nature reviews genetics

https://doi.org/10.1038/s41576-024-00709-x

Review article

Check for updates

Genetic variation across and within individuals

Zhi Yu[1,2,4], Tim H. H. Coorens[1,4], Md Mesbah Uddin[1,2], Kristin G. Ardlie[1], Niall Lennon[1] & Pradeep Natarajan[1,2,3]

Abstract

Germline variation and somatic mutation are intricately connected and together shape human traits and disease risks. Germline variants are present from conception, but they vary between individuals and accumulate over generations. By contrast, somatic mutations accumulate throughout life in a mosaic manner within an individual due to intrinsic and extrinsic sources of mutations and selection pressures acting on cells. Recent advancements, such as improved detection methods and increased resources for association studies, have drastically expanded our ability to investigate germline and somatic genetic variation and compare underlying mutational processes. A better understanding of the similarities and differences in the types, rates and patterns of germline and somatic variants, as well as their interplay, will help elucidate the mechanisms underlying their distinct yet interlinked roles in human health and biology.

Sections

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Cardiovascular Research Center and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [3]Department of Medicine, Harvard Medical School, Boston, MA, USA. [4]These authors contributed equally: Zhi Yu, Tim H. H. Coorens. e-mail: pnatarajan@mgh.harvard.edu

Introduction

Genetic variations between (that is, germline variants) and within (that is, somatic mutations) individuals underpin many phenotypic differences. Germline variants are inherited and templated in all cells of an individual (Fig. 1a), with evolutionary forces such as selection, recombination and drift shaping their frequency and distribution in a population over generations[1]. By contrast, somatic mutations accumulate in a mosaic manner within an individual from conception onward as a result of DNA damage or errors in DNA repair (Fig. 1a). Somatic mutations originate in a single cell and are propagated within an individual by DNA replication and cell division[2–6] (Fig. 1b). These fundamental differences between germline and somatic variants underlie their analogous modes of study either between individuals (germline) or between cells within an individual (somatic) (Fig. 1c).

Both germline and somatic variants contribute to human disease. More specifically, germline variants underlie inherited genetic conditions, such as Huntington disease[7], familial hypercholesterolaemia[8] and breast cancer predisposition syndromes[9,10]. Although most somatic mutations do not have a noticeable phenotypic effect, some can alter key cellular functions and potentially culminate in cancer[11]. Furthermore, somatic mutations in multiple normal tissues play profound roles in non-oncologic diseases (reviewed in ref. 12), such as somatic mutations that contribute to Alzheimer disease in neurons[13]. Moreover, somatic mutations that lead to clonal expansion in non-neoplastic blood cells are associated with a range of non-oncologic conditions, such as atherosclerosis and chronic liver disease[14–16].

Germline and somatic variants are inherently connected. All germline variants originate as de novo somatic mutations either in parental germ cells or very early in embryonic development. As such, mutations introduced and retained in germ cells, if passed on to the next generation, effectively become de novo germline variants. Additionally, germline variants, especially those in genes encoding DNA repair proteins, can influence somatic mutation rates and patterns[17].

This Review draws on insights gained from decades of DNA sequencing and more recent omics approaches (Box 1) to delineate the fundamental features of germline and somatic variants. More specifically, we define how germline and somatic variants compare in their mutation rates, types and patterns, and we describe approaches to their detection and analytical considerations for genetic association studies. We also highlight instances of interplay between germline and somatic variants, including somatic reversal of germline variants and germline predisposition for somatic mutations. Finally, we underscore how the biological and clinical significance of germline and somatic variants will continue to be elucidated by diverse and longitudinal human studies, single-cell multi-omics and causal inference methods.

Rates and patterns

Germline and somatic variants have many shared characteristics, given that germline variants originate as somatic mutations in germ cells. However, fundamental differences in their heritability, effect and prevalence can affect their respective mutational types, rates and patterns.

Genetic variants can be divided into four main classes: substitutions, which are mainly comprised of single-nucleotide variants (SNVs); short (<50 bp) insertions and deletions (indels); structural variants, including large deletions, segmental duplications, inversions, translocations and transposable element insertions; and other large chromosomal abnormalities, including whole-chromosome losses and gains. Segmental duplications, large deletions and often whole-chromosome alterations are also referred to as copy number variations. These distinct classes are consequences of different mutagenic processes: SNVs and indels are a consequence of erroneous DNA damage repair or replication, whereas structural variants can result from errors introduced during DNA double-strand break repair, mitotic or meiotic recombination, chromosome lagging or chromosomal missegregation, the latter of which can also cause somatic aneuploidy. Notably, acquired or inherited deficiencies (through somatic or germline variants, respectively) in DNA repair pathway components or DNA polymerases can change the frequency and type of mutations[18–20]. Finally, the genome may be mutated by the insertion of mobile genetic elements, such as retrotransposons[21].

Somatic mutations

Different mutagenic processes produce distinct rates and patterns (known as mutational signatures) of somatic mutations. For example, mutational signatures specifically for SNVs typically refer to the distribution of base changes in specific trinucleotide contexts[22,23], and similar signatures are defined for other classes of mutations, including indels[22], chromosomal alterations[24] and structural variants[25] (see COSMIC database). Recent studies have shown that somatic mutation rates differ across tissues and age ranges in the same individual[26–36] (Fig. 2a). Notably, somatic mutation rates in different tissues are influenced by exposure to both cell-intrinsic and cell-extrinsic causes, such as smoking, ultraviolet light and chemical mutagens. Even when restricted to endogenous mutagenic sources, the mutation rate of SNVs greatly varies across cell types. For example, approximately 17 SNVs are acquired per year in neurons[30] and haematopoietic stem cells[32], whereas 28 and 44 SNVs are acquired per year in endometrial stem cells[34] and colonic stem cells, respectively[31]. The pattern of somatic mutations, or mutational signature, also varies based on the source of mutations (Fig. 2b). Normal tissues vary in their mutational signatures, but all tissues have some trace of the linear, clock-like mutational signatures[27] referred to as single base substitution signature 1 (SBS1; C>T mutations in a CpG context) and SBS5 (a flat signature) in the COSMIC database[22]. In addition, many tissues also exhibit SBS18, which is characterized by C>A mutations and is a consequence of oxidative damage. Expectedly, some mutagens are restricted to certain organs; for example, SBS7, the consequence of ultraviolet light damage, is a common mutational signature in the skin, and it consists of C>T mutations. Normal tissues also show mutational patterns due to other exposures, such as smoking (SBS4, dominated by C>A mutations)[35], chemotherapy (including SBS31, dominated by C>T mutations in CCC and CCT contexts; and SBS35, a collection of C>A, C>T and T>A mutations), and a genotoxic strain of *Escherichia coli* (SBS88; substitutions of T when preceded by A or T)[27,31].

The vast majority of somatic mutations will have no phenotypic effect on the cells that harbour them, especially those that affect non-coding regions of the genome or induce a synonymous change in genes. However, occasionally a cell will acquire a mutation that carries a selective advantage, such as increased proliferation or survival. These clonal expansions become more widespread with age (as mutations accumulate over time) and are enriched in rapidly dividing tissues in which clones are unconstrained, such as the sheet-like epithelia of the skin[37], oesophagus[38] and bladder[29], as well as blood[32,39]. Clonal expansions in the blood-forming system are referred to as clonal haematopoiesis. By contrast, large clonal expansions are rare in glandular epithelium, such as the colon, because of physical constraints imposed by the tissue architecture, barring a history of damage and regrowth, such as in inflammatory bowel disease or the normal endometrial epithelium.

