RB1* (MIM: 614041), *PTEN* (MIM: 601728), *BRCA2* (MIM: 600185), *STK11* (MIM: 602216), *PTCH1* (MIM: 601309), *CDKN1C* (MIM: 600856), *BRCA1* (MIM: 113705), *NBN* (MIM: 602667), *GATA1*, and *NF2* (MIM: 607379), found in individuals with a median age of 52 years. Another 49 genes had fewer than 10 mosaic P/LP variants each.

The second largest group of mosaic P/LP variants were in individuals referred for genetic testing for pediatric disorders and for rare diseases such as immune deficiencies or ophthalmological disorders (Figure 3). Notably, several individuals had mosaic P/LP variants in genes associated with early-onset cancer syndromes; these variants were found more often among individuals referred for a pediatrics and rare disease gene panel than among those referred for a hereditary cancer or cardiology gene panel (Figure S5). Lastly, relatively few mosaic P/LP variants were observed in individuals referred for genetic testing for cardiovascular or neurological disorders (Figure 3).

After normalizing the number of mosaic P/LP variants observed in each clinical area by the number of individuals tested in that clinical area, hereditary cancer still had the highest frequency of mosaic P/LP variants, followed by pediatrics and rare disease, neurology, and cardiology. For genes in which mosaic P/LP variants were detected, the mosaic P/LP variants contributed to as little as $<0.01\%$ of all P/LP results in the gene (in *CFTR*) to as much as 100% (in *AKT1* [MIM: 164730]) (Figures 4A–4E). Hereditary cancer genes were more than four times as likely as genes associated with other clinical disorders to have a mosaic P/LP variant (47% vs. $\leq 11\%$) (Table S7).

Mosaicism in carrier screening

Among 94,899 individuals who underwent reproductive carrier screening for autosomal-recessive or X-linked disease, only 38 harbored at least one mosaic P/LP variant: 8 mosaic P/LP variants were in *FMR1* (MIM: 300805), 2 were in *DMD* (MIM: 300377), and 31 others were individually present in other autosomal or X-linked genes. With one exception (in *FMR1*), mosaic variants were not observed during carrier screening in commonly tested and professional guidelines-recommended genes, such as *CFTR* (MIM: 602421), *SMN1* (MIM: 600354), and *HBA1* (MIM: 141800).

Mosaicism in X-linked genes

Of the 191 mosaic variants detected specifically in X-linked genes through diagnostic testing or carrier screening, 110 were in euploid males (ages 0–90 years) and 81 in euploid females (ages 0–89 years). The ABs for mosaic variants in X-linked genes ranged from 0.08 to 0.94 in male individuals and from 0.08 to 0.58 in female individuals. Mosaic

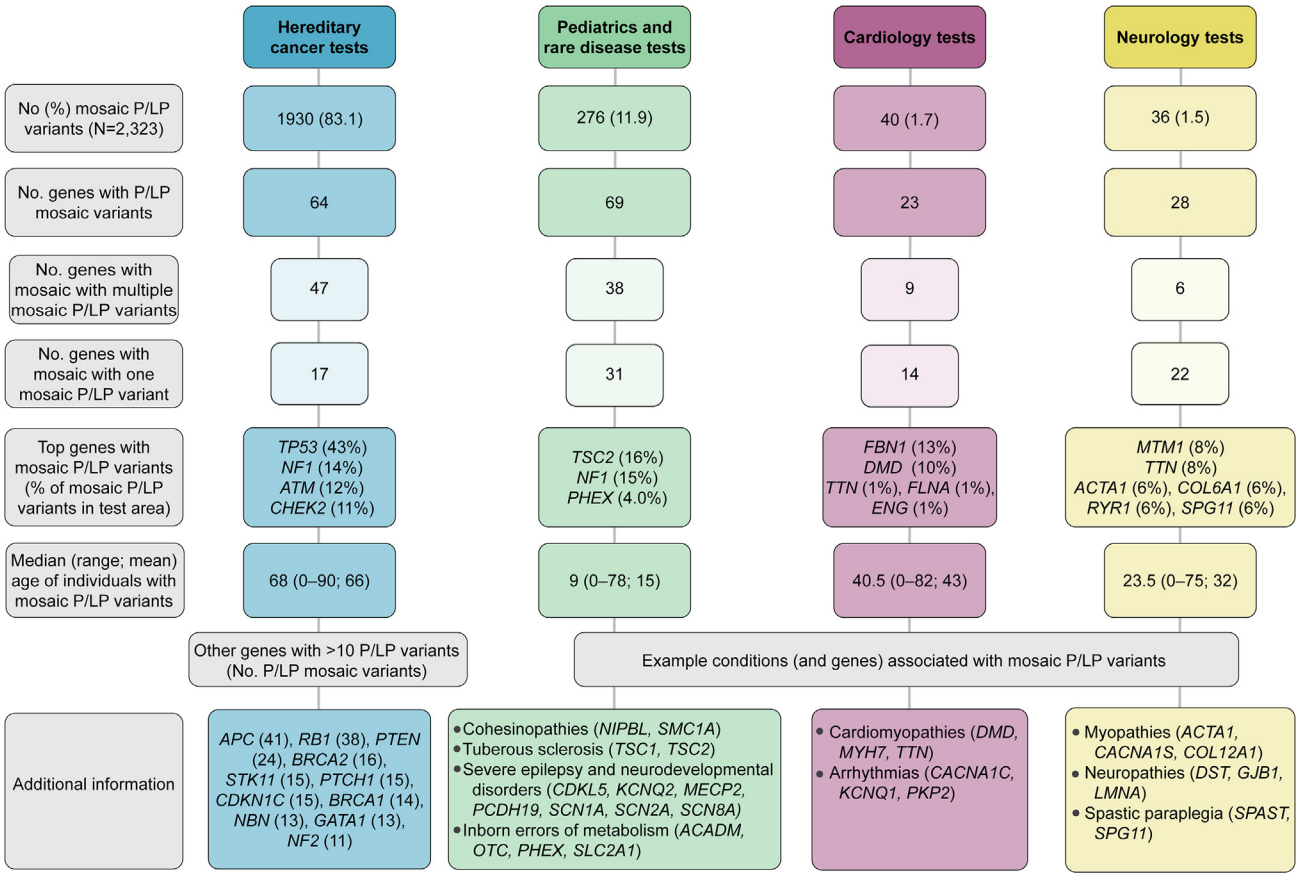


Figure 3. Mosaic P/LP variants by test clinical area
Summary information about mosaic P/LP variants observed in individuals referred for genetic testing for hereditary cancer, pediatric conditions or rare diseases, cardiac conditions, or neurological conditions. Some genes are associated with multiple clinical areas (e.g., *TTN* is found in panels related to both cardiological and neurological disorders). P/LP, pathogenic/likely pathogenic.

variants were found in 46 X-linked genes, primarily *PHEX, GPC3* (MIM: 300037)*, CDKL5* (MIM: 300203)*, GATA1, PCDH19, WDR45* (MIM: 300526), *DMD*, *SMC1A* (MIM: 300040), *FMR1*, *FLNA* (MIM: 300017), and *MECP2* (MIM: 300005), and more than half of these variants were classified as P/LP. Eleven mosaic P/LP variants in X-linked genes were identified through carrier screening: 8 in *FMR1*, 2 in *DMD*, and 1 in *ATRX* (MIM: 300032). Three individuals each carried two mosaic variants in *FMR1,* either as a combination of two unique premutation mosaic alleles and a normal repeat non-mosaic allele (in one female individual) or as a combination of one premutation mosaic allele and a mosaic full-mutation allele (in one female and one male individual).

Transmission of mosaic variants

To confirm the expected *de novo* occurrence of mosaic variants, we examined results for 103 probands with mosaic P/LP or mosaic VUS whose parents were both available for targeted variant testing. In 99 of the cases, neither parent harbored the mosaic variant, confirming *de novo* occurrence in the probands. For all four remaining probands, one parent harbored the variant as a non-mosaic heterozygous change; the genes involved in these four cases were *LIG4* (MIM: 601837), *KANSL1* (MIM: 612452), *WDR45*, and *FANCA* (MIM: 607139). Further examination revealed that the *LIG4* case likely involved mosaic uniparental disomy in a proband who had two mosaic pathogenic truncating variants in *trans*, each inherited from a non-mosaic heterozygous parent. The *KANSL1* and *WDR45* cases were likely due to variant calling artifacts caused by segmental duplication and a co-existing large X chromosome abnormality in the proband, respectively. Lastly, the *FANCA* case appeared to be an example of mosaicism resulting from somatic reversion of the variant in the proband.

Separately, for 216 individuals who had a mosaic P/LP variant or mosaic VUS, we were able to test at least one sibling. The proband and the sibling shared the same variant in only two families, and in both instances the variant was non-mosaic in the sibling. Additional analysis suggested suboptimal sequencing due to a hematologic malignancy in one proband involving *TERC* and a very rare case of mosaic maternal uniparental disomy confirmed through parental testing in the aforementioned case involving *LIG4*.

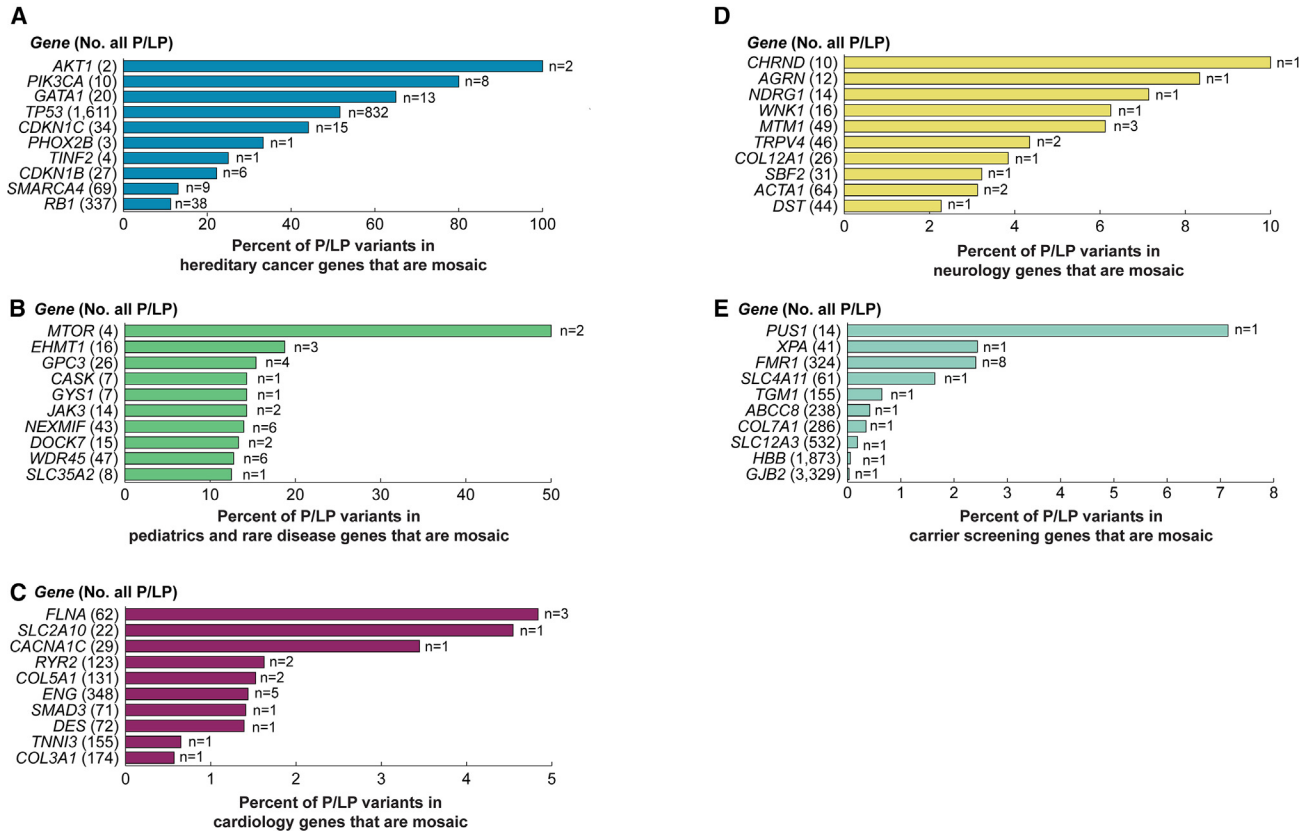


Figure 4. Percentage of all observed P/LP variants that were mosaic in the 10 genes with the highest mosaic burden for each gene clinical area

(A) Hereditary cancer.
(B) Pediatrics and rare disease.
(C) Cardiology.
(D) Neurology.
(E) Carrier screening.