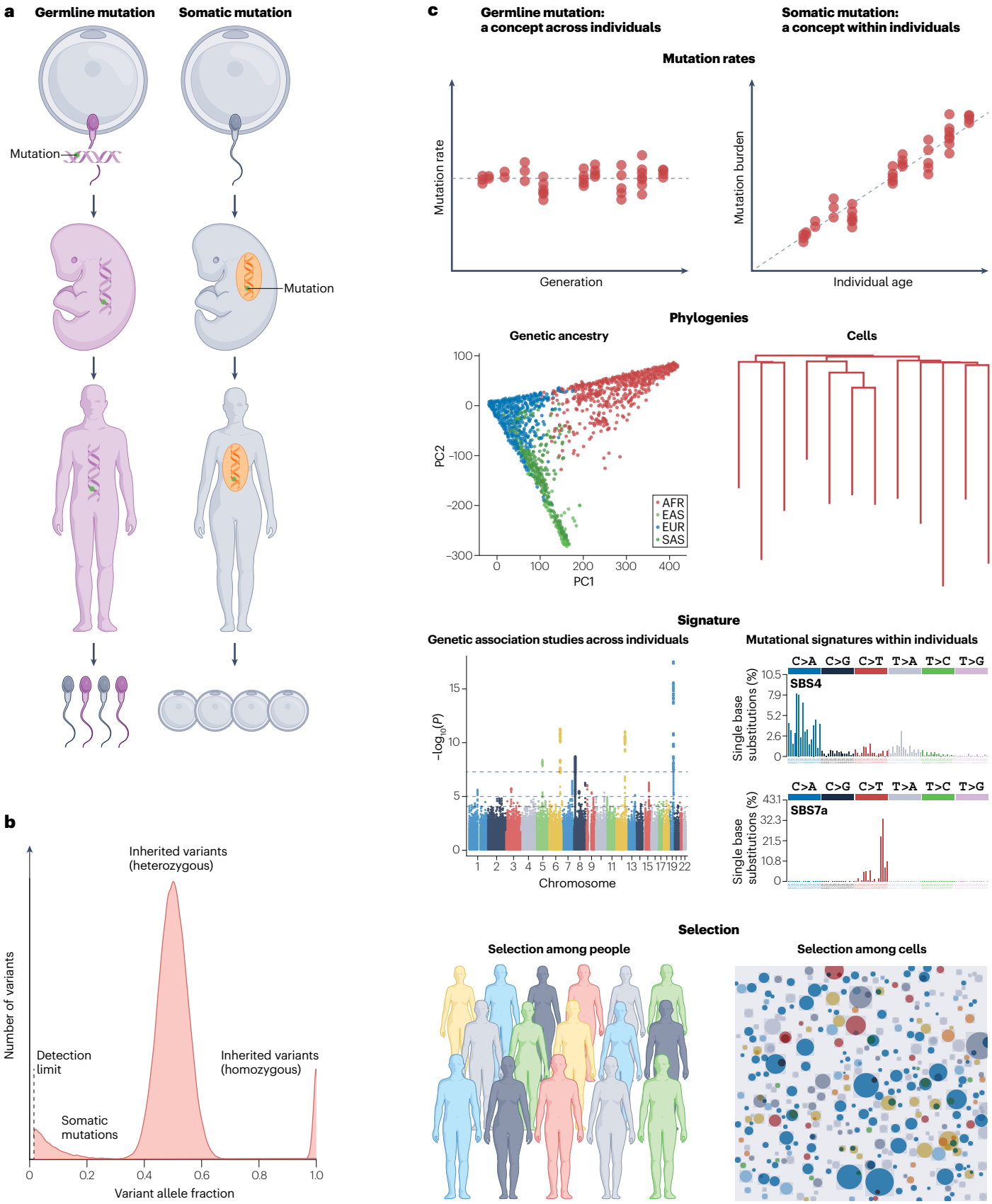

a
Germline mutation
Somatic mutation
Mutation
Mutation
b
Inherited variants (heterozygous)
Number of variants
Detection limit
Somatic mutations
Inherited variants (homozygous)
0
0.2
0.4
0.6
0.8
1.0
Variant allele fraction
c
Germline mutation: a concept across individuals
Somatic mutation: a concept within individuals
Mutation rates
Mutation rate
Generation
Mutation burden
Individual age
Phylogenies
Genetic ancestry
PC2
PC1
AFR
EAS
EUR
SAS
Cells
Signature
Genetic association studies across individuals
$-\log_{10}(P)$
Chromosome
Mutational signatures within individuals
C>A C>G C>T T>A T>C T>G
SBS4
Single base substitutions (%)
SBS7a
Selection
Selection among people
Selection among cells

Fig. 1 | Comparison of variants across (germline) and within (somatic) individuals. a, Origins of germline and somatic variants. Germline variants are inherited at conception from a parent and are transmissible to the offspring; somatic variants are only present in a single cell and its progeny, and they cannot be inherited (unless they occur in germ cells). **b**, The distribution of allelic fractions observed in sequencing data. The distributions that are close to 50% and 100% represent germline variants (present in all cells) that are heterozygous and homozygous, respectively, whereas the smaller distribution, close to 0%, represents putative somatic variants (present in some cells)[152]. **c**, Analogous aspects of studies on germline and somatic variants: rates of mutation acquisition across generations (germline) and individual age (somatic), ancestry (germline) and cell lineage (somatic) tracing, genetic association (germline) and mutational (somatic) signatures, and selection of specific variants in populations comprised of human individuals (germline) or cells (somatic)[2,153]. AFR, African; EAS, East Asian; EUR, European; PC1, principal component 1; PC2, principal component 2; SAS, South Asian; SBS, single base substitution. Part **c** reprinted with permission from ref. 153, Wolters Kluwer.

In many normal tissues, clones with somatic mutations that are under positive selection and have been previously associated with cancer may lead to malignant transformation[34,37,38]. However, some somatic mutations can induce clonal expansion without leading to cancer, such as those underlying clonal haematopoiesis that is linked to many non-cancer diseases[15,39,40]. Somatic mutations can even protect against cancer; *NOTCH1*-mutant clones in the oesophageal epithelium are known to outcompete pre-cancer clones[41,42]. Alternatively, recurrent somatic mutations may mitigate the effects of a disease. Such mitigation was recently demonstrated in chronic liver diseases, in which different clones recurrently and independently acquired somatic mutations that lead to escaping disease-related toxicity[43].

Somatic mutations in germ cells

Somatic mutations that accumulate in germ cells, when passed on to the next generation, become de novo germline variants. Therefore, the mutational rates and patterns specifically within these germ cells greatly influence the generation of germline variants. The seminiferous tubules of the testes, which produce the spermatocytes, accrue approximately 2.7 SNVs per year, the lowest observed mutation rate among the examined tissues[27]. Notably, this mutation rate is estimated for the diploid genome of spermatogonial stem cells rather than the haploid spermatocytes. Therefore, spermatocytes should accumulate somatic mutations at half the rate of spermatogonial stem cells. Indeed, the mutation rate in seminiferous tubules reflects the estimated paternal age effect on de novo germline variants, which is approximately 1.4 per year of the father's age, in contrast to 0.37 per year of the mother's age[44–46]. Although these tissues mostly exhibit age-related mutational signatures (SBS1 and SBS5), mutagenic exposures, such as chemotherapy treatment, can increase the mutation rates in germ cells[47].

The effects of somatic variants in germ cells can greatly influence the rate of transmission to the next generation. Some somatic variants are termed 'selfish' because they can subvert normal germline processes to increase the likelihood of propagating to the next generation, which often leads to major developmental disorders[48,49]. For example, some mutations in the *FGFR2* gene are under positive selection in the seminiferous tubules of the father and lead to large clonal expansions; however, when passed on to the next generation, these *FGFR2* mutations cause Crouzon syndrome, a genetic disorder characterized by the premature fusion of certain skull bones[50]. Alternatively, somatic variants in germ cells that are ultimately fertilized can result in embryonic lethality and therefore never manifest as germline variants. For example, most germline aneuploidies are lethal, with the notable exception of trisomy 21[51]. In addition, many variants classically associated with cancer, such as *BRAF* V600E, have never been reported as germline variants, suggesting that they are lethal despite assumedly driving increased propagation in germ cells[52]. In summary, the mutation rates of germ cells and the effects of these variants on embryogenesis are major determinants of de novo germline variants. Future work is needed to compare somatic variants in germ cells with de novo germline variants to understand the contribution of embryonic lethality to rates and patterns of germline variants.

Germline variants

Germline variants are mostly static throughout an individual's lifetime, but the pattern of these variants is dynamic across populations and generations. Evolutionary forces, such as selection and genetic drift, can shape germline variant patterns, leading to distinct genetic signatures among different populations and species. These genetic signatures can offer insights into evolutionary processes, adaptation and the genetic basis of heritable diseases. An illustrative case is the positive selection observed on variants in the *G6PD* gene, which confer a protective advantage against malaria and are more prevalent in regions where this disease is endemic[53].

Advancements in detection

Understanding the intricate landscape of germline and somatic variation depends on their accurate identification. Not only have advancements in detection technologies improved the accuracy of variant identification, but they have also enhanced our ability to delve into the genetic underpinnings of disease at a more granular level.

Germline variants

Given that germline variants are present in most cells within an individual, the variants most challenging to detect are those in regions difficult to resolve through whole-genome sequencing, such as repeat-rich regions. However, recent technological advancements, such as long-read sequencing technologies and novel algorithms[54], have improved overall germline variant detection (Box 1). More specifically, long reads can cover repetitive genomic regions that are inaccessible to short reads[55] and deep learning models, such as DeepVariant, enhance variant detection by minimizing the dependence on arbitrary rules and filters[56]. Despite the remaining challenges of highly polymorphic and duplicate-rich regions of the genome that often complicate calling variants in clinically relevant genes (for example, *PMS2* and *SMN1*), current germline variant calling achieves high accuracy, especially for SNVs (>99.9% in benchmark regions[57]).