Note that each panel uses a different scale on the x axis. The number in parentheses after each gene name denotes the number of P/LP variants (mosaic and non-mosaic) detected. The number to the right of each bar denotes the number of mosaic P/LP variants detected. Peds, pediatrics; P/LP, pathogenic/likely pathogenic.

We next explored variant transmission rates in 376 probands with a mosaic P/LP variant who had at least one offspring tested. Among 563 offspring tested, 549 did not have the parental mosaic variant. The remaining 14 offspring inherited the parental variant in a non-mosaic heterozygous state, as would be expected from germline transmission from a germline mosaic parent. The level of mosaicism in blood samples submitted for transmitting parents ranged from 0.10 to 0.31 AB. The parental variants inherited by the 14 offspring were in 10 genes: *ACTA1*, *LMNA* (2 individuals), *MYH11*, *TSC2*, *TTN* (2 individuals), *BRCA1*, *DSC2*, *PHEX* (2 individuals), *RET*, and *STK11* (2 individuals).

Mosaicism and clinical presentation

To explore the relationship between mosaicism and clinical phenotype, we focused on 25 individuals who had a mosaic variant in one of 13 genes associated with a distinctive phenotype that had been previously curated and reported to predict variant pathogenicity (Table S8).[31] The hypothesis we addressed was that the clinical phenotype information available for individuals with mosaic variants was less specific or insufficiently representative of the classic phenotype in comparison to the information available for individuals with non-mosaic variants. Indeed, the phenotype information reported for these 25 individuals, as determined by the phenotype-related evidence codes applied during variant interpretation, was less specific than that reported for 8,788 individuals who had a non-mosaic variant in one of the same genes ($p = 1e{-}12$).

Mosaicism and age

The percentage of mosaic SNVs appeared to increase overall with the age of the individuals, with the majority present in those older than 50 years (Table 1), but mosaic CNVs did not show this pattern. However, when mosaic findings in hereditary cancer-related genes were excluded in order to negate any potential effects of CH, we observed that both mosaic SNVs and mosaic CNVs were enriched in the youngest individuals.

Table 1. Mosaic variants by variant type and age

	No. mosaic variants[a]	No. (%) mosaic variants observed in individuals <50 years	No. (%) mosaic variants observed in individuals 50–65 years old	No. (%) mosaic variants observed in individuals >65 years old
SNV—all genes	5,046	957 (19)	1,270 (25)	2,819 (56)
CNV—all genes	203	77 (38)	39 (19)	87 (43)
SNV—hereditary cancer genes and tests excluded	785	461 (59)	111 (14)	213 (27)
CNV—hereditary cancer genes and tests excluded	66	49 (74)	6 (9)	11 (17)

The numbers in the lower two rows exclude variants in genes with a primary clinical area of hereditary cancer and variants in individuals whose test referral clinical area was hereditary cancer. SNV, single-nucleotide variant; CNV, copy-number variant.
[a]Includes variants classified as pathogenic or likely pathogenic and uncertain.

Overall, the percentage of individuals who carried any type of mosaic variant increased with age across clinical areas (Figure 5). However, when we restricted our analysis to variants clinically classified as P/LP (i.e., excluding VUSs), the only clinical area in which the mean age at testing was higher for individuals with mosaic variants than for those with non-mosaic variants was hereditary cancer, in which CH was a likely and prominent contributor to mosaicism (Table 2). In all other clinical areas, the mean age of individuals with mosaic P/LP variants was, overall, comparable to the mean age of those with non-mosaic P/LP variants, even when individuals ≥50 years old and individuals who had variants in genes associated with hereditary cancer were excluded.

We also uncovered relationships between level of mosaicism and age among all individuals with mosaic P/LP variants. Individuals who were <18 years of age had a significantly higher level of mosaicism, based on ABs of the mosaic variants, than those who were ≥18 years of age (mean AB, 0.22 versus 0.16; $p = 2e{-}17$) (Figure S6A). This held true across all clinical areas (Table 3) and was also observed specifically among genes involved in early-onset disorders (mean AB, 0.23 for individuals <18 year old versus 0.15 for individuals ≥18 year old; $p = 5e{-}12$) (Figure S6B).

In addition to investigating correlations in aggregate between mosaicism and age across all genes and clinical areas, we specifically focused on mosaic variants and age in the context of molecular diagnoses. Individuals of different ages had diagnostic mosaic P/LP variants in one of eleven genes; the level of mosaicism for these variants was slightly higher in individuals <18 years of age than in those ≥18 (mean AB, 0.194 versus 0.175) (Table S9). Furthermore, we compared individuals harboring diagnostic mosaic P/LP variants with individuals harboring diagnostic non-mosaic P/LP variants in the same genes (134 genes in total). For 17 of the 134 genes, individuals with diagnostic mosaic P/LP variants were significantly older than those with diagnostic non-mosaic P/LP variants ($p < 0.05$). For example, in the case of neurofibromatosis, the mean age of those harboring diagnostic P/LP variants in *NF1* was 59 years when the variants were mosaic versus 21 years when they were non-mosaic ($p = 3e{-}92$). Although the same pattern was evident for another 76 genes (e.g., *DMD*, *KCNQ2* [MIM: 602235], and *NIPBL*), the difference in age between the two groups did not reach statistical significance because of small sample sizes (Table S9). For the remaining 41 genes, either there was no difference in mean age between the two groups or the mosaic individuals were actually younger. When all 134 genes were considered together, individuals with the diagnostic mosaic variants in these genes were significantly older overall than those with the diagnostic non-mosaic variants ($p = 8.3e{-}96$).

Discussion

Data from high-depth NGS in a very large clinical cohort afforded us a unique opportunity to address key questions about the types and frequencies of mosaic variants that occur in hereditary disease, the distribution of mosaic variants among genes and their cognate disorders, and the correlations among mosaic variation, clinical phenotypes, and age of affected individuals. Across 509 genes, we not only identified many mosaic variants in genes previously reported to have a high prevalence of mosaicism (e.g., *TP53*, *CHEK2*, *NF1*, *CDKL5*) but also discovered many mosaic sequence variants and intragenic CNVs not previously reported in ClinVar.

Mosaic variants in this study appeared to explain disease etiology in roughly 2% of individuals referred for genetic testing for hereditary diseases, including cancer syndromes. In a previous study by Cao et al., clinical exome sequencing in 12,000 samples from individuals with unselected clinical phenotypes demonstrated that roughly 1.5% of molecular diagnoses were due to a mosaic variant.[24] Because we analyzed a limited number of disease-related genes and Cao et al.'s study had reduced sensitivity for mosaicism resulting from lower coverage sequencing, both observations likely underestimate the overall prevalence of mosaicism in a single genome, but for different reasons. Future high-depth exome or whole

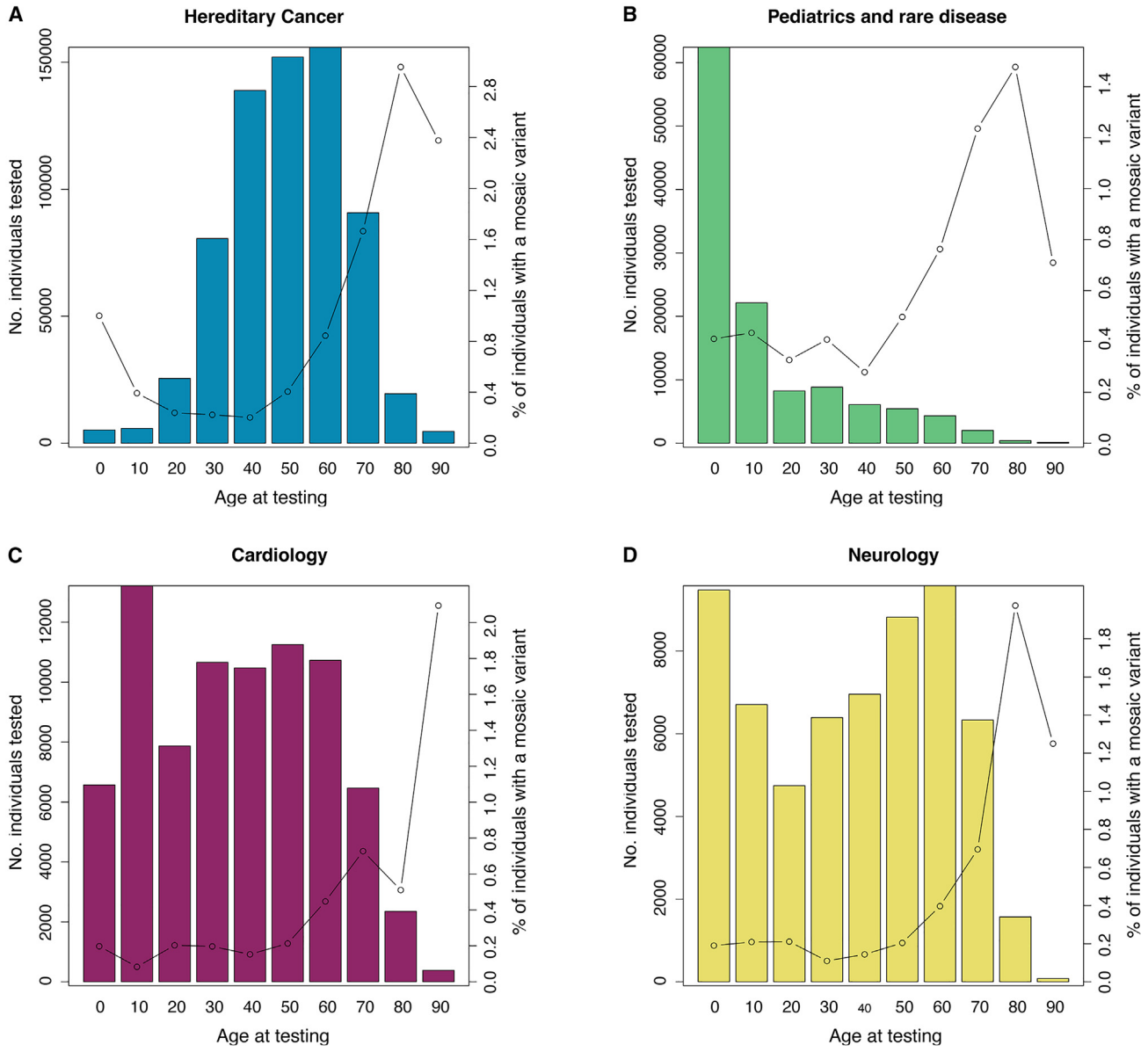


Figure 5. Relationships between age and frequency of mosaic variants observed in clinical genetic testing
Trend lines show the proportion of individuals who harbored a mosaic variant among all who were referred for genetic testing with gene panels associated with (A) hereditary cancer, (B) pediatrics and rare disease, (C) cardiology, or (D) neurology. Columns show the number of tests within each clinical area by age group. Peds, pediatrics.

genome sequencing studies in larger cohorts will likely correct these underestimates.