Somatic mutations

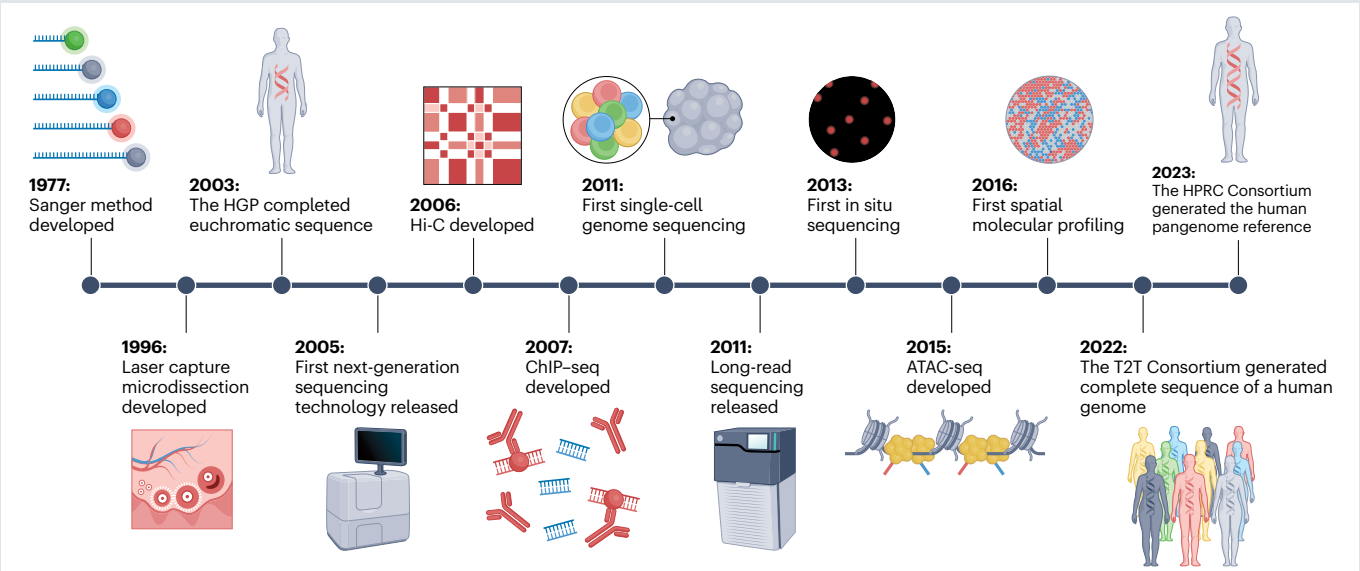
The detection of somatic mutations requires additional considerations, because they are present in relatively few cells. Many insights into somatic mutations originally came from whole-genome sequencing of cancer tissue, which are large single-cell-derived clones. The genome sequence of the cancer cells is then compared to that of a large aggregate of normal cells (often blood) and any sequence differences are attributed to somatic mutations. However, somatic variants in normal

Box 1

Milestones in the study of somatic and germline variants

Advancements in genetic research from 1977 to 2023 have enabled high-resolution variant identification, large-scale DNA sequencing, cell-type-specific regulation understanding and breakthroughs in repetitive regions and structural variants (see the figure). Together, these innovations have enhanced our understanding of the genetics underlying human history and health.

Since the Human Genome Project (HGP) completed its goal to sequence all human euchromatin in 2003[158,159], whole genome and exome sequencing — combined with advanced bioinformatics techniques — have facilitated the high-resolution identification of variants and the detailed study of somatic and germline variants. Decreasing costs have led to large-scale bulk DNA sequencing studies, such as the UK Biobank and NHLBI Trans-Omics for Precision Medicine (TOPMed) projects[160–162], that have supplied the population-scale genetic data required to link genetics to both human history and health[61,163–165]. Concurrently, the emergence of single-cell DNA sequencing, complemented with laser capture microdissection and other in situ techniques, provides the means to identify somatic variants at the granularity of cellular subpopulations and individual cells[61,163–169]. These techniques can be integrated with other functional genomics approaches, such as Hi-C, assay for transposase-accessible chromatin with sequencing (ATAC-Seq), chromatin immunoprecipitation followed by sequencing (ChIP–seq) or spatial molecular profiling[79–85], to inform mechanisms of cell-type specific regulation and the positional context of these cells within a tissue[170–172]. In addition, the advent of long-read sequencing technologies has enabled more accurate assembly and detection of complex genomic elements, including structural variants and repetitive sequences. Initiatives such as the Telomere-to-Telomere (T2T) Consortium, the Human Pangenome Reference Consortium (HPRC) and the Human Genome Structural Variation Consortium (HGSVC) are using long-read sequencing technology to capture the full spectrum of genomic diversity, including the heterochromatic regions previously unfinished by the Human Genome Project 20 years ago. Our understanding of germline and somatic variants is poised to substantially expand as these novel technologies are adopted more widely. This paradigm shift is expected to enhance genomic maps of both germline and, consequently, somatic variation, particularly with respect to structural variants that are currently under-represented.



tissues are not typically found in large clonal aggregates, so they tend to be rare and difficult to differentiate from errors introduced during library preparation and sequencing. This error rate depends on a variety of factors, including the chemistry and specific technology used for sequencing. Recent technological advancements have enabled researchers to surmount these obstacles and increase the detection sensitivity for somatic variants (Table 1).

Increased sequencing depth. The most straightforward method to improve the detection of somatic variants is to increase the sequencing depth. However, to avoid dramatically increasing sequencing costs, this approach is often paired with a 'bait capture' of genomic regions of interest, such as genes known to be under selection or often harbouring cancer-associated mutations. Bait capture relies on hybridization of oligonucleotide probes to enrich the DNA of the genomic regions of interest. Prior work using targeted sequencing experiments can reliably detect variants in ~0.1–1.0% of cells[37,38], below which the mutation detection is constrained by the error rate of sequencing. Because of its sensitivity and depth, this approach excels at cell population-level inferences on the selection of mutations in specific

genes, which is important to identify 'driver' mutations that confer a selective advantage to cells. However, given the limited genomic footprint, relatively few mutations are detected overall, which hampers the study of mutational burdens and signatures. Sequencing polyclonal populations of cells also precludes the precise reconstruction of phylogenetic trees, as it is impossible to prove that different somatic mutations co-occur in the same cells or occur in different cell populations. Of note, when the sample consists of a dominant clone with subclones, as is the case for cancers, a rudimentary phylogeny can be reconstructed through clustering mutations by their variant allele frequency into clones[58].

Sequencing single cells or clones. An alternative approach to increasing detection sensitivity of somatic variants is to sequence the DNA of a single cell to obtain mutational readouts of a single lineage. However, a single cell does not have enough DNA to for reliable whole-genome sequencing and somatic variant detection. Single-cell DNA can be amplified using one of three methods: first, directly isolating the DNA of a single cell and biochemically amplifying the DNA[59]; second, isolating single cells and expanding into clones in vitro[32,60]; third, isolating the clonal progeny of a single cell in vivo, often using laser capture microdissection[27,31,36,61]. However, each of these approaches has downsides. Amplifying the DNA of a single cell can introduce artefactual

Fig. 2 | Patterns and rates of somatic variants. a, Current estimates of single-nucleotide variant (SNV) somatic mutation rates are variable between stem cells of the seminiferous tubules[27]; neurons[30,59]; haematopoietic stem cells[32]; stem cells of the bronchial[35], endometrial[34], colonic[31] and small intestinal[33] epithelium; and trophoblasts of the placenta[26]. **b**, Profiles of three COSMIC reference mutational signatures showing the single base substitutions (SBS) and flanking 5′ and 3′ bases. Profiles differ by source of mutation: top (SBS5) is a flat, clock-like signature associated with age; middle (SBS7) is linked to ultraviolet light (UV) damage; and bottom (SBS88) is induced by colibactin produced by a genotoxic strain of *Escherichia coli*. Reference mutational signatures from refs. 22,154.

Table 1 | Approaches for optimizing variant detection

Aim	Method	Description	Application	Pros	Cons
Increasing the depth of sequencing	Deep targeted or panel sequencing	Enriched sequencing of specific areas or genes of interest in the genome	Variant detection in known genes, for example cancer genes (only for SNVs and indels)	Sensitive and cost-effective for targeted areas	Limited to predetermined genomic regions
Lowering the error rate of sequencing	Duplex sequencing	Sequencing and comparison of both strands of DNA to detect mutations	Detection of average mutation burden and mutation signature (only for SNVs and indels)	Highly accurate and cost-effective	May cover the genome unevenly; may require a large amount of input DNA; high sequencing depth needed to call specific, rare variants
Isolating DNA from a single clone	Biochemical amplification of single-cell DNA	Isolation and sequencing of DNA from a single cell	Suitable for phylogenetic reconstruction (for all classes of mutations, if whole-genome sequencing implemented)	Can reveal true cell-to-cell genetic variability	Technically challenging; many amplification errors; possibility of allelic dropout
	In vitro expansions	Cells are grown in a lab to create a larger clonal sample for sequencing		Truly monoclonal sample	Possible culture-induced changes; time-consuming; bias towards stem cells
	Laser capture microdissection	Isolation of specific cells from a tissue sample using a laser		Can isolate specific cell types and retain spatial coordinates	Low throughput; not well-suited to tissues without clonal units; possible stromal contamination

This table summarizes various approaches implemented in the detection of somatic and germline variants. For each approach, the techniques, their specific applications and the advantages and disadvantages associated are provided. Indels, insertions and deletions; SNV, single-nucleotide variant.

variants and can exclude genomic regions that are difficult to amplify[59]. Cloning cells in vitro is laborious and works better for stem cells than differentiated cell types. The microdissection approach relies on the presence of typically rare clonal units in a tissue, but it yields high-quality genomic readouts and retains spatial information on tissue architecture[27,31,36,61]. Obtaining single-cell-derived readouts is crucial for lineage tracing and reconstructing the mutation-based phylogenies of normal cells.

Minimized error rate. Prior work has minimized the error rate of library preparation and sequencing by sequencing both the forward and reverse strands of a DNA duplex molecule[30,62]. A true somatic variant will be observed in both strands, whereas variants introduced during library preparation and sequencing will only appear in one of the two strands. This approach quadratically lowers the probability that a variant is observed erroneously and can lower the error rate to approximately 10^{-8} per base from the usual 10^{-4} per base on commonly used short read sequencing platforms. Notably, this increased sensitivity only applies to the detection of SNVs and indels. By achieving a lower error rate, duplex sequencing provides better estimations of average mutational burdens and signatures in cell populations.