Mosaicism related to hereditary cancer and clonal hematopoiesis

Although it is well established that certain types of DNA variants can arise in tumors and promote their growth, constitutional mosaic variants acquired in cancer-related genes have received less attention. The majority of mosaic P/LP variants in our study were in cancer-related genes (e.g., *TP53*, *ATM*, and *CHEK2*), even after normalizing for the number of individuals tested across clinical areas. Similar to other reports,[35,36] half of the P/LP variants in *TP53* were mosaic in our cohort. Several groups have discussed the biological and clinical significance of mosaic variants in this gene, including the importance of follow-up testing in a second sample type or in relevant family members to guide proper clinical interpretation,[17,36] and the potential for clonal populations harboring *TP53* P/LP variants to expand preferentially following cancer treatments.[9,37,38] Given the direct impact on treatment strategies for some individuals, it is important to determine whether a P/LP variant in *TP53* is non-mosaic, constitutionally mosaic, or the result of CH.

Many mosaic variants detected in older individuals with cancer are somatic variants arising from CH or age-related changes. These can be found in up to 10% of people by age 65 and up to 30% of people by age 80.[15,39,40] Small

Table 2. Mean age at testing among individuals with a pathogenic or likely pathogenic result, shown by clinical area

Clinical area of test order	Age at testing	Mean age of individuals with non-mosaic P/LP variants	Mean age of individuals with mosaic P/LP variants	p value for difference in mean ages
Hereditary cancer	any	53	63	4.79e−120
Hereditary cancer	<50 years	37	32	2.99e−10
Pediatrics and rare disease	any	19	15	0.0004
Pediatrics and rare disease[a]	<50 years	13	11	0.003
Cardiology	any	40	43	0.416
Cardiology[a]	<50 years	26	26	0.812
Neurology	any	32	32	0.983
Neurology[a]	<50 years	21	17	0.194

[a]Individuals with genetic findings in a gene associated with hereditary cancer are excluded.

sequence variants that confer a cellular growth advantage are predominant contributors to CH, and chromosomal abnormalities and uniparental disomy are also observed as rare mutational mechanisms in CH.[3] Although the involvement of intragenic deletions and duplications in CH has been unclear, our study preliminarily suggests that intragenic CNVs are not major contributors to CH, based on the observation that their prevalence was not higher in older individuals in whom CH is expected to be more common. Further research is needed to explore the extent to which CNVs may contribute to CH.

Mosaicism in pediatric, cardiovascular, and neurological disorders

Mosaic variants in *DMD*, *CDKL5*, and other epilepsy-related genes associated with pediatric neurological disorders have been well documented.[41] This study has expanded the number of genes in which mosaic variants are observed in such early-onset disorders. Individuals with mosaic variants associated with pediatric and rare diseases appeared in several instances to have non-classic phenotypic features, possibly due to milder phenotypes resulting from the mosaicism. For example, a mosaic truncating pathogenic variant was observed in *DMD* in a 64-year-old man with impaired cardiac function and non-classic Duchenne muscular dystrophy, while severely disruptive non-mosaic loss-of-function variants in this gene lead to the classic phenotype within the first decade of life. Likewise, a mosaic splice-disrupting pathogenic variant in *RB1* was found in a 42-year-old individual with bilateral retinoblastoma, whereas most non-mosaic individuals with *RB1*-related retinoblastoma are diagnosed between early infancy and six years of age.

Complexities in interpreting the clinical significance of mosaicism

This study highlights several important challenges in conveying the clinical significance of mosaic variants. First, a parent who carries a mosaic pathogenic variant should be made aware of the risk that it can be transmitted through the germline to offspring. Among 376 cases in which a proband with a mosaic P/LP variant also had offspring tested, 14 of 563 tested offspring harbored a non-mosaic heterozygous version of the variant, including two cases of sibling pairs who had inherited the same variant from a mosaic parent. Given the limited number of cases of transmission and the fact that the transmitting parents showed a wide range of ABs (0.10–0.31), we were not able to make comparisons to address questions related to level of mosaicism and transmission. Future studies on this topic are warranted. Similarly, mosaic variants detected through carrier screening can complicate reproductive decision-making because of uncertainty around whether the variants are also present in the germline. Second, for mosaic variants in genes associated with disorders that show reduced penetrance, correlations with disease and clinical prognosis are not straightforward. We observed 2,814 mosaic P/LP variants or mosaic VUSs in genes associated with disorders that have reduced penetrance. Third, mosaic variants in X-linked genes found in clinically affected women may be difficult to correlate with disease because of the influence of X-inactivation. We identified 81 mosaic P/LP variants or mosaic VUSs in such genes in this study cohort. A fourth challenge is that when both a mosaic P/LP variant and a non-mosaic P/LP variant are discovered in the same gene associated with an autosomal-dominant or X-linked disease, it is difficult to clearly discern which variant is contributing to disease. This, in turn, has implications for cascade testing of family members and for reproductive planning. Separately, when multiple mosaic variants are detected in the same gene, especially in older individuals, the variants could be the result of CH and have little or no effect on health. In our study, when two or more mosaic variants were observed together, e.g., in either *TP53* or *ATM,* the ABs for the variants were frequently similar among older individuals, supporting the notion that co-origination of the variants allowed age-related clonal expansion. Finally,

Table 3. Mean allele balance of mosaic P/LP variants

Gene clinical area	<18 years		18–50 years		>50 years	
	Allele balance of P/LP variants, mean (range)	Allele balance of P/LP variants and VUSs, mean (range)	Allele balance of P/LP variants, mean (range)	Allele balance of P/LP variants and VUSs, mean (range)	Allele balance of P/LP variants, mean (range)	Allele balance of P/LP variants and VUSs, mean (range)
Hereditary cancer	0.19 (0.08–0.38)	0.18 (0.08–0.38)	0.18 (0.06–0.39)	0.17 (0.06–0.39)	0.16 (0.06–0.40)	0.16 (0.06–0.40)
Pediatrics and rare disease	0.19 (0.07–0.39)	0.19 (0.07–0.39)	0.16 (0.08–0.32)	0.15 (0.08–0.32)	0.15 (0.07–0.36)	0.15 (0.07–0.38)
Cardiology	0.19 (0.14–0.19)	0.15 (0.08–0.31)	0.18 (0.08–0.28)	0.18 (0.07–0.32)	0.16 (0.10–0.33)	0.15 (0.07–0.33)
Neurology	0.20 (0.10–0.32)	0.19 (0.09–0.33)	0.14 (0.08–0.28)	0.16 (0.07–0.32)	0.17 (0.09–0.38)	0.15 (0.08–0.38)
All genes	0.19 (0.07–0.39)	0.19 (0.07–0.39)	0.18 (0.06–0.39)	0.17 (0.06–0.39)	0.16 (0.06–0.40)	0.16 (0.15–0.40)

Mosaic individuals with variants with allele balances (ABs) < 0.06 were excluded due to limited reliability and those with ABs >0.4 were excluded because 0.4 is the highest allowable AB for a standard non-mosaic heterozygous single nucleotide variant in an autosome. All variants with higher ABs are X-linked in male individuals or have other exceptional circumstances (e.g., overlapping with a large deletion).

albeit extremely rare, a mosaic variant can arise from real biological phenomena that are related to DNA repair or chromosomal behavior, such as revertant mosaicism or mosaic somatic uniparental disomy. The presence of an apparently revertant mosaic variant in *FANCA* in our study supports previous observations of this phenomenon in this gene.[42] We also identified a case of mosaic uniparental disomy whereby the sequence variant did not originate from DNA synthesis or repair, but rather from abnormal chromosomal recombination. All of these scenarios pose significant challenges to genetic counseling and prognostication, requiring investigation of additional samples from the probands and testing of family members.

Effect of mosaicism on clinical phenotype and age of onset

In a hereditary disease setting, true mosaic P/LP variants appear to be present predominantly in younger individuals in whom CH contribution is very unlikely. Mosaic P/LP variants are also present at a higher AB in younger individuals than in older individuals, suggesting that once a certain threshold of variant burden is reached, disease manifestation and clinical recognition is unavoidable. We particularly observed higher levels of mosaicism in early-onset disorders. However, this probably reflected a bias of ascertainment since the young individuals who unknowingly harbored mosaic P/LP variants were sufficiently affected with disease to be seen by clinicians, who then ordered genetic testing and discovered the variants. Individuals with mosaic P/LP variants who have mild phenotypes or are unaffected would be a useful comparator group to test these assumptions.

Since generalizing our analysis of mosaic variants across all genes and diseases may have obscured important correlations, we specifically compared individuals of different ages who had mosaic P/LP variants that contributed to molecular diagnoses within the same genes. When compared with non-mosaic P/LP variants, the presence of mosaic P/LP variants would be expected to reduce phenotypic severity, through either a milder phenotype or a later age of onset, because of the limited number of affected tissues or cells.[1,3] This was corroborated by our analysis of individuals with diagnostic mosaic variants in the same genes, showing an inverse correlation between level of mosaicism and age of the affected individual. When comparing individuals with mosaic versus non-mosaic diagnostic variants in the same genes, we noted that those with mosaic variants were older. In these cases, the mosaic individuals may have had a milder phenotype or a later onset of recognizable disease, leading to genetic testing at a later age. Corroborating this observation, our exploratory analysis of genes associated with rare diseases with distinctive clinical features suggested that mosaic variants were indeed associated with milder or atypical phenotypes. However, correlations between mosaicism and phenotypic severity are not uniformly predictable for each individual and, depending on the distribution of mosaicism within the individual, the phenotypic severity can range from non-existent to mild to classic disease presentation.

Technical considerations for detecting mosaicism

Ambiguity in detecting mosaic variants can arise from both biological and technical factors. For instance, it was not possible in some rare cases for us to determine whether a variant spuriously appeared mosaic due to (1) a co-existing chromosomal or subchromosomal aneuploidy (e.g., the *WDR45* variant), (2) a mosaic focal CNV at the same location, (3) poor quality of sequencing from a suboptimal blood or DNA specimen, (4) ambiguity in the position of the observed variant, or (5) an inability to discern whether a variant is present in a pseudogene sequence (e.g., in *KANSL1*). For example, at least some of the multiple mosaic variants observed in 12 individuals with hematologic malignancies were likely spurious, possibly due to the disrupting effects of cancer treatments on DNA obtained from blood specimens, to residual disease, or to the fact that these malignancies can be associated with mosaic chromosomal rearrangements that compromise

detection of certain sequence variants. Similarly, in another 11 individuals with multiple mosaic variants, NGS-based CNV analysis could not be completed, suggesting again that these observations were spurious due to suboptimal DNA and sequencing quality. Lastly, in two individuals, a sequence variant was observed at the same location as a mosaic intragenic CNV, making the sequence variant appear falsely as mosaic in the NGS data. These cases point to instances that may not always be easy to resolve with NGS alone in any clinical laboratory.

Notwithstanding the aforementioned challenges, it is worth re-emphasizing that some genes appear to have a higher propensity than others for harboring disease-causing mosaic variants. Therefore, the distribution of ABs for the mosaic variants observed in our study reflects both the clinical prevalence of mosaic variants at different ABs and the sensitivity and specificity of the NGS and bioinformatics methods used. For these types of genes, it would be useful to optimize the NGS chemistry and include a specific algorithm in the bioinformatics pipeline to identify mosaic variants with greater sensitivity.

Our validation experiments (supplemental methods) revealed a broad range of ABs and variant types that can be reliably attributed to mosaic variation when deep-coverage sequencing methods are used. These methods are most effective for small mosaic variants such as SNVs. Other types of variants, including large indels, those within repetitive sequences, those in genes with high-similarity copies, and intragenic CNVs (especially those involving small segments of DNA), can be more difficult to detect. Although our methods could confidently call mosaic CNVs with ABs of 0.15–0.30, additional work is needed to refine methods to increase the range of mosaic CNV identification. Our validation experiments also allowed us to develop an algorithm that identifies variant-specific and sequencing-performance-specific parameters for calling mosaic variants that can be applied to new genes added to the NGS assay. Exome sequencing or other methods that typically use lower depth of sequencing than targeted gene panel sequencing can still uncover mosaic variants but are less sensitive. Further investigation of mosaicism could benefit from specially designed assays, such as those based on anchored multiplex PCR (AMP) chemistry,[43] which can reliably detect very low-level mosaicism. Explicit practice guidelines would be useful for standardizing how laboratories identify and report mosaicism.

Limitations

Our findings probably represent an underestimate of mosaic variation in the cohort sampled, because we examined only those genes covered by our targeted NGS panels. Even though we used a higher sequencing depth than is typical for hereditary disease testing with gene panels, a sequencing assay designed for mosaicism would likely uncover more mosaic variants. The retrospective analysis also limited our examination of constitutional mosaicism because multiple samples from different tissue sources were difficult to obtain from each individual with a mosaic variant. A key limitation to standard genetic testing for hereditary disease is that the methods typically utilize blood- or saliva-derived DNA and therefore can miss mosaic variants present in a hard-to-access affected tissue (e.g., brain or heart). Another limitation in this study was the inability to routinely distinguish constitutional mosaic variants that represent actual molecular diagnoses for hereditary cancer syndromes from somatic variants related to malignancies or CH in older individuals. Our observation that approximately half of the individuals with multiple mosaic variants had two or three of them within a single gene, often within *TP53* or *ATM*, may be because some individuals who undergo germline testing for hereditary cancer conditions have been treated with chemotherapies, which by themselves increase the risk of myeloid neoplasms. As a result, some of these individuals may have an evolving and undiagnosed myelodysplastic syndrome at the time of testing. Finally, age in our study was limited to age at testing and not age at disease onset or diagnosis in the clinic, which may have obfuscated the true effects of mosaic variants on the natural history of disease in some individuals. Finally, at least with respect to hereditary cancer genetic testing, healthcare providers referred affected individuals but also those who were unaffected but at risk for a familial variant identified in a proband relative or had strong family history of cancer without a known genetic etiology; it is possible that the relative proportions of these individuals may have affected the rate of mosaic variants detected in cancer-related genes.

Conclusion

Understanding mosaic variation in the human genome has been a long-standing and challenging effort, and the versatility of NGS is enabling deeper research of this phenomenon. The results of this exhaustive analysis of disease-related genes in a very large clinical cohort expands our knowledge of the overall incidence of mosaic variation in hereditary disease; the correlations between mosaic variation and age of clinically affected individuals, severity of clinical phenotypes, and onset of disease; and the complicated implications of mosaicism in carrier screening or in X-linked genes. Observations from this study can support researchers working to unravel the mechanisms through which mosaic variants affect the natural history of hereditary diseases, such as neurofibromatosis or primary immune deficiencies. These results can help clinicians better diagnose hereditary diseases, provide prognoses, and manage the care of individuals with mosaicism, including by recognizing opportunities for additional testing in affected individuals and their family members.

Data and code availability

Variants were made publicly available in ClinVar when the data-sharing preferences of the sequenced individual allowed: https://www.ncbi.nlm.nih.gov/clinvar/submitters/500031/. The algorithms for mosaicism detection described in the supplemental

methods are not publicly available because they are part of a bioinformatics pipeline specifically coupled to the customized next-generation sequencing method and process at Invitae.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.02.013.

Acknowledgments

This work was a retrospective study of available data from clinical genetic testing and was not specifically funded. We thank Kerry Aradhya of Invitae for scientific editing and Dr. Sarah Poll for assistance with data analysis. S.C.C. acknowledges support from the Intramural Research Program of the National Human Genome Research Institute at the National Institutes of Health.

Declaration of interests

R.T., S.R., K.O., C.K., M.K., D.P.-A., B.J., A.S., L.B.-S., R.L.N., and S.A. are current or former employees and shareholders of Invitae Corporation.

Received: September 22, 2022
Accepted: February 23, 2023
Published: March 17, 2023

Web resources

ClinVar, https://www.ncbi.nlm.nih.gov/clinvar/
Online Mendelian Inheritance in Man, http://www.omim.org

References

1. Campbell, I.M., Shaw, C.A., Stankiewicz, P., and Lupski, J.R. (2015). Somatic mosaicism: implications for disease and transmission genetics. Trends Genet. *31*, 382–392.
2. De, S. (2011). Somatic mosaicism in healthy human tissues. Trends Genet. *27*, 217–223.
3. Forsberg, L.A., Gisselsson, D., and Dumanski, J.P. (2017). Mosaicism in health and disease - clones picking up speed. Nat. Rev. Genet. *18*, 128–142.
4. Siri, B., Varesio, C., Freri, E., Darra, F., Gana, S., Mei, D., Porta, F., Fontana, E., Galati, G., Solazzi, R., et al. (2021). CDKL5 deficiency disorder in males: five new variants and review of the literature. Eur. J. Paediatr. Neurol. *33*, 9–20.
5. Rydzanicz, M., Glinkowski, W., Walczak, A., Koppolu, A., Kostrzewa, G., Gasperowicz, P., Pollak, A., Stawiński, P., and Płoski, R. (2022). Postzygotic mosaicism of a novel PTPN11 mutation in monozygotic twins discordant for metachondromatosis. Am. J. Med. Genet. *188*, 1482–1487.
6. Myers, C.T., Hollingsworth, G., Muir, A.M., Schneider, A.L., Thuesmunn, Z., Knupp, A., King, C., Lacroix, A., Mehaffey, M.G., Berkovic, S.F., et al. (2018). Parental mosaicism in "De Novo" epileptic encephalopathies. N. Engl. J. Med. *378*, 1646–1648.
7. Pareja, F., Ptashkin, R.N., Brown, D.N., Derakhshan, F., Selenica, P., da Silva, E.M., Gazzo, A.M., Da Cruz Paula, A., Breen, K., Shen, R., et al. (2022). Cancer-causative mutations occurring in early embryogenesis. Cancer Discov. *12*, 949–957.
8. Steinke-Lange, V., de Putter, R., Holinski-Feder, E., and Claes, K.B. (2021). Somatic mosaics in hereditary tumor predisposition syndromes. Eur. J. Med. Genet. *64*, 104360.
9. Slavin, T.P., Coffee, B., Bernhisel, R., Logan, J., Cox, H.C., Marcucci, G., Weitzel, J., Neuhausen, S.L., and Mancini-DiNardo, D. (2019). Prevalence and characteristics of likely-somatic variants in cancer susceptibility genes among individuals who had hereditary pan-cancer panel testing. Cancer Genet. *235-236*, 31–38.
10. Mirzaa, G., Graham, J.M., Jr., and Keppler-Noreuil, K. (2013). PIK3CA-Related Overgrowth Spectrum. In GeneReviews®, M.P. Adam, D.B. Everman, G.M. Mirzaa, R.A. Pagon, S.E. Wallace, L.J.H. Bean, K.W. Gripp, and A. Amemiya, eds. (University of Washington, Seattle).
11. Brioude, F., Toutain, A., Giabicani, E., Cottereau, E., Cormier-Daire, V., and Netchine, I. (2019). Overgrowth syndromes - clinical and molecular aspects and tumour risk. Nat. Rev. Endocrinol. *15*, 299–311.
12. Renaux-Petel, M., Charbonnier, F., Théry, J.C., Fermey, P., Lienard, G., Bou, J., Coutant, S., Vezain, M., Kasper, E., Fourneaux, S., et al. (2018). Contribution of de novo and mosaic TP53 mutations to Li-Fraumeni syndrome. J. Med. Genet. *55*, 173–180.
13. Weitzel, J.N., Chao, E.C., Nehoray, B., Van Tongeren, L.R., LaDuca, H., Blazer, K.R., Slavin, T., Facmg, D.A.B.M.D., Pesaran, T., Rybak, C., et al. (2018). Somatic TP53 variants frequently confound germ-line testing results. Genet. Med. *20*, 809–816.
14. Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. N. Engl. J. Med. *371*, 2488–2498.
15. Shlush, L.I. (2018). Age-related clonal hematopoiesis. Blood *131*, 496–504.
16. Batalini, F., Peacock, E.G., Stobie, L., Robertson, A., Garber, J., Weitzel, J.N., and Tung, N.M. (2019). Li-Fraumeni syndrome: not a straightforward diagnosis anymore-the interpretation of pathogenic variants of low allele frequency and the differences between germline PVs, mosaicism, and clonal hematopoiesis. Breast Cancer Res. *21*, 107.
17. Coffee, B., Cox, H.C., Bernhisel, R., Manley, S., Bowles, K., Roa, B.B., and Mancini-DiNardo, D. (2020). A substantial proportion of apparently heterozygous TP53 pathogenic variants detected with a next-generation sequencing hereditary pan-cancer panel are acquired somatically. Hum. Mutat. *41*, 203–211.
18. Rohlin, A., Wernersson, J., Engwall, Y., Wiklund, L., Björk, J., and Nordling, M. (2009). Parallel sequencing used in detection of mosaic mutations: comparison with four diagnostic DNA screening techniques. Hum. Mutat. *30*, 1012–1020.
19. Tsiatis, A.C., Norris-Kirby, A., Rich, R.G., Hafez, M.J., Gocke, C.D., Eshleman, J.R., and Murphy, K.M. (2010). Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications. J. Mol. Diagn. *12*, 425–432.
20. Campbell, I.M., Yuan, B., Robberecht, C., Pfundt, R., Szafranski, P., McEntagart, M.E., Nagamani, S.C.S., Erez, A., Bartnik, M., Wiśniowiecka-Kowalnik, B., et al. (2014). Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. Am. J. Hum. Genet. *95*, 173–182.
21. Acuna-Hidalgo, R., Bo, T., Kwint, M.P., van de Vorst, M., Pinelli, M., Veltman, J.A., Hoischen, A., Vissers, L.E.L.M., and