Optimized variant calling. Finally, the bioinformatic approach to variant calling from aligned sequence data must be considered. Many different variant callers, including GATK MuTect2 (refs. 63,64), VarScan2 (ref. 65) and CaVEMan[66], are used to detect somatic variants. To distinguish between germline and somatic variants, the sample may be compared with another normal tissue sample, and any common variants are assumed to be from the germline[34]. Alternatively, unmatched calling approaches avoid the risk of filtering out shared variants that arose during very early embryogenesis, in which case the variant allele frequency across tissues can be used to distinguish between germline and somatic variants[2] (Fig. 1b). Furthermore, some variant callers, such as MoChA[40] and DeepMosaic[67], are specifically designed for mosaic variant calling by explicitly modelling or training on data of mosaic mutations. Generally, variant callers filter out recurrent artefacts by utilizing a panel of unmatched normal samples, ideally subjected to the exact same library preparation and sequencing protocol as the sample of interest[63,64].

Genetic association analyses

Genetic association analyses play a crucial role in uncovering the relationships between genetic variation and phenotypic traits, both in the context of germline and somatic variants. Germline analyses typically involve large-scale population studies to identify variants associated with specific traits or diseases. Conversely, many somatic mutation studies are focused on mutations within individual cells or tissues, often utilizing single-cell sequencing to discern their association with phenotypic changes.

Germline variants

Multiple advancements have drastically increased the breadth and depth of association analyses of germline variants (Fig. 3). First, studies have transitioned from single-population investigations to well-powered studies of mega-biobanks or multiple diverse cohorts due in part to sequencing cost reductions[68] (Fig. 3a). Second, high-throughput sequencing technologies and bioinformatics tools[69,70] have facilitated a shift from focusing solely on common coding variants to exploring the complex polygenic architecture across diverse populations, as well as rare coding, non-coding and structural variants[71–78] (Fig. 3b). Third, the functional effects of germline variants with phenotypic associations are now explored through the integration of techniques such as single-cell sequencing, CRISPR, Hi-C, assay for transposase-accessible chromatin with sequencing (ATAC-seq), chromatin immunoprecipitation followed by sequencing (ChIP–seq) and spatial molecular profiling[79–86] (Fig. 3c and Box 1). These innovative integrations have helped dissect regulatory and disease pathways acting at cell-specific or tissue-specific levels[87–90]. Fourth, further integration of these new types of data with enhanced statistical and machine learning methods has enabled researchers to identify disease causal variants and target genes[91–93], infer cell developmental trajectories[94] and predict gene expression accurately[95]. The recently introduced STAAR (variant-set test for association using annotation information) series exemplifies these advancements by developing

novel statistical methods for aggregated rare variant association tests. These methods incorporate multiple functional annotations and are scalable to large whole-genome sequencing datasets, as demonstrated using multi-ancestry whole-genome sequencing data from the TOPMed program[75,78,96].

As germline variant research evolves, conventional analytical considerations, such as statistical power, remain pertinent. Strategies to enhance power include increasing sample size and diversity of ancestries, utilizing more accurate models (such as those that account for gene–gene or gene–environment interactions), and aggregating genomic information (such as polygenic risk scores)[97–101]. Innovations accounting for linkage disequilibrium, population structure using principal component analysis, and relatedness using linear mixed models have more recently focused on scalability, given ongoing development and expansion of large-scale biobanks. These analytical considerations have been extensively addressed in the literature[102–104].

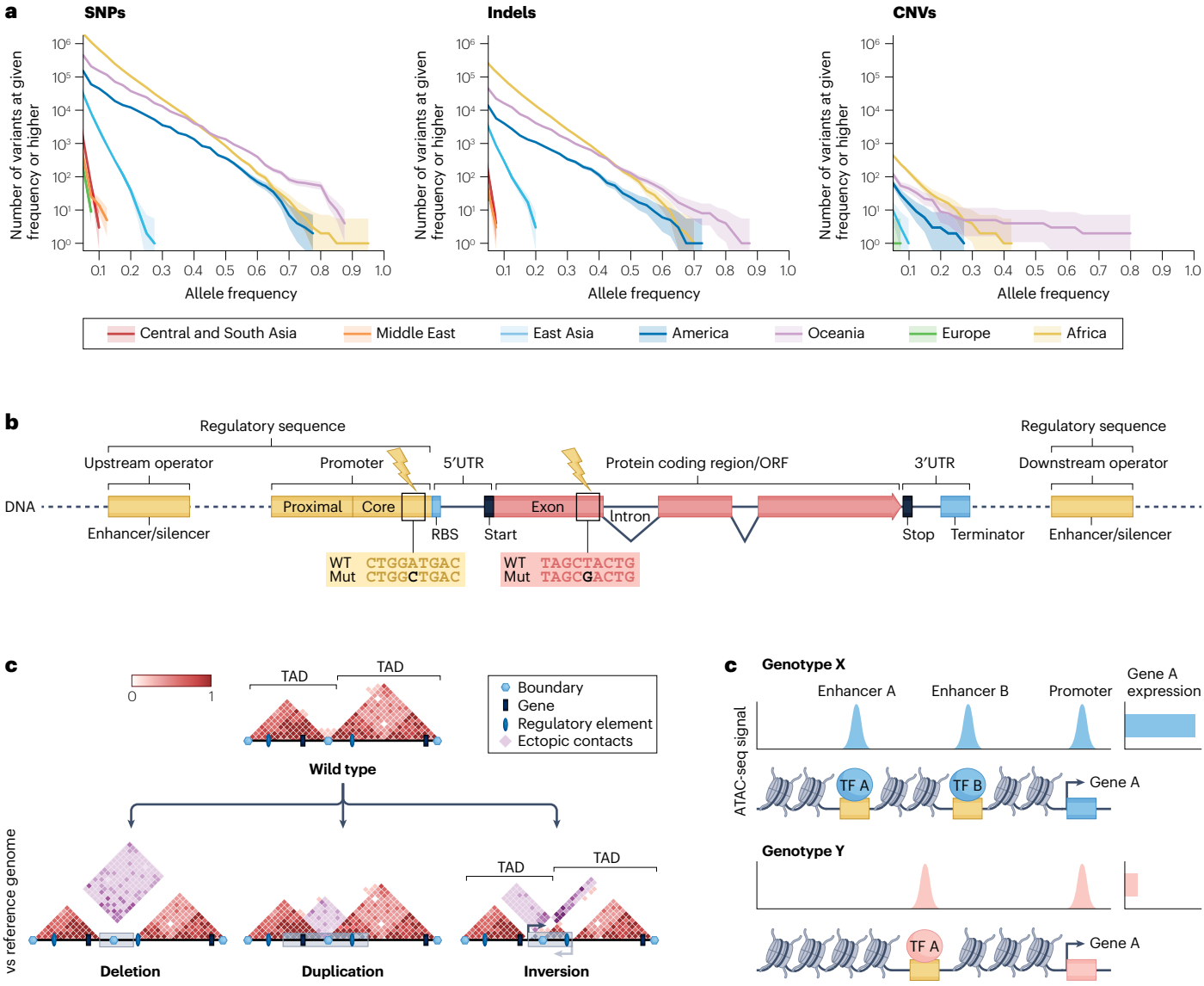


Fig. 3 | Advancements in association analyses of germline variants. **a**, Studying diverse populations helps identify additional variants that are not present in a single population. Counts of SNPs, insertions and deletions (indels), and copy number variations (CNVs) grouped by the geographical location of populations[155]. **b**, Technical and methodological advancements have facilitated a shift from focusing solely on common coding variants to exploring rare coding, non-coding and structural variants. The protein-coding sequence is shown in red, and the regulatory features that determine where and when the protein coding sequence will be expressed are shown in yellow. **c**, New technologies capture various mechanistic levels that can differ by genotype or cell type. Left: structural variation can induce dramatic changes in chromatin organization and thus create specific signatures that are noticeable by visual inspection of Hi-C interaction maps[156]. Right: cell type-specific assay for transposase-accessible chromatin with sequencing (ATAC-seq) peak due to differential chromatin accessibility between cell types X and Y[157]. Mut, mutant; ORF, open reading frame; RBS, ribosomal binding site; TAD, topologically associated domain; TF, transcription factor; UTR, untranslated region; WT, wild type. Part **a** reprinted with permission from ref. 155, AAAS. Part **c** reprinted from ref. 157, Springer Nature Limited; adapted from ref. 156, Springer Nature Limited.