Gilissen, C. (2015). Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. Am. J. Hum. Genet. *97*, 67–74.
22. Qin, L., Wang, J., Tian, X., Yu, H., Truong, C., Mitchell, J.J., Wierenga, K.J., Craigen, W.J., Zhang, V.W., and Wong, L.-J.C. (2016). Detection and Quantification of Mosaic Mutations in Disease Genes by Next-Generation Sequencing. J. Mol. Diagn. *18*, 446–453.
23. Brewer, C.J., Gillespie, M., Fierro, J., Scaringe, W.A., Li, J.M., Lee, C.-Y., Yen, H.-Y., Gao, H., and Strom, S.P. (2020). The Value of Parental Testing by Next-Generation Sequencing Includes the Detection of Germline Mosaicism. J. Mol. Diagn. *22*, 670–678.
24. Cao, Y., Tokita, M.J., Chen, E.S., Ghosh, R., Chen, T., Feng, Y., Gorman, E., Gibellini, F., Ward, P.A., Braxton, A., et al. (2019). A clinical survey of mosaic single nucleotide variants in disease-causing genes detected by exome sequencing. Genome Med. *11*, 48.
25. Kurian, A.W., Hare, E.E., Mills, M.A., Kingham, K.E., McPherson, L., Whittemore, A.S., McGuire, V., Ladabaum, U., Kobayashi, Y., Lincoln, S.E., et al. (2014). Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. J. Clin. Oncol. *32*, 2001–2009.
26. Truty, R., Paul, J., Kennemer, M., Lincoln, S.E., Olivares, E., Nussbaum, R.L., and Aradhya, S. (2019). Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. Genet. Med. *21*, 114–123.
27. Lincoln, S.E., Truty, R., Lin, C.-F., Zook, J.M., Paul, J., Ramey, V.H., Salit, M., Rehm, H.L., Nussbaum, R.L., and Lebo, M.S. (2019). A Rigorous Interlaboratory Examination of the Need to Confirm Next-Generation Sequencing-Detected Variants with an Orthogonal Method in Clinical Genetic Testing. J. Mol. Diagn. *21*, 318–329.
28. Lincoln, S.E., Hambuch, T., Zook, J.M., Bristow, S.L., Hatchell, K., Truty, R., Kennemer, M., Shirts, B.H., Fellowes, A., Chowdhury, S., et al. (2021). One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. Genet. Med. *23*, 1673–1680.
29. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.
30. Nykamp, K., Anderson, M., Powers, M., Garcia, J., Herrera, B., Ho, Y.-Y., Kobayashi, Y., Patil, N., Thusberg, J., Westbrook, M., et al. (2017). Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. Genet. Med. *19*, 1105–1117.
31. Johnson, B., Ouyang, K., Frank, L., Truty, R., Rojahn, S., Morales, A., Aradhya, S., and Nykamp, K. (2022). Systematic use of phenotype evidence in clinical genetic testing reduces the frequency of variants of uncertain significance. Am. J. Med. Genet. *188*, 2642–2651.
32. ClinGen Low Penetrance/Risk Allele Working Group (2020). Recommended terminology when describing variants with decreased penetrance for Mendelian conditions. https://www.clinicalgenome.org/site/assets/files/4531/clingenrisk_terminology_recomendations-final-02_18_20.pdf.
33. Mundschau, G., Gurbuxani, S., Gamis, A.S., Greene, M.E., Arceci, R.J., and Crispino, J.D. (2003). Mutagenesis of GATA1 is an initiating event in Down syndrome leukemogenesis. Blood *101*, 4298–4300.
34. Webster, R., Cho, M.T., Retterer, K., Millan, F., Nowak, C., Douglas, J., Ahmad, A., Raymond, G.V., Johnson, M.R., Pujol, A., et al. (2017). De novo loss of function mutations in KIAA2022 are associated with epilepsy and neurodevelopmental delay in females. Clin. Genet. *91*, 756–763.
35. Coffee, B., Cox, H.C., Kidd, J., Sizemore, S., Brown, K., Manley, S., and Mancini-DiNardo, D. (2017). Detection of somatic variants in peripheral blood lymphocytes using a next generation sequencing multigene pan cancer panel. Cancer Genet. *211*, 5–8.
36. Mester, J.L., Jackson, S.A., Postula, K., Stettner, A., Solomon, S., Bissonnette, J., Murphy, P.D., Klein, R.T., and Hruska, K.S. (2020). Apparently Heterozygous TP53 Pathogenic Variants May Be Blood Limited in Patients Undergoing Hereditary Cancer Panel Testing. J. Mol. Diagn. *22*, 396–404.
37. Wong, T.N., Ramsingh, G., Young, A.L., Miller, C.A., Touma, W., Welch, J.S., Lamprecht, T.L., Shen, D., Hundal, J., Fulton, R.S., et al. (2015). Role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. Nature *518*, 552–555.
38. Ruijs, M.W.G., Verhoef, S., Rookus, M.A., Pruntel, R., van der Hout, A.H., Hogervorst, F.B.L., Kluijt, I., Sijmons, R.H., Aalfs, C.M., Wagner, A., et al. (2010). TP53 germline mutation testing in 180 families suspected of Li-Fraumeni syndrome: mutation detection rate and relative frequency of cancers in different familial phenotypes. J. Med. Genet. *47*, 421–428.
39. Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. *371*, 2477–2487.
40. Arends, C.M., Galan-Sousa, J., Hoyer, K., Chan, W., Jäger, M., Yoshida, K., Seemann, R., Noerenberg, D., Waldhueter, N., Fleischer-Notter, H., et al. (2018). Hematopoietic lineage distribution and evolutionary dynamics of clonal hematopoiesis. Leukemia *32*, 1908–1919.
41. Stosser, M.B., Lindy, A.S., Butler, E., Retterer, K., Piccirillo-Stosser, C.M., Richard, G., and McKnight, D.A. (2018). High frequency of mosaic pathogenic variants in genes causing epilepsy-related neurodevelopmental disorders. Genet. Med. *20*, 403–410.
42. Nicoletti, E., Rao, G., Bueren, J.A., Río, P., Navarro, S., Surrallés, J., Choi, G., and Schwartz, J.D. (2020). Mosaicism in Fanconi anemia: concise review and evaluation of published cases with focus on clinical course of blood count normalization. Ann. Hematol. *99*, 913–924.
43. Zheng, Z., Liebers, M., Zhelyazkova, B., Cao, Y., Panditi, D., Lynch, K.D., Chen, J., Robinson, H.E., Shim, H.S., Chmielecki, J., et al. (2014). Anchored multiplex PCR for targeted next-generation sequencing. Nat. Med. *20*, 1479–1484.

The American Journal of Human Genetics, Volume *110*

Supplemental information

Patterns of mosaicism for sequence and copy-number variants discovered through clinical deep sequencing of disease-related genes in one million individuals

Rebecca Truty, Susan Rojahn, Karen Ouyang, Curtis Kautzer, Michael Kennemer, Daniel Pineda-Alvarez, Britt Johnson, Amanda Stafford, Lina Basel-Salmon, Sulagna Saitta, Anne Slavotinek, Settara C. Chandrasekharappa, Carlos Jose Suarez, Leslie Burnett, Robert L. Nussbaum, and Swaroop Aradhya

SUPPLEMENTAL INFORMATION

Supplemental Figures

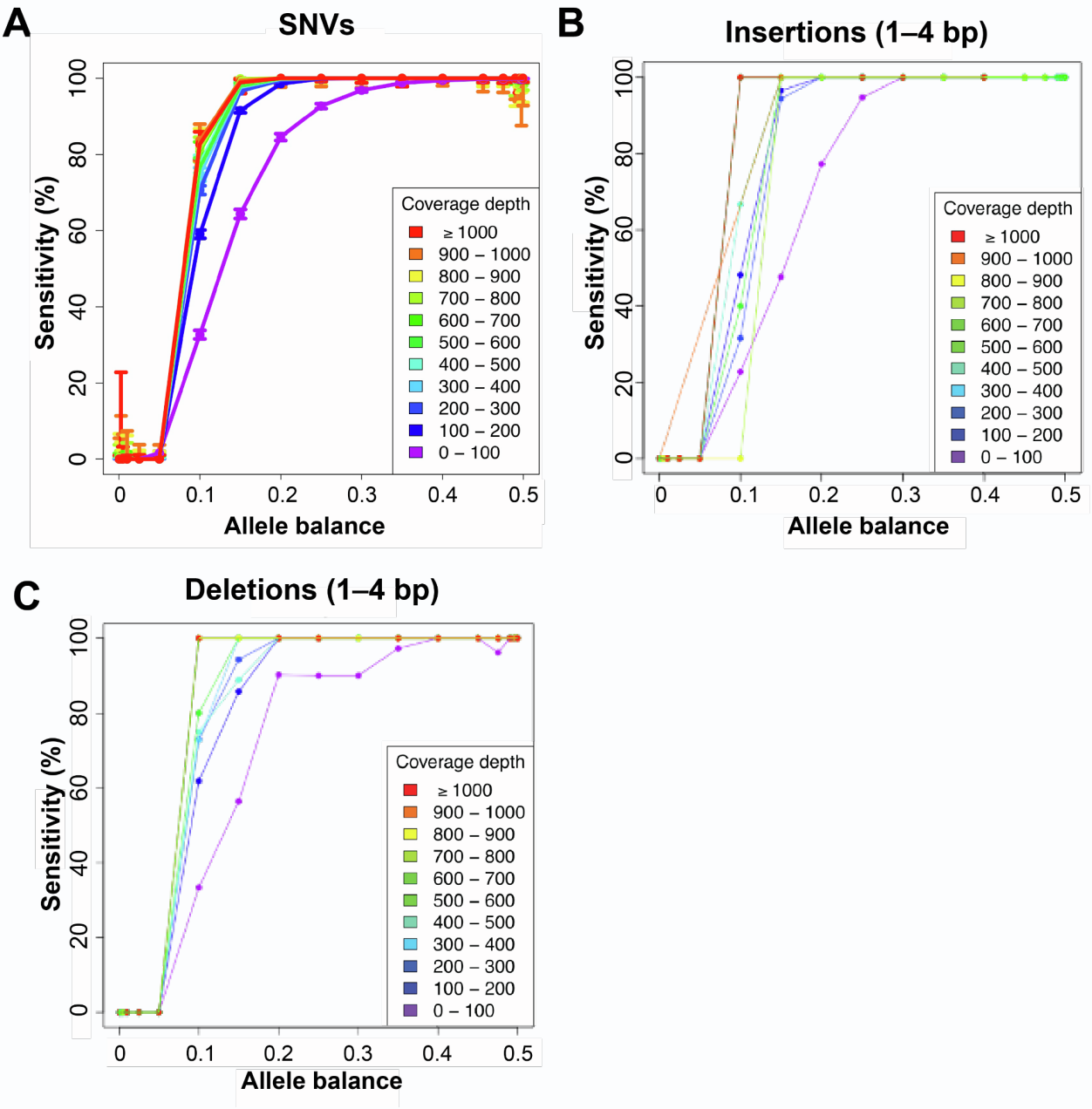
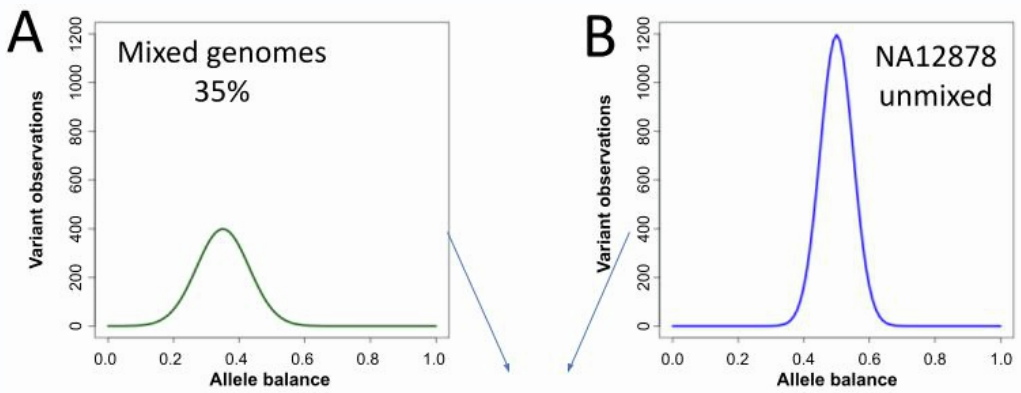
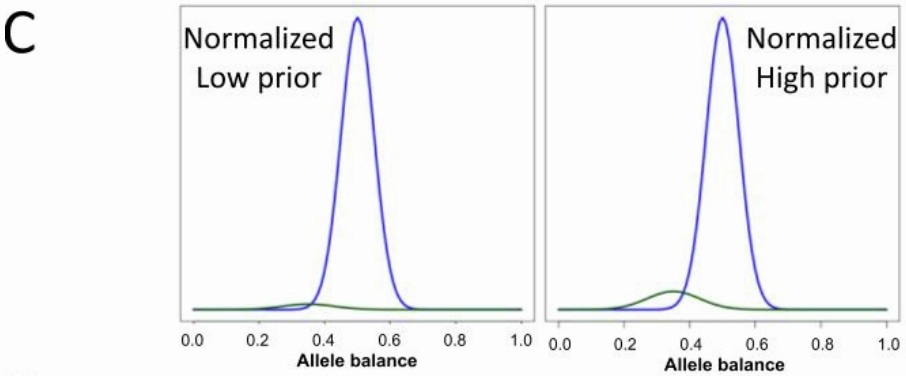
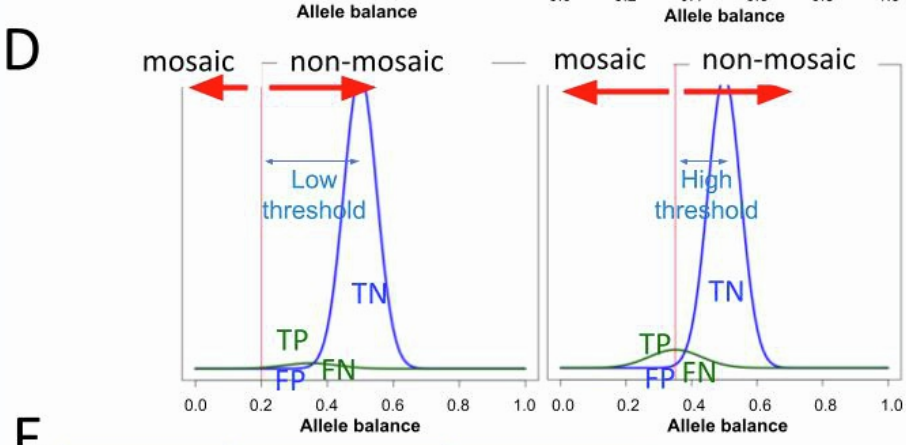


Figure S1. Sensitivity for detecting mosaic sequence variants in uncomplicated genomic regions. Sensitivity as a function of variant depth of sequencing coverage (color) and benchmark allele balance for **(A)** single nucleotide variants (SNVs), **(B)** insertions of 1–4 base pairs, and **(C)** deletions of 1–4 base pairs.







E

	Below threshold = mosaic	Above threshold = non-mosaic
True het	FP	TN
True mosaic	TP	FN

PPV = TP / (TP + FP)
Sensitivity = TP / (TP + FN)

Figure S2. Process to determine the threshold between mosaic and non-mosaic variants. **(A)** Hypothetical example of a single nucleotide variant (SNV) allele balance (AB) distribution in a mixed genome experiment. **(B)** Example of a non-mosaic heterozygous variant in an unmixed genome. **(C)** Genomes normalized to simulate mosaic variants at clinically observed levels for low and high prior genes, where prior refers to the probability that a gene would harbor a mosaic variant. **(D)** An example of the AB distribution of non-mosaic heterozygous SNVs is shown in blue. The vertical lines mark the high threshold and low threshold values computed from the width of this distribution. In green is the AB distribution for similar SNVs with a target AB of 0.35. Any variant below the threshold would be classified as mosaic.

A

	AB below threshold	AB above threshold
True non-mosaic variant	False positive	True negative
True mosaic variant	True positive	False negative

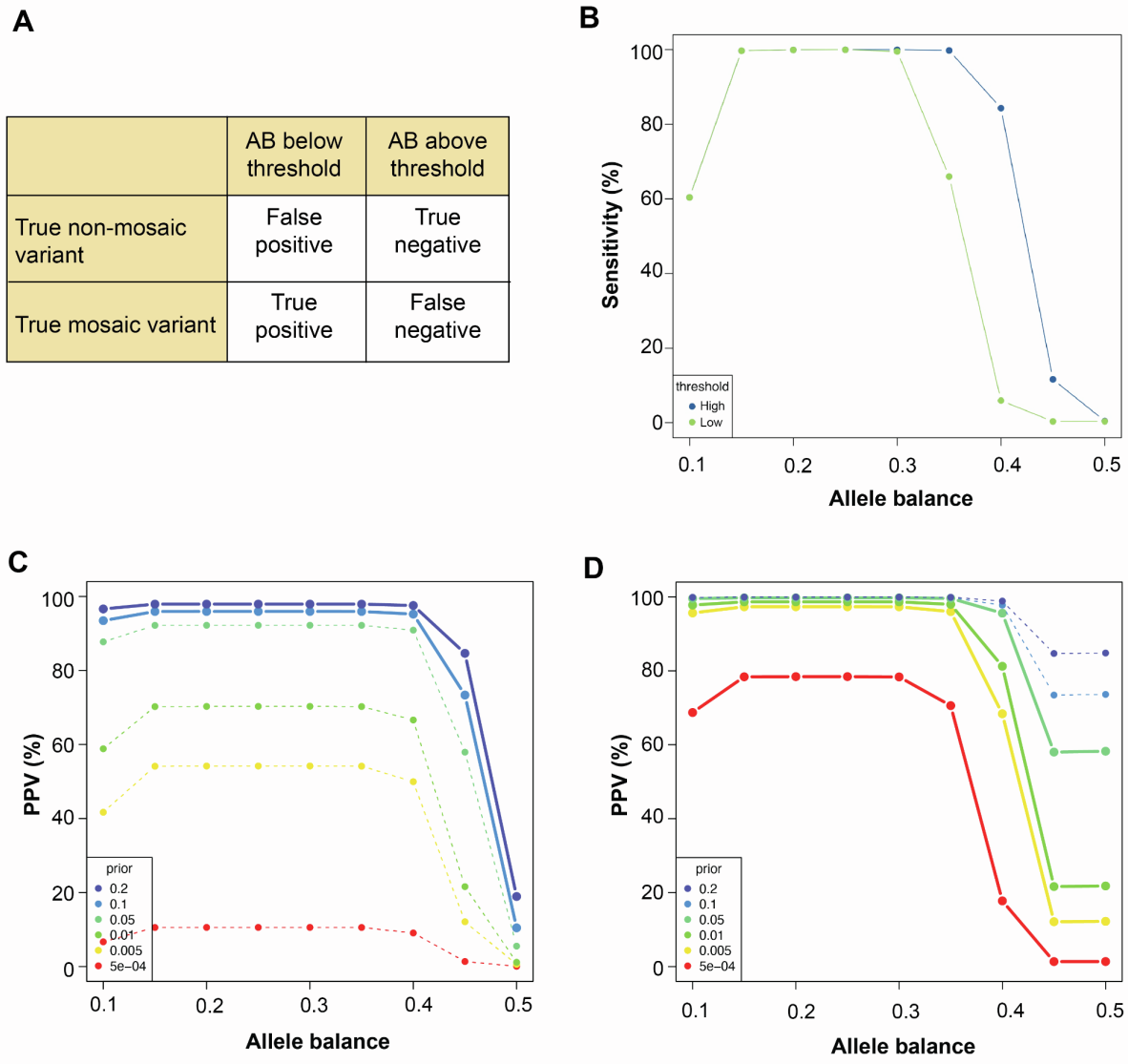


Figure S3. Data for SNVs within an uncomplicated genomic context with at least 600x coverage. (A) Guide to validation experiment outcomes, where the allele balance (AB) of an observed variant determines whether it is classified as mosaic or non-mosaic. (B) Sensitivity as a function of benchmark AB, where sensitivity is calculated as the number of true positives divided by the sum of true positives and false negatives. Blue indicates a high threshold for high-prior genes with known propensity for mosaicism (where ≥10% of clinically significant variants are mosaic). Green indicates a low threshold for other genes (where <10% of clinically significant variants are mosaic). True positives are variants from the mosaic AB distribution that fall below the threshold and false negatives are variants from the mosaic AB distribution that fall above the threshold. (C) Positive predictive value (PPV) as a function of benchmark AB, using a high threshold, for a variety of prior probabilities. PPV is calculated as the number of true positives divided by the sum of true positives and true negatives. (D) PPV as a function of target AB, using a low threshold, for a variety of prior probabilities. Thicker colored lines indicate PPV values anticipated for reported mosaic variants in genes with high prior probability of mosaicism (C) and low prior probability of mosaicism (D).