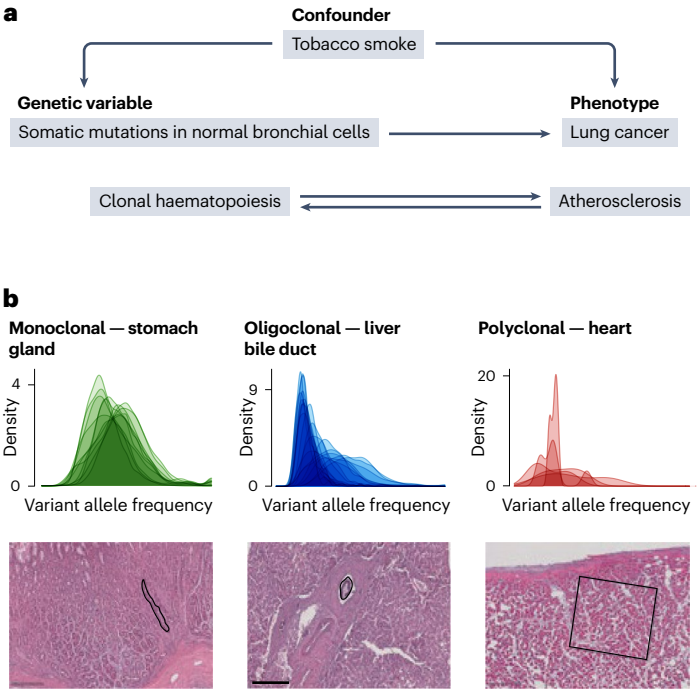


Fig. 4 | Factors to consider in association analyses of somatic variants. **a**, Causal inference. The effect (on human health) of somatic mutations can be confounded by environmental exposures and lifestyle, and can have potentially bi-directional causal relationships (that is, evidence supports clonal haematopoiesis as a causal risk factor of atherosclerosis, and atherosclerosis has also been shown to accelerate clonal haematopoiesis), necessitating careful consideration in analytical models. For example, tobacco smoking is associated with both increased somatic burden in normal bronchial cells and increased risk of lung cancer[35]; therefore, it confounds the association between somatic mutation and lung cancer. Murine evidence shows both clonal haematopoiesis leading to atherosclerosis and atherosclerosis leading to clonal haematopoiesis[106,107]. **b**, Tissue and cell specificity. Unlike germline variants, which are present in all cells, somatic mutations are present in a subset of cells within tissues. The typical pattern and clonal structure can be different between tissues and their architecture, with distribution of variant allele frequencies and histological sections shown for stomach (monoclonal tissue units), liver bile ducts (oligoclonal structure) and heart (polyclonal structure). Scale bars, 500 μm[27]. **c**, Heterogeneous impact. The molecular and clinical consequences of somatic mutations can differ depending on the driver gene involved, suggesting that a nuanced, gene-specific approach is often more informative than broad categorizations. For example, *TET2*-mutant clonal haematopoiesis of indeterminate potential causes atherosclerosis in both animal experiments and human studies, whereas the role of *DNMT3A*-mutant clonal haematopoiesis of indeterminate potential in atherosclerosis is less clear. Part **b** reprinted from ref. 27, Springer Nature Limited.

Somatic mutations

Association analyses of somatic variants differ from those of germline variants in several ways (Fig. 4). For somatic mutation studies, the sample collection required to identify mutations across tissues is inherently invasive, and thus tissue-comprehensive studies are often limited in sample size. To date, datasets with a wider range of tissues have typically been procured from a limited number of deceased donors[89]. For example, recent efforts to analyse somatic mosaicism across developmental stages have used tissue samples from deceased paediatric donors[105]. Therefore, given that many tissue types from living research participants are inaccessible, studies of population-based somatic mosaicism at scales comparable to germline mutation studies have focused mainly on blood DNA in adults.

Given the acquired and dynamic nature of somatic mutations, association analyses for somatic mutations need to also account for potential exposures that may jointly or separately influence the acquisition, fitness and clinical outcome of a somatic variant (Fig. 4a). For example, associations between somatic mutations and diseases may be confounded by factors such as smoking[35] or exhibit potential bi-directional causal relations (that is, each of two traits may be causal to the other at the same time)[14,106,107].

Given that somatic variants occur only in a subset of cells, association analyses of somatic variants often require tissue-specific and single-cell analyses (Fig. 4b). These analyses can identify rare subclones, track clonal evolution and investigate cellular phenotypes associated with specific variants[32,60,108]. Furthermore, recent advancements have surmounted a limitation of conventional single-cell technologies, so that cells are no longer destroyed in the process of sequencing. More specifically, DNA can be sequenced while simultaneously measuring other 'omics' phenotypes at the single-cell level, enriching our insights into the mechanisms of somatic variants[109]. Additionally, algorithmic innovations can now directly detect somatic variants in single-cell RNA sequencing (RNA-seq) and ATAC-seq reads without the need for matched DNA sequencing data, which allows repurposing many previously collected single-cell data sets that went through RNA-seq and ATAC-seq for somatic mutation calling[110]. These developments enable the capture and integration of multiple data modalities to inform how somatic mutations affect cellular function and regulation[111,112].

In addition, different driver genes in somatic mutations have diverse biological and clinical consequences and are typically better considered separately rather than as one. For example, biologically, two commonly mutated genes in clonal haematopoiesis, *DNMT3A* and *TET2*, are involved in methylation and demethylation, respectively[113,114]. Clinically, *TET2* mutations cause atherosclerosis in both animal experiments and human studies[14,107] whereas the role of *DNMT3A* mutations in atherosclerosis is less clear[115] (Fig. 4c).

Interplay of germline and somatic variants

Genetic variation across and within individuals is a dynamic mosaic of germline and somatic variants that can influence one another to shape

health and disease trajectories (Fig. 5). Their interplay has important clinical implications through contributions to the pathogenesis of various diseases, but it also has the potential to provide more effective, personalized therapeutic strategies.

Somatic rescue of germline syndrome

Somatic variants can partially or fully reverse the pathogenic effects of inherited germline variants, a phenomenon known as somatic genetic rescue (SGR; Fig. 5a). SGR has now been reported in over 30 different

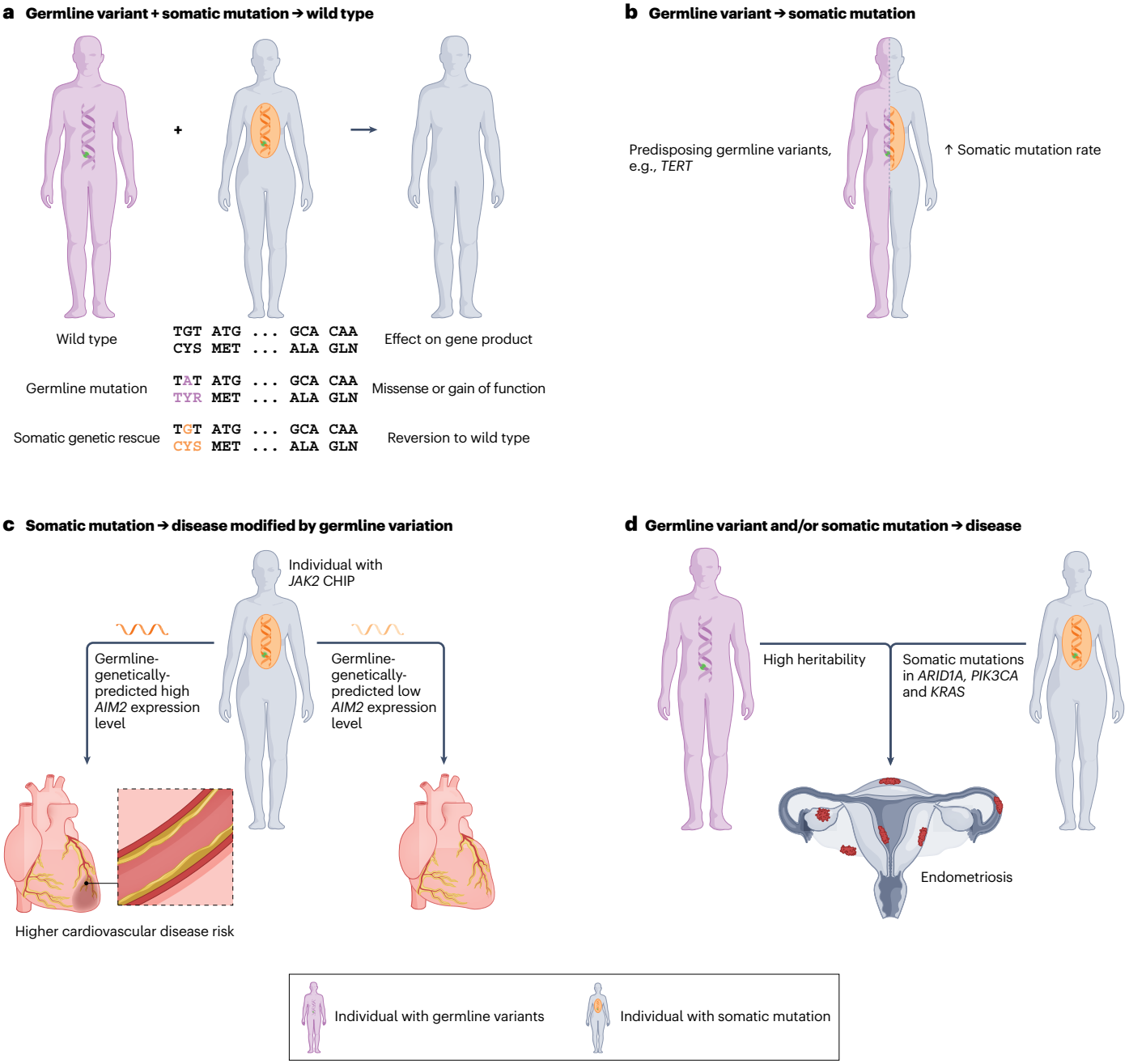


Fig. 5 | Interplay between germline and somatic variants. a, Somatic variants can reverse the pathogenic effects of germline variants (known as somatic genetic rescue) by either reverting to the non-pathogenic sequence or compensating for the germline defect through changes elsewhere, thus leading to variable disease phenotypes and therapeutic resistance[117]. **b**, Germline variants can predispose individuals to an increased rate of somatic mutations, as seen in clonal haematopoiesis[130,133] (which can contribute to the development of haematological cancer and many other non-cancer diseases). **c**, Germline gene expression levels can modify somatic mutation-associated disease risks. For example, *JAK2*-mutant clonal haematopoiesis of indeterminate potential (CHIP) is associated with increased cardiovascular disease risk, and that risk is reduced among individuals with genetically predicted low expression of *AIM2* (ref. 135). **d**, The risks of various cancers and other diseases can be increased by germline and/or somatic variants. For example, based on family studies, endometriosis has high heritability, and it is also associated with somatic variants in *ARID1A*, *PIK3CA* and *KRAS*[139,140].