CNV source sample	CNV type	Affected gene region	Allele balances observed in mosaic CNV benchmark samples						
Clinical sample 1	Whole gene duplication	*NIPA1*	0.09	0.10	0.10	0.13	0.17	0.22	0.29
Clinical sample 2	Partial gene duplication	*AARS* exons 12-21	0.14	0.20	0.20	0.25	0.30	0.34	0.39
Clinical sample 2	Single exon deletion	*CTNNA3* exon 10	0.14	0.20	0.20	0.25	0.30	0.34	0.39
Clinical sample 3	Whole gene deletion	*NPHP1*	0.13	0.19	0.19	0.24	0.29	0.34	0.39
Clinical sample 3	Partial gene deletion	*ALG1* exons 4-13	0.13	0.19	0.19	0.24	0.29	0.34	0.39

Figure S4. Performance of mosaic CNV detection methods in mixed genome samples. Allele balances (ABs) shown were derived from the ABs of single nucleotide variants in the sample. To create the mosaic copy number variant (CNV) benchmark samples, clinical samples known to harbor a CNV were mixed at varying concentrations with a control sample (NA24385, see Table S1). Target ABs of the benchmark mosaic CNVs were 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, and 0.40. Green boxes indicate benchmark samples in which a mosaic CNV could be detected. In these samples, our custom CNV caller flagged the variants as CNV of low quality. In clinical practice, such samples would receive manual review for confirmation of a mosaic CNV. Red and orange boxes indicate benchmark samples in which a mosaic CNV could not be detected. For each red sample, the CNV caller flagged the region as low quality and therefore no CNV was called. For each orange sample, a CNV was called without any quality flag and therefore was not subjected to manual review to confirm that it was mosaic.

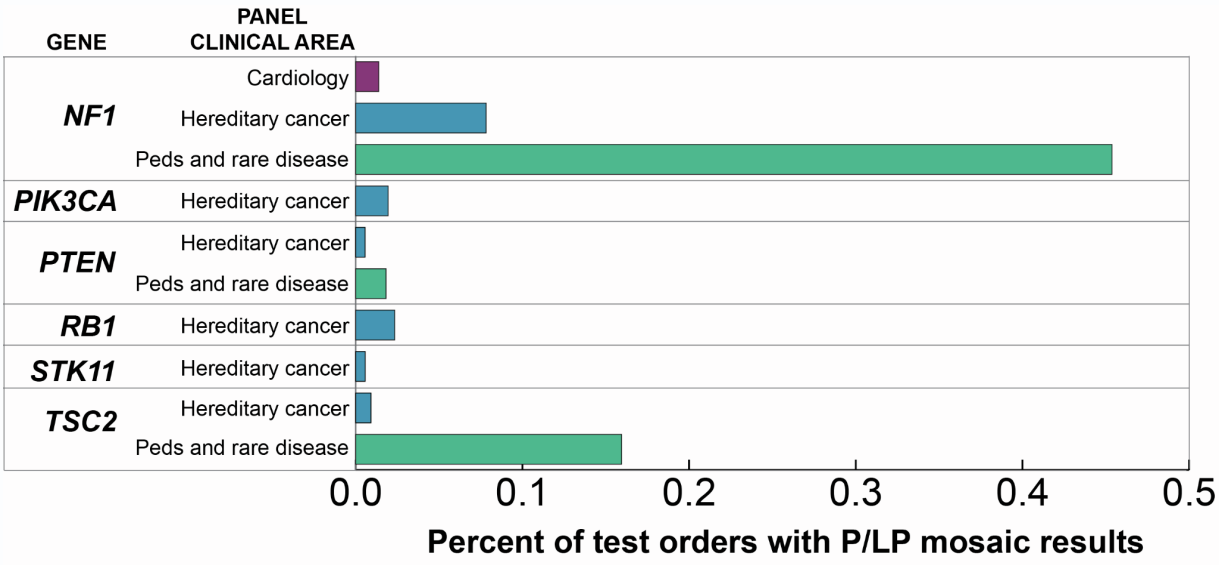


Figure S5. Mosaic variants in genes associated with childhood cancer. Genes associated with childhood cancer were tested in both children and adults, depending on which type of gene panel was requested by the ordering healthcare provider. Peds, Pediatrics.

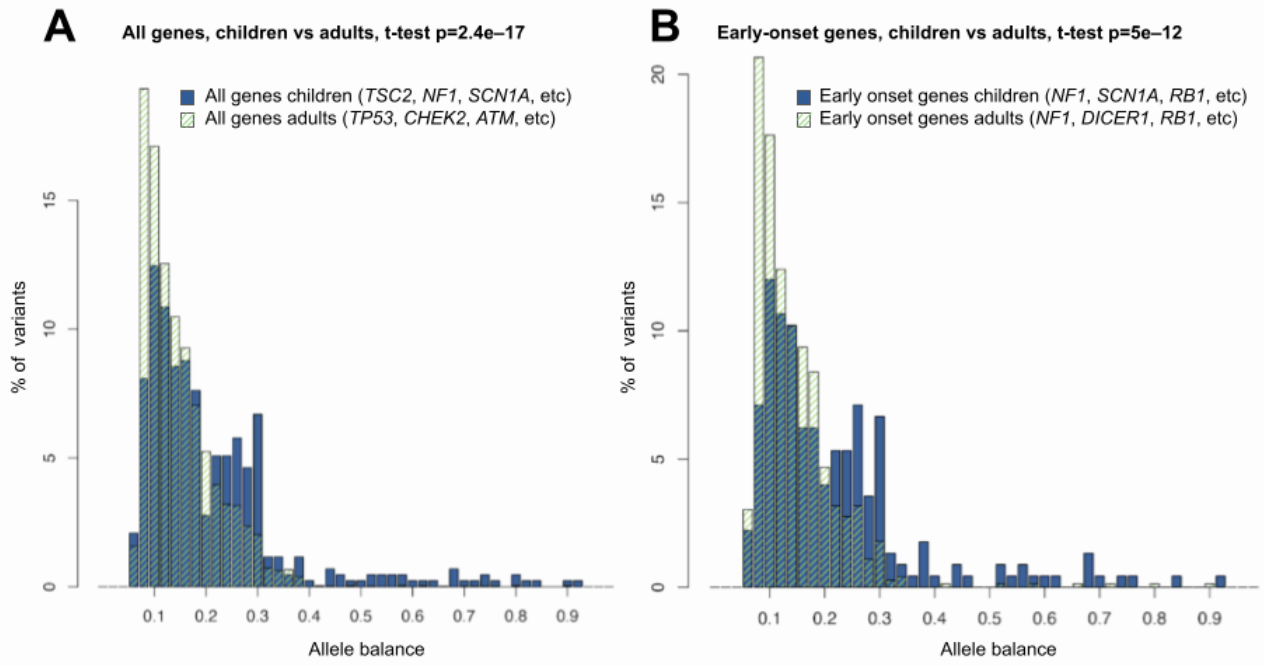


Figure S6. Associations between age at testing and levels of mosaicism. (A) Among all mosaic P/LP variants, those in individuals tested as children (<18 years of age) had higher allele balances than those in individuals tested as adults. (B) Among all mosaic P/LP variants observed in early-onset genes, those in individuals tested as children (<18 years of age) had higher allele balances than those in individuals tested as adults. Both comparisons showed a statistically significant difference in patterns between children and adults, as shown by the t test p value on top of each chart.

Supplemental Tables

Genome A	Genome B	No. of unique variants (genome A / genome B)	Percentages of benchmark sample comprised of genome A across dilution series, %	Allele balances of benchmark mosaic variants across dilution series
NA24385	NA12878	615 / 544	10, 20, 30, 40, 50, 60, 70, 80	0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40
NA24385	NA24631	628 / 593	10, 20, 30, 40, 50, 60, 70, 80	0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40
NA12878	NA24631	628 / 631	10, 20, 30, 40, 50, 60, 70, 80	0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40
Clinical Sample 1	NA24385	Whole gene duplication *NIPA1 (chr15:22839430-23095572)*	20, 30, 40, 50, 60, 70, 80	0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40
Clinical Sample 2	NA24385	Exons 12-21 duplication *AARS (chr16:70286199-70296541)* Exon 10 deletion *CTNNA3 (chr10:68,381,430-68,381,544)*	20, 30, 40, 50, 60, 70, 80	0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40
Clinical Sample 3	NA24385	Whole gene deletion *NPHP1 (chr2:110858517-110983174)* Exons 4–13 deletion *ALG1 (chr16:5127103-5137339)*	20, 30, 40, 50, 60, 70, 80	0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40

Table S1. Benchmark samples for validating mosaic variant detection. Samples NA24385, NA12878, NA24631, and NA12878 are genomic DNA samples that have been well characterized by the Genome in a Bottle (GIAB) consortium.[1] For the GIAB samples, unique variants were heterozygous variants (i.e., in positions that differed from the reference genome) that were not observed in the mixture-partner sample. For the clinical samples, unique variants were copy number variants. For both sample types, the unique variants met three criteria: 1) the variant was heterozygous in genome A and reference-matching in genome B (or vice versa), 2) the variant was called with high quality ("PASS" filter value), and 3) the variant was within the assay target region. In the GIAB samples only, variants were also required to be in high-confidence regions (as determined by GIAB) for both genomes in a mixture. Estimated genomic coordinates for CNVs in clinical samples are based on human genome build GRCh37/hg19.

Coverage level	Quality filter flags[a]	Prior	AB threshold
300–400x	No	High	0.38
300–400x	No	Low	0.28
300–400x	Yes	NA, no mosaic variants called	
400–500x	No	High	0.39
400–500x	No	Low	0.30
300–400x	Yes	NA, no mosaic variants called	
>500x	No	High	0.40
>500x	No	Low	0.32
>500x	Yes	High	0.40
>500x	Yes	Low	0.32

Table S2. Allele balance thresholds for mosaic single nucleotide variants. Allele balance (AB) thresholds for single nucleotide variants (SNVs) were determined based on sequencing depth, quality filters, prior probability that the affected gene contained a mosaic variant (i.e., prior), and the AB distribution of the non-mosaic heterozygous state of the variant. SNVs falling below the AB threshold were predicted to be mosaic. When a variant was flagged for quality concerns ("Yes" in the second column), only variants with at least 500x coverage (sequencing depth) were eligible for mosaic variant calls. AB, allele balance; NA, not applicable. [a]"Yes" indicates that a quality measure such as sequencing depth or mapping quality was below standard thresholds.

Category	No. patients (%)
Age range (y)	
0–17	124,495 (12)
18–90	910,020 (88)
Biological sex	
Female	759,013 (73)
Male	275,550 (27)
Unknown	17 (<1)
Race/ethnicity reported on order form	
Ashkenazi Jewish	30,796 (3)
Asian	44,139 (4)
Black/African American	69,255 (7)
Hispanic	89,707 (9)
Other/Multiple[a]	121,457 (12)
White	587,508 (57)
Unknown	91,718 (9)
Referral clinical area	
Cardiology	79,962 (8)
Hereditary cancer	678,964 (66)
Neurology	60,648 (6)
Pediatrics and rare disease	120,107 (12)
Reproductive carrier screening	94,899 (9)

Table S4. Characteristics of individuals in cohort (N=1,034,580).

[a:] Other includes categories for which <20 individuals were identified, including but not limited to French Canadian, Mediterranean, Native American, Pacific Islander, and Sephardic Jewish. Multiple includes 2 or more combinations of categories shown in the table or other categories not presented.

Inheritance	No. cases of variant combination	Classification of mosaic variants[a]	Classification of non-mosaic variants[a]	Significance of mosaic variant
Autosomal dominant	51	P/LP	P/LP	Possible explanation for disease
	48	P/LP	VUS	Probable explanation for disease
	83	VUS	P/LP	Likely incidental
	61	VUS	VUS	Uncertain
Autosomal recessive	6	P/LP	P/LP	Likely explanation for disease
	1	P/LP	VUS	Possible explanation for disease
	1	VUS	P/LP	Possible explanation for disease
	19	VUS	VUS	Uncertain
X-linked	2	P/LP	P/LP	Likely explanation for disease in a female individual; likely incidental in a male individual
	0	P/LP	VUS	Possible explanation for disease in female or male individual
	1	VUS	P/LP	Possible explanation for disease in a female individual; likely incidental in a male individual
	1	VUS	VUS	Uncertain
All inheritance types	59	P/LP	P/LP	
	50	P/LP	VUS	
	85	VUS	P/LP	
	94[b]	VUS	VUS	

Table S6. Cases with mosaic variants alongside non-mosaic variants in the same gene. Counts are by gene; individuals with mosaic variants in multiple genes appear multiple times in this table. [a]Results that include P/LP variants or P/LP variants in combination with VUS are counted once as a P/LP variant. [b]Includes 1 count of a mosaic VUS + non-mosaic VUS in a gene associated with a condition of unknown inheritance. P/LP, pathogenic/likely pathogenic; VUS, variant(s) of uncertain significance.

Test clinical area	No. genes in clinical area	No. genes with any P/LP variant	No. genes with P/LP mosaic variant(s)	% genes with P/LP mosaic variants among genes with any P/LP variant
Cardiology	640	230	23	10
Carrier	301	299	33	11
Hereditary cancer	435	137	64	47
Neurology	591	267	28	10
Pediatrics and rare disease	1,810	1,040	69	7

Table S7. Pathogenic and likely pathogenic mosaic variants by test clinical area observed in a clinical cohort of 1,034,580 unrelated individuals. Pathogenic and likely pathogenic variants include pre-mutation alleles and increased-risk alleles. Many carrier panel genes are found in other test (referral) clinical areas. The 301 genes in the carrier screening test clinical area are those sequenced as part of patient referrals for carrier screening only; many other genes are screened in both carrier screening and another clinical area, but are only assigned to the non-carrier screening test clinical area.
P/LP, pathogenic/likely pathogenic.

Gene	Associated condition(s)
ACADM	Medium-chain acyl-coenzyme A dehydrogenase deficiency
ATM	Ataxia telangiectasia
CFTR	Cystic fibrosis
EXT1	Hereditary multiple osteochondromas; Trichorhinophalangeal syndrome type II
KRIT1	Cerebral cavernous malformation
MKS1	Joubert syndrome; Bardet-Biedl syndrome; Meckel syndrome
FBN1	Marfan syndrome
NF1	Neurofibromatosis
PAX6	Aniridia; coloboma
RB1	Retinoblastoma
PDCD10	Cerebral cavernous malformation
SCN1A	Dravet syndrome; Generalized epilepsy with febrile seizures plus, type 2
SLC2A1	GLUT1 deficiency syndrome

Table S8. Distinctive phenotype genes with mosaic variants in a clinical cohort of 1,034,580 unrelated individuals. The genes above are associated with diseases that have explicitly defined clinical criteria for diagnosis. When these criteria are met, the clinical information for a patient can be used to predict variant pathogenicity. See Johnson et al. 2022.[2]

Supplemental Methods

Validation of mosaic SNV and indel detection

To evaluate our ability to detect mosaic single nucleotide variants (SNVs) and insertions and deletions (indels), we created a set of benchmark mosaic samples by mixing previously sequenced genomic DNA samples in a series of concentrations to simulate varying levels of mosaicism (Table S1). To achieve this, genomic DNA isolated from two well-characterized cell lines and mixed to create benchmark samples that harbored two different bases at a variety of positions. These mixed-base positions were considered benchmark mosaic variants and had allele balances (ABs) ranging from 0.05 to 0.40 in increments of 0.05 according to the relative concentrations of the two input genomes. (AB is the number of reads containing a specific allele divided by the total number of reads aligning to the specific genomic locus.)

In the benchmark samples, we evaluated 3,745 variants (3,639 SNVs and 106 indels) detected in 823 genes. We bioinformatically downsampled the sequencing data and repeated the variant calling to mimic mosaic variant detection at lower sequencing coverage depths. We evaluated variants that were flagged with any of our internal quality filters (e.g., strand bias, proximity to repetitive genomic regions) separately from those without any quality flags. Regions with known pseudogenes or other duplications (e.g., *PMS2*/*PMS2CL*, *SMN1*/*SMN2*, *NEB*) were excluded from the validation study.

We first used a simple count to confirm that we could observe the benchmark mosaic variants at the expected ABs. Overall, we found that all benchmark mosaic SNVs and indels were detected within 0.05 AB of their predicted levels; the majority were within 0.02 AB.

We next evaluated our analytic sensitivity for detecting mosaic SNVs and indels by calculating the percentage of expected benchmark mosaic variants that were confidently observed in the mixed samples:

$$(No.\ observed\ benchmark\ mosaic\ variants \div No.\ known\ benchmark\ mosaic\ variants) \times 100$$

In each mixed genome dataset, the coverage depth (i.e., the number of sequencing reads) at each location with a benchmark mosaic variant was assigned to a coverage-depth bin, with bins created in increments of 100x coverage. Sensitivity was evaluated by coverage depth, variant type (i.e., SNV, insertion, deletion), size (1–4bp, 5–10bp, 11–20bp, >20bp), expected AB, and genomic context (whether near a repetitive region or GC-rich "bad promoter").[3,4] The sensitivity of our detection of benchmark mosaic SNVs was 100% above 0.15 AB, was reduced at 0.05–0.15 AB, and dropped to 0 below 0.05 AB (Figure S1A). Although there was reduced statistical precision in the indel data (Figure S1B–C), the overall pattern was consistent with SNVs and showed ~90% sensitivity above 0.15 AB and 0% sensitivity below 0.05 AB.

We then determined the lower and upper bounds of AB between which we could call mosaicism with high confidence. The lower bound of detection was straightforward to determine because our standard next-generation sequencing (NGS) sequencing method does not allow reliable detection of ABs below ~0.05–0.10 due to technical limitations such as read depth, read and call quality, and potential strand bias. On the other hand, establishing an upper bound of mosaic variant calls that separates them from non-mosaic variant calls in heterozygotes (with ABs near 0.5) required more detailed consideration. For the upper bound, we empirically determined variant-specific AB thresholds below which a given variant would be predicted to be mosaic and above which the variant would be called non-mosaic.

These AB thresholds were evaluated for specificity and sensitivity to discriminate between mosaic and non-mosaic variants in our mixed-genome validation experiments. Using this approach we were able to balance the goals of maximizing the detection of true mosaic variants while minimizing false positives in non-mosaic heterozygotes. We began by comparing the AB distributions of benchmark

mosaic variants with those of non-mosaic variants (Figure S2). The AB distribution of mosaic variants was empirically determined from the mixed benchmark samples (Figure S2A), while the AB distribution for the non-mosaic variants was empirically determined from sequence data from the unmixed GIAB specimen NA12878, which was sequenced 3,007 times (Figure S2B). The ratio of the two integrals (i.e., total area under the curve) of the AB distributions of the benchmark mosaic and non-mosaic variants were adjusted with a correction factor (CF) so that the ratio reflected the relative prevalence of mosaic and non-mosaic variants as observed by our lab and in the literature (Figure S2C). In other words, the CF was applied to adjust for the *a priori* likelihood that a given gene would harbor a mosaic variant, as certain genes such as *TP53* have a greater propensity for mosaicism than others.

We then considered high and low candidate thresholds that would distinguish a mosaic variant from a non-mosaic variant based on an observed variant's AB (Figure S2D). For a given variant type, depth of coverage, and prior probability that a gene would harbor a mosaic variant (i.e., "prior"), we sought thresholds that best balanced the positive predictive value (PPV) and sensitivity. Multiple factors affected the performance of a given threshold. As expected, there was a tradeoff between specificity (and therefore PPV) and sensitivity: higher sensitivity resulted in lower PPV, and higher PPV resulted in lower sensitivity. We also found that the prior impacted the balance between true positives and false positives in specific genes, which in turn affected the performance of the threshold (Figure S3B–D).

After investigating thresholds across a wide range of priors, we found we could most easily optimize performance of our mosaic-calling pipeline by creating two categories of thresholds: one for high-prior genes and another for low-prior genes. For high-prior genes, a higher AB threshold was selected as it maximized sensitivity without compromising PPV. For low-prior genes, a lower, more conservative AB threshold was required to minimize false positives and maintain a high PPV (>80% for most genes) without substantial loss of sensitivity. Final AB thresholds were determined as a function of

variant type, prior probability of mosaicism, non-mosaic heterozygous AB distributions, sequencing coverage, and presence or absence of variant call quality filter flags. The resulting threshold values are shown in Table S2.

Based on our benchmarking studies, we assigned a high prior to six genes in which ≥10% of previously reported variants were mosaic: *ACTB* [MIM: 102630], *GATA1* [MIM: 305371], *PITX3* [MIM: 602669], *TP53* [MIM: 191170], *WDR45* [MIM: 300526], and *ZIC2* [MIM: 603073]. All other genes were considered low prior genes in which <10% of previously reported variants were mosaic.

Validation of mosaic CNV detection

To evaluate our ability to call mosaic copy number variants (CNVs), we used three clinical samples previously sequenced by the lab and known to harbor CNVs in five genes (Table S1). These samples were diluted with GIAB genomic DNA samples to simulate seven levels of mosaicism, with expected ABs ranging from 0.10 to 0.40. Due to the limited number of benchmark mosaic CNVs available, we could not statistically analyze the performance of our mosaic CNV detection method and instead present a qualitative assessment. Following our laboratory's standard processes, CNVs that were flagged as low quality (e.g., due to AB, strand bias, depth of coverage) were manually reviewed by lab personnel for the presence of mosaic CNVs. For benchmark mosaic CNVs, the observed ABs (computed from ABs of SNVs outside the affected gene/gene region) were often lower than the ABs expected based on the dilution of the DNA in mixed genome samples. Qualitatively, our method performed best for ABs of 0.14–0.30, for which the sensitivity was roughly 43% (Figure S4). There was no clear correlation between sensitivity and CNV size or copy number.

Validation of other mosaic variants

Sensitivity for larger indels and variants in more complex parts of the genome (e.g., in repetitive regions) was difficult to determine due to the small number of these variants in the mixed genome samples. Qualitatively, sensitivity for calling these variants was reduced compared with sensitivity for calling SNVs and indels.

Supplemental References

1. Zook, J.M., McDaniel, J., Olson, N.D., Wagner, J., Parikh, H., Heaton, H., Irvine, S.A., Trigg, L., Truty, R., McLean, C.Y., et al. (2019). An open resource for accurately benchmarking small variant and reference calls. Nat. Biotechnol. *37*, 561–566.

2. Johnson, B., Ouyang, K., Frank, L., Truty, R., Rojahn, S., Morales, A., Aradhya, S., and Nykamp, K. (2022). Systematic use of phenotype evidence in clinical genetic testing reduces the frequency of variants of uncertain significance. Am. J. Med. Genet. A. 188, 2642-2651.

3. Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome Biol. *14*, R51.

4. Krusche, P., Trigg, L., Boutros, P.C., Mason, C.E., De La Vega, F.M., Moore, B.L., Gonzalez-Porta, M., Eberle, M.A., Tezak, Z., Lababidi, S., et al. (2019). Best practices for benchmarking germline small-variant calls in human genomes. Nat. Biotechnol. *37*, 555–560.