haematopoietic disorders and several other diseases, such as breast cancer, which is caused by autosomal recessive, autosomal dominant and X-linked mutations[116–118]. The genetic mechanisms underlying SGR are diverse, including site-specific mutations that revert the original germline variant to the 'wild type' (that is, its common form found in the general population that is typically associated with the absence of the disease phenotype), second-site mutations that compensate for the germline defect, copy-neutral loss of heterozygosity and chromosomal deletions or rearrangements. Clinically, SGR can lead to milder disease phenotypes and delayed diagnoses, but it can also have neutral or negative effects[117]. A classic example with therapeutic relevance is germline variants in *BRCA1* or *BRCA2* that are related to cancer. Individuals with such variants might be initially responsive to poly (ADP-ribose) polymerase (PARP) inhibitors due to DNA repair defects, but they then become resistant to these inhibitors following additional mutations that restore some DNA repair function[119,120]. Looking ahead, CRISPR–Cas9 gene editing has been proposed as a method to controllably induce SGR for the targeted treatment of inherited monogenic disorders, such as Duchenne muscular dystrophy and cystic fibrosis[121]. In summary, understanding the genetic mechanisms and clinical effects of SGR has important diagnostic and therapeutic value.

Germline variants predisposing to somatic mutation risk

Prior work has identified two mechanisms by which germline variants can influence the risks of developing somatic mutations. First, some germline variants can increase the baseline rate of somatic mutations through, for example, inherited defects in DNA repair pathways. This mechanism is responsible for Bloom syndrome and Fanconi anaemia pathways, which are diseases associated with genomic instability and a higher likelihood of somatic mutation[122,123]. Second, individuals might develop malignancies following somatic mutations because of certain germline variants that predispose an individual to the clonal expansion of these cells. The occurrence of this phenomenon is evidenced by the increased risk of cancer development among first-degree relatives of cancer patients[9,10,124]. Another example of this phenomenon is that families with long telomere syndrome from *POT1* mutations also have increased familial risk of clonal haematopoiesis[125,126].

Further efforts to identify such germline variants have been enabled in recent years by larger biobanks and patient cohorts. However, most population-level studies have focused solely on somatic variants in the blood system, clonal haematopoiesis, due to the difficulty and cost of collecting non-blood tissue samples[17,127]. These population-level studies revealed heterogeneous germline genetic basis across different types of clonal haematopoiesis. For instance, genome-wide association studies on clonal haematopoiesis of indeterminate potential have identified over 20 loci near genes involved in haematopoietic stem cell self-renewal, proliferation, telomere maintenance and DNA damage response pathways[128–130] (Fig. 5b). By contrast, germline variants influencing mosaic loss of the X chromosome within blood cells, another type of clonal haematopoiesis, are primarily linked to genes with established roles in chromosomal missegregation, cancer predisposition and autoimmune diseases[131]. In addition, even within one type of clonal haematopoiesis, somatic mutations at different driver genes can have different germline genetic underpinnings. A relevant example is a germline locus on *TCL1A* where alleles associated with an increased risk of developing *DNMT3A*-mutant clonal haematopoiesis of indeterminate potential are also associated with a decreased risk of developing *TET2*-mutant clonal haematopoiesis of indeterminate potential[129,132].

Germline modifiers of somatic mutations

Germline variants, especially those regulating inflammatory pathways, can modify disease risks or treatment effects associated with somatic mutations. For example, mice with *Tet2*-mutant clonal haematopoiesis develop larger atherosclerotic burdens than mice under normal conditions, and prior work demonstrated marked blunting of *Tet2*-mutant clonal haematopoiesis's atherogenic effect upon chemically abrogating interleukin (IL)-1B secretion[107]. This finding led subsequent work to test and confirm in humans that IL-6 pathway inhibition (a downstream event of IL-1B secretion inhibition), proxied by an *IL6R*-disruptive coding mutation, substantially modifies the clonal haematopoiesis-associated cardiovascular disease risk[133]. Another study extended these findings to another driver mutation of clonal haematopoiesis, *Jak2*^{V617F}. The authors demonstrated that atherogenic mice with *Jak2*^{V617F} clonal haematopoiesis had increased atherosclerosis, which was then reduced in the presence of *Aim2* deficiency (induced through *Aim2* knockout bone marrow transplantation)[134]. More recently, additional human genetics findings have validated the *JAK2–AIM2* interaction in humans and shown that germline genetically determined expression levels of several other genes can selectively modify the associations between specific driver genes of clonal haematopoiesis and cardiovascular disease risk[135] (Fig. 5c).

Diseases linked to both somatic and germline variants

Many cancers, spanning from haematological malignancies to solid tumours, owe their pathogenesis to the complex interplay between germline and somatic variants. Although cancer typically results from the accumulation of somatic variants, germline variants can predispose individuals to developing cancer, both directly, such as the case of *BRCA1* and *BRCA2* genes increasing the risk of breast cancer[10,136], and indirectly, by increasing the risk of developing somatic mutations, as discussed above. A recent example includes germline variants in *TP53*, which can cause Li–Fraumeni syndrome, a rare genetic disorder that predisposes individuals to multiple cancers[137]. A recent study across 14 cancers identified a highly polygenic architecture, involving germline variants at thousands of loci, and suggested that polygenic risk prediction has potential for patient stratification[138].

Numerous non-cancer diseases, ranging from developmental diseases in early life to geriatric conditions, are shaped by the combination of somatic and germline variants. For instance, endometriosis, a condition exhibiting nearly 50% heritability based on family studies[139], is also associated with somatic mutations in *ARID1A*, *PIK3CA* and *KRAS*[140] (Fig. 5d). Additionally, germline variants in *STAT3* have been found in rheumatoid arthritis cases, and somatic variants in *STAT3* are common in T cells from patients with Felty's syndrome, a complication of rheumatoid arthritis[141]. Furthermore, rare genetic disorders such as autoimmune lymphoproliferative syndrome are associated with both germline and somatic mutations in *FAS*[142]. The observation that clonal haematopoiesis is linked to atherosclerosis[14,107,133] subsequently led to clonal haematopoiesis being associated with many other non-cancer, heritable conditions, such as chronic liver diseases and neurodegenerative disorders[15,40,143,144]. Finally, a recent study found that recurrent non-missense somatic mutations in blood cells are individually not oncogenic, but they are associated with blood cell traits, such as altered monocyte counts comparable to those of Mendelian variants in *RASGRP1* and *ELANE*, that cause severe congenital neutropenia[145]. These mutations are not readily explained by other clonal phenomena and seem to have a germline genetic basis related to adaptive immune function, pro-inflammatory cytokine production and lymphoid lineage,

highlighting the complex interplay between germline and somatic variation patterns and disease risk[57].

Conclusions and future directions

Our understanding of the contribution of both somatic and germline mutations to human diseases has substantially progressed in recent years. This progress has been applied to several therapeutic treatments for cancer and to increasing prevention and treatment options for largely monogenic non-cancer diseases, such as sickle cell anaemia, for which gene therapies were recently approved by the FDA[146].

Continued progress, particularly towards precision prevention and treatment, will necessitate integrating high-depth multi-omics data, as well as information on social determinants of health, lifestyle factors, health status and the environment. Given the dynamic nature of somatic variants, longitudinal analyses will also yield a more nuanced understanding of mutation dynamics over time[108,147]. Additionally, diverse representation in human genetic studies examining both germline and somatic variants will be necessary for a comprehensive understanding of the mutational landscape and to ensure the research findings and their subsequent applications are equitable and beneficial across populations. Finally, studies of somatic variants in human tissues or cell types beyond blood are currently sparse, but they will yield important new observations, especially on patterns of somatic mosaicism in normal tissues and on the role of somatic mutations in human disease. Notably, the recently established NIH Somatic Mosaicism across Human Tissues (SMaHT) Consortium, which intends to comprehensively study all classes of somatic variants by short-read and long-read sequencing across human tissues, is a pivotal step in this direction[148].

Future directions for technical progress include leveraging methodological advancements from studies of germline variants towards a better understanding of somatic mosaicism. For example, studies of the causal relationship between somatic variants and diseases can extend beyond current mice experiments into human genetics by utilizing causal inference methods from epidemiological and statistical domains[149]. Additionally, in situ omics technologies remain underutilized outside of neurobiology[150,151] and could thus become invaluable tools. For example, omics technologies could help track the evolution of somatic variants within an individual and dissect the molecular consequences over time. Finally, comprehensive reference databases and catalogues need to be constructed to account for the accelerating pace of research on somatic mutations. Building a robust and accessible database would facilitate cross-disciplinary studies and accelerate the translation of somatic variation research into biological and clinical applications.

As research expands to include more diverse and longitudinal human studies, as well as more single-cell multi-omics and causal inference methods, new insights will continue to emerge on the similarities, differences and interplay of germline and somatic variations. These insights can, in combination with large-scale collaborative efforts to facilitate translation, lead to deeper biological understanding, therapeutic innovations and clinical care applications.

Published online: 28 March 2024

References

1. Lynch, M. et al. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**, 704–714 (2016).
2. Coorens, T. H. H. et al. Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**, 387–392 (2021).
In this study, clones from many different normal tissues are sequenced, and phylogenetic trees of these normal cells are reconstructed, revealing embryonic lineages and somatic evolution.
3. Bizzotto, S. et al. Landmarks of human embryonic development inscribed in somatic mutations. *Science* **371**, 1249–1253 (2021).
4. Spencer Chapman, M. et al. Lineage tracing of human development through somatic mutations. *Nature* **595**, 85–90 (2021).
5. Fasching, L. et al. Early developmental asymmetries in cell lineage trees in living individuals. *Science* **371**, 1245–1248 (2021).
6. Park, S. et al. Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature* **597**, 393–397 (2021).
7. Bates, G. P. History of genetic disease: the molecular genetics of Huntington disease — a history. *Nat. Rev. Genet.* **6**, 766–773 (2005).
8. Berberich, A. J. & Hegele, R. A. The complex molecular genetics of familial hypercholesterolaemia. *Nat. Rev. Cardiol.* **16**, 9–20 (2019).
9. Wooster, R. et al. Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* **378**, 789–792 (1995).
10. Miki, Y. et al. A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science* **266**, 66–71 (1994).
11. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
12. Mustjoki, S. & Young, N. S. Somatic mutations in "benign" disease. *N. Engl. J. Med.* **384**, 2039–2052 (2021).
13. Miller, M. B. et al. Somatic genomic changes in single Alzheimer's disease neurons. *Nature* **604**, 714–722 (2022).
14. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
15. Wong, W. J. et al. Clonal haematopoiesis and risk of chronic liver disease. *Nature* **616**, 747–754 (2023).
16. Niroula, A. et al. Distinction of lymphoid and myeloid clonal hematopoiesis. *Nat. Med.* **27**, 1921–1927 (2021).
17. Silver, A. J., Bick, A. G. & Savona, M. R. Germline risk of clonal haematopoiesis. *Nat. Rev. Genet.* **22**, 603–617 (2021).
18. Robinson, P. S. et al. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat. Genet.* **53**, 1434–1442 (2021).
19. Lee, B. C. H. et al. Mutational landscape of normal epithelial cells in Lynch Syndrome patients. *Nat. Commun.* **13**, 2710 (2022).
20. Robinson, P. S. et al. Inherited *MUTYH* mutations cause elevated somatic mutation rates and distinctive mutational signatures in normal human cells. *Nat. Commun.* **13**, 3949 (2022).
21. Kazazian, H. H. Jr Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632 (2004).
22. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
23. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
24. Macintyre, G. et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.* **50**, 1262–1270 (2018).
25. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
26. Coorens, T. H. H. et al. Inherent mosaicism and extensive mutation of human placentas. *Nature* **592**, 80–85 (2021).
27. Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
28. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
29. Lawson, A. R. J. et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020).
This study is one of the first to use organoid cultures of stem cells from different human tissues to study somatic mutations in normal cells by whole-genome sequencing.
30. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
31. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
32. Mitchell, E. et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
33. Wang, Y. et al. APOBEC mutagenesis is a common process in normal human small intestine. *Nat. Genet.* **55**, 246–254 (2023).
34. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
35. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
36. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
37. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
This study identifies large clonal expansions carrying driver mutations in normal skin.
38. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
39. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
This is a landmark study demonstrating population-level associations of somatic mutation with both cancer and non-cancer health conditions.

40. Zekavat, S. M. et al. Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat. Med.* **27**, 1012–1024 (2021).
41. Colom, B. et al. Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. *Nat. Genet.* **52**, 604–614 (2020).
42. Abby, E. et al. *Notch1* mutations drive clonal expansion in normal esophageal epithelium but impair tumor growth. *Nat. Genet.* **55**, 232–245 (2023).
43. Ng, S. W. K. et al. Convergent somatic mutations in metabolism genes in chronic liver disease. *Nature* **598**, 473–478 (2021).
This study identifies selection for recurrent somatic mutations as an adaptive mechanism to chronic liver disease.
44. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
45. Jónsson, H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
46. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
47. Kaplanis, J. et al. Genetic and chemotherapeutic influences on germline hypermutation. *Nature* **605**, 503–508 (2022).
This study, based on whole-genome data of over 20,000 families, identifies accelerated rates of de novo germline mutations and determines the likely causes of this hypermutation.
48. Maher, G. J. et al. Visualizing the origins of selfish de novo mutations in individual seminiferous tubules of human testes. *Proc. Natl Acad. Sci. USA* **113**, 2454–2459 (2016).
49. Goriely, A., McGrath, J. J., Hultman, C. M., Wilkie, A. O. M. & Malaspina, D. 'Selfish spermatogonial selection': a novel mechanism for the association between advanced paternal age and neurodevelopmental disorders. *Am. J. Psychiatry* **170**, 599–608 (2013).
50. Goriely, A., McVean, G. A. T., Röjmyr, M., Ingemarsson, B. & Wilkie, A. O. M. Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science* **301**, 643–646 (2003).
51. Pena, S. D. J. Advances of aneuploidy research in the maternal germline. *Nat. Rev. Genet.* **24**, 274 (2023).
52. Champion, K. J. et al. Germline mutation in *BRAF* codon 600 is compatible with human development: de novo p.V600G mutation identified in a patient with CFC syndrome. *Clin. Genet.* **79**, 468–474 (2011).
53. Sabeti, P. C. et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
54. Olson, N. D. et al. Variant calling and benchmarking in an era of complete human genome sequences. *Nat. Rev. Genet.* **24**, 464–483 (2023).
55. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
56. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
57. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
58. Grigoriadis, K. et al. CONIPHER: a computational framework for scalable phylogenetic reconstruction with error correction. *Nat. Protoc.* **19**, 159–183 (2024).
59. Luquette, L. J. et al. Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. *Nat. Genet.* **54**, 1564–1571 (2022).
60. Williams, N. et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162–168 (2022).
61. Ellis, P. et al. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021).
62. Bae, J. H. et al. Single duplex DNA sequencing with CODEC detects mutations with high sensitivity. *Nat. Genet.* **55**, 871–879 (2023).
63. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
64. Benjamin, D. et al. Calling somatic SNVs and indels with Mutect2. Preprint at *bioRxiv* 861054 https://doi.org/10.1101/861054 (2019).
65. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
66. Jones, D. et al. cgpCaVEManwrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
67. Yang, X. et al. Control-independent mosaic single nucleotide variant detection with DeepMosaic. *Nat. Biotechnol.* **41**, 870–877 (2023).
68. Zhou, W. et al. Global Biobank Meta-analysis Initiative: powering genetic discovery across human disease. *Cell Genom.* **2**, 100192 (2022).
69. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat. Genet.* **55**, 1243–1249 (2023).
70. Rubinacci, S., Hofmeister, R. J., Sousa da Mota, B. & Delaneau, O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nat. Genet.* **55**, 1088–1090 (2023).
71. Weiner, D. J. et al. Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614**, 492–499 (2023).
72. Fiziev, P. P. et al. Rare penetrant mutations confer severe risk of common diseases. *Science* **380**, eabo1131 (2023).
73. Hujoel, M. L. A. et al. Influences of rare copy-number variation on human complex traits. *Cell* **185**, 4233–4248.e27 (2022).
74. Mukamel, R. E. et al. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**, 1499–1505 (2021).
75. Li, Z. et al. A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nat. Methods* **19**, 1599–1611 (2022).
76. Selvaraj, M. S. et al. Whole genome sequence analysis of blood lipid levels in >66,000 individuals. *Nat. Commun.* **13**, 5995 (2022).
77. Natarajan, P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun.* **9**, 3391 (2018).
78. Li, X. et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**, 969–983 (2020).
This study introduces the STAAR series, which exemplifies multiple aspects of advancements in germline association studies: multi-ancestry study population, rare variants, multiple functional annotations and novel methods.
79. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
80. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
81. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
82. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
83. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
84. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
85. Ståhl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
86. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
87. Haniffa, M. et al. A roadmap for the human developmental cell atlas. *Nature* **597**, 196–205 (2021).
88. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
89. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
90. Zhang, M. J. et al. Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat. Genet.* **54**, 1572–1580 (2022).
91. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
92. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B* **82**, 1273–1300 (2020).
93. Gazal, S. et al. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).
94. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
95. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
This study showcases how leveraging deep learning advancement can improve our understanding of genomic biology.
96. Li, X. et al. Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nat. Genet.* **55**, 154–164 (2023).
97. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
98. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
99. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
100. Klarin, D. & Natarajan, P. Clinical utility of polygenic risk scores for coronary artery disease. *Nat. Rev. Cardiol.* **19**, 291–301 (2022).
101. Patel, A. P. et al. A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat. Med.* **29**, 1793–1803 (2023).
102. Weir, B. S., Anderson, A. D. & Hepler, A. B. Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* **7**, 771–780 (2006).
103. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**, 477–485 (2008).
104. Lawson, D. J. et al. Is population structure in the genetic biobank era irrelevant, a challenge, or an opportunity? *Hum. Genet.* **139**, 23–41 (2020).
105. Jyoti, G., Dayal, M. S. & Arguello, A. Developmental genotype-tissue expression (dGTEx). *National Human Genome Research Institute* https://www.genome.gov/Funded-Programs-Projects/Developmental-Genotype-Tissue-Expression (2020).

106. Heyde, A. et al. Increased stem cell proliferation in atherosclerosis accelerates clonal hematopoiesis. *Cell* **184**, 1348–1361.e22 (2021).
107. Fuster, J. J. et al. Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. *Science* **355**, 842–847 (2017).
108. Fabre, M. A. et al. The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* **606**, 335–342 (2022).
109. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
110. Muyas, F. et al. De novo detection of somatic mutations in high-throughput single-cell profiling data sets. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01863-z (2023).
111. Miles, L. A. et al. Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature* **587**, 477–482 (2020).
112. Nam, A. S., Chaligne, R. & Landau, D. A. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat. Rev. Genet.* **22**, 3–18 (2021).
113. Uddin, M. D. M. et al. Clonal hematopoiesis of indeterminate potential, DNA methylation, and risk for coronary artery disease. *Nat. Commun.* **13**, 5350 (2022).
114. Izzo, F. et al. DNA methylation disruption reshapes the hematopoietic differentiation landscape. *Nat. Genet.* **52**, 378–387 (2020).
115. Gumuser, E. D. et al. Clonal hematopoiesis of indeterminate potential predicts adverse outcomes in patients with atherosclerotic cardiovascular disease. *J. Am. Coll. Cardiol.* **81**, 1996–2009 (2023).
116. Schratz, K. E. et al. Somatic reversion impacts myelodysplastic syndromes and acute myeloid leukemia evolution in the short telomere disorders. *J. Clin. Investig.* **131**, e147598 (2021).
117. Revy, P., Kannengiesser, C. & Fischer, A. Somatic genetic rescue in Mendelian haematopoietic diseases. *Nat. Rev. Genet.* **20**, 582–598 (2019).
118. Banda, K., Swisher, E. M., Wu, D., Pritchard, C. C. & Gadi, V. K. Somatic reversion of germline *BRCA2* mutation confers resistance to poly(ADP-ribose) polymerase inhibitor therapy. *JCO Precis. Oncol.* **2**, 1–6 (2018).
119. Ashworth, A. Drug resistance caused by reversion mutation. *Cancer Res.* **68**, 10021–10023 (2008).
120. Sakai, W. et al. Secondary mutations as a mechanism of cisplatin resistance in *BRCA2*-mutated cancers. *Nature* **451**, 1116–1120 (2008).
121. Saha, K. et al. The NIH somatic cell genome editing program. *Nature* **592**, 195–204 (2021).
122. Biswas, P. & Verma, R. S. Somatic mosaicism in inherited bone marrow failure and chromosomal instability syndrome. *Genome Instab. Dis.* **2**, 150–163 (2021).
123. Sebert, M. et al. Clonal hematopoiesis driven by chromosome 1q/MDM4 trisomy defines a canonical route toward leukemia in Fanconi anemia. *Cell Stem Cell* **30**, 153–170.e9 (2023).
124. Steinberg, G. D., Carter, B. S., Beaty, T. H., Childs, B. & Walsh, P. C. Family history and the risk of prostate cancer. *Prostate* **17**, 337–347 (1990).
125. DeBoy, E. A. et al. Familial clonal hematopoiesis in a long telomere syndrome. *N. Engl. J. Med.* **388**, 2422–2433 (2023).
126. McNally, E. J., Luncsford, P. J. & Armanios, M. Long telomeres and cancer risk: the price of cellular immortality. *J. Clin. Investig.* **129**, 3474–3481 (2019).
127. Franch-Expósito, S. et al. Associations between cancer predisposition mutations and clonal hematopoiesis in patients with solid tumors. *JCO Precis. Oncol.* **7**, e2300070 (2023).
128. Bick, A. G. et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).
This is a landmark study examining the germline genetic basis of one type of somatic mutation using population-level data.
129. Uddin, M. M. et al. Germline genomic and phenomic landscape of clonal hematopoiesis in 323,112 individuals. Preprint at *medRxiv* https://doi.org/10.1101/2022.07.29.22278015 (2022).
130. Kessler, M. D. et al. Common and rare variant associations with clonal haematopoiesis phenotypes. *Nature* **612**, 301–309 (2022).
131. Liu, A. et al. Population analyses of mosaic X chromosome loss identify genetic drivers and widespread signatures of cellular selection. Preprint at *medRxiv* https://doi.org/10.1101/2023.01.28.23285140 (2023).
132. Weinstock, J. S. et al. Aberrant activation of TCL1A promotes stem cell expansion in clonal haematopoiesis. *Nature* **616**, 755–763 (2023).
133. Bick, A. G. et al. Genetic interleukin 6 signaling deficiency attenuates cardiovascular risk in clonal hematopoiesis. *Circulation* **141**, 124–131 (2020).
134. Fidler, T. P. et al. The AIM2 inflammasome exacerbates atherosclerosis in clonal haematopoiesis. *Nature* **592**, 296–301 (2021).
135. Yu, Z. et al. Genetic modification of inflammation and clonal hematopoiesis-associated cardiovascular risk. *J. Clin. Investig.* **133**, e168597 (2023).
136. Hall, J. M. et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–1689 (1990).
137. Pareja, F. et al. Cancer-causative mutations occurring in early embryogenesis. *Cancer Discov.* **12**, 949–957 (2022).
138. Zhang, Y. D. et al. Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nat. Commun.* **11**, 3353 (2020).
139. Saha, R. et al. Heritability of endometriosis. *Fertil. Steril.* **104**, 947–952 (2015).
140. Anglesio, M. S. et al. Cancer-associated mutations in endometriosis without cancer. *N. Engl. J. Med.* **376**, 1835–1848 (2017).
141. Savola, P. et al. Somatic mutations in clonally expanded cytotoxic T lymphocytes in patients with newly diagnosed rheumatoid arthritis. *Nat. Commun.* **8**, 15869 (2017).
142. Magerus, A., Bercher-Brayer, C. & Rieux-Laucat, F. The genetic landscape of the FAS pathway deficiencies. *Biomed. J.* **44**, 388–399 (2021).
143. Bouzid, H. et al. Clonal hematopoiesis is associated with protection from Alzheimer's disease. *Nat. Med.* **29**, 1662–1670 (2023).
144. Weeks, L. D. et al. Age-related diseases of inflammation in myelodysplastic syndrome and chronic myelomonocytic leukemia. *Blood* **139**, 1246–1250 (2022).
145. Weinstock, J. S. et al. The genetic determinants of recurrent somatic mutations in 43,693 blood genomes. *Sci. Adv.* **9**, eabm4945 (2023).
146. Office of the Commissioner. FDA approves first gene therapies to treat patients with sickle cell disease. *U.S. Food and Drug Administration* https://www.fda.gov/news-events/press-announcements/fda-approves-first-gene-therapies-treat-patients-sickle-cell-disease (2023).
147. Robertson, N. A. et al. Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects. *Nat. Med.* **28**, 1439–1446 (2022).
148. National Institutes of Health. Somatic Mosaicism across Human Tissues (SMaHT). *NIH* https://commonfund.nih.gov/smaht (2021).
149. Hernan, M. A. & Robins, J. M. *Causal Inference: What If* 1st edn (Taylor & Francis Group, 2023).
150. Zeng, H. et al. Integrative in situ mapping of single-cell transcriptional states and tissue histopathology in a mouse model of Alzheimer's disease. *Nat. Neurosci.* **26**, 430–446 (2023).
151. Zeng, H. et al. Spatially resolved single-cell translatomics at molecular resolution. *Science* **380**, eadd3067 (2023).
152. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
153. Yu, Z. et al. Polygenic risk scores for kidney function and their associations with circulating proteome, and incident kidney diseases. *J. Am. Soc. Nephrol.* **32**, 3161–3173 (2021).
154. Sondka, Z. et al. COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res.* **52**, D1210–D1217 (2024).
155. Bergström, A. et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
156. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018).
157. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat. Protoc.* **17**, 1518–1552 (2022).
158. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
159. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
160. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
161. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
162. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
163. Espina, V. et al. Laser-capture microdissection. *Nat. Protoc.* **1**, 586–603 (2006).
164. Lan, F., Demaree, B., Ahmed, N. & Abate, A. R. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol.* **35**, 640–646 (2017).
165. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
166. Emmert-Buck, M. R. et al. Laser capture microdissection. *Science* **274**, 998–1001 (1996).
167. Bonner, R. F. et al. Laser capture microdissection: molecular analysis of tissue. *Science* **278**, 1481–1483 (1997).
168. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
169. Payne, A. C. et al. In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science* **371**, eaay3446 (2021).
170. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
171. Jagadeesh, K. A. et al. Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. *Nat. Genet.* **54**, 1479–1492 (2022).
172. Lomakin, A. et al. Spatial genomics maps the structure, nature and evolution of cancer clones. *Nature* **611**, 594–602 (2022).

Author contributions

T.H.H.C. and Y.Z. researched the literature. T.H.H.C., P.N. and Y.Z. contributed substantially to discussion of the content. T.H.H.C., M.M.U. and Y.Z. wrote the article. All authors reviewed and/or edited the manuscript before submission.

Competing interests

P.N. reports investigator-initiated grants from Amgen, Apple, Boston Scientific, Novartis and AstraZeneca; personal fees from Allelica, Apple, AstraZeneca, Blackstone Life Sciences, Foresite Labs, Genentech and Novartis; scientific board membership for Esperion Therapeutics, geneXwell and TenSixteen Bio; and spousal employment at Vertex, all unrelated

to the present work. P.N. is a scientific co-founder of TenSixteen Bio, which is a company focused on clonal haematopoiesis but had no role in the present work. The other authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Genetics* thanks Jan O. Korbel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Related links

COSMIC: https://cancer.sanger.ac.uk/cosmic