

The Missing Somatic Genome: Mosaic Variation in Repetitive and Structurally Complex DNA Across Mammalian Tissues

Birupaksha Biswas

drbiswasassted.medicalglory@gmail.com

Independent Researcher

Systematic Review

Keywords: long-read sequencing, mammalian genomics, repetitive DNA, somatic mosaicism, structural variation, telomere-to-telomere assembly

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11042204/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Somatic mosaicism is increasingly recognised as a fundamental property of mammalian tissues, yet its genomic architecture has been reconstructed predominantly from regions that can be represented accurately by reference genomes and interrogated reliably by short-read sequencing. This creates a potentially important ascertainment problem. Centromeric and pericentromeric satellites, acrocentric short arms and ribosomal DNA arrays, tandem repeats, segmental duplications, mobile elements, telomeric and subtelomeric regions, and other structurally complex loci have historically been incompletely assembled, ambiguously mapped, or excluded from conventional somatic-variant analyses. Telomere-to-telomere assemblies, pangenomes, long-read sequencing, single-molecule methods and increasingly powerful single-cell approaches are now making substantial parts of this previously inaccessible sequence experimentally tractable. This Review proposes the “Missing Somatic Genome Hypothesis”: repetitive and structurally complex genomic compartments contain a quantitatively and functionally important burden of postzygotic variation that has been systematically underestimated. Evidence from tandem-repeat instability, mobile-element mobilisation, centromeric diversity, segmental duplications, ribosomal DNA biology and tissue-specific structural mosaicism supports the biological plausibility of this framework, although direct body-wide measurements remain limited. The Review distinguishes established evidence from inference and defines experimental strategies capable of determining whether incorporation of the previously inaccessible genome materially changes current understanding of mammalian somatic mutation, tissue evolution, ageing and disease.

Introduction

The mammalian body is not genomically uniform. From the first postzygotic divisions onwards, cells acquire substitutions, insertions, deletions, copy-number alterations, structural rearrangements, changes in repetitive sequence and mobile-element insertions. These events generate a branching genomic history within a single organism. Some mutations remain confined to one cell, some mark developmental lineages, some undergo neutral drift, and others alter cellular fitness sufficiently to produce clonal expansion or disease. Somatic mosaicism has therefore moved from being regarded as an exceptional state in classical genetics to a central principle of mammalian genome biology.

Large-scale studies of histologically normal tissues have transformed the quantitative understanding of this process. Single-molecule sequencing has revealed mutation landscapes at extremely low variant frequencies, multi-organ studies have mapped somatic mutagenesis across morphologically normal human tissues, and comparative studies across mammals have shown that somatic mutation rates vary systematically with lifespan (Abascal et al. 2021; Cagan et al. 2022; Li et al. 2021). Earlier tissue-comparison studies also established that differentiated human tissues can contain mosaic copy-number changes, demonstrating that structural variation is not restricted to constitutional genomes or neoplasia (Piotrowski et al. 2008). More recent pan-tissue analyses have shown age-associated mosaic chromosomal alterations across multiple normal organs, although large somatic copy-number changes remain less common than substitutions in many tissues (Gao et al. 2023). The Somatic Mosaicism

across Human Tissues Network is now developing a multi-platform framework combining deep short-read sequencing, long-read sequencing, single-cell approaches and donor-specific genome reconstruction to interrogate mosaicism more comprehensively (Coorens et al. 2025).

A conceptual asymmetry nevertheless remains. Most foundational maps of the somatic genome were produced using technologies that perform best in genomic regions that can be assembled accurately, mapped uniquely and compared confidently with a linear reference sequence. That methodological constraint gradually became embedded in the biological description of somatic mosaicism. In practical terms, the “somatic genome” often came to mean the portion of somatic genomic variation that available sequencing and reference systems could recognise with acceptable specificity.

That equivalence is becoming increasingly difficult to defend. The telomere-to-telomere assembly of the human genome added nearly 200 Mb of sequence relative to previous reference representations and resolved extensive centromeric, acrocentric, duplicated and repetitive sequence that had historically remained absent or incomplete (Nurk et al. 2022). Complete genomic and epigenomic maps subsequently established tractable sequence and chromatin frameworks across human centromeres (Altemose et al. 2022). The human pangenome demonstrated that even a complete single reference cannot adequately represent population-level structural diversity, revealing substantial additional sequence and improving representation of complex loci (Liao et al. 2023). Haplotype-resolved long-read assemblies likewise showed that a large fraction of structural variants detectable by assembly had not been recovered by short-read sequencing (Ebert et al. 2021), while population-scale long-read sequencing has now resolved more than 100,000 sequence-defined structural variants and hundreds of thousands of multiallelic VNTRs across diverse human genomes (Schloissnig et al. 2025).

The transition from isolated complete genomes to diverse near-complete genomes has further sharpened the problem. Analysis of 65 individuals closed most previous assembly gaps, resolved thousands of complex structural variants and completely assembled more than one thousand human centromeres, demonstrating extensive structural diversity in precisely those regions most poorly represented historically (Logsdon et al. 2025). Complete centromere comparisons have exposed striking sequence and size diversity, elevated nucleotide variation and major differences in higher-order repeat architecture (Logsdon et al. 2024). Most recently, population-scale analysis of thousands of complete centromeres has revealed hundreds of centromere haplotypes, thousands of alpha-satellite higher-order repeat variants and marked variation in centromeric mutation rates (Gao et al. 2026).

These developments create a question distinct from the established observation that difficult genomic regions vary among individuals. The more fundamental question is whether those same genomic compartments also accumulate meaningful variation within an individual during development, ageing and disease, and whether contemporary estimates of mammalian somatic variation have been conditioned by the genomic territory that sequencing technology happened to make observable.

The term **missing somatic genome** is used here to describe somatic variation occurring within repetitive, duplicated, structurally complex or historically unresolved genomic sequence that has been

disproportionately inaccessible to conventional reference-dependent short-read analysis. The term does not imply that the relevant sequence remains absent from modern genome assemblies. Rather, it identifies a persistent measurement deficit in which biological observability depends strongly on genomic architecture.

No dedicated synthesis identified in the current literature has organised mammalian somatic mosaicism specifically around this genomic ascertainment problem. The objective of this Review is therefore not to argue that all repetitive DNA is unusually unstable, nor to suggest that established estimates of somatic mutation are invalid. Instead, it asks whether genomic regions that were historically hardest to measure may harbour forms of somatic variation sufficiently distinct or abundant to alter the biological picture once they are incorporated systematically.

Scope and review approach

This article is a hypothesis-driven narrative Review rather than a systematic review or meta-analysis. Relevant literature was evaluated through targeted searches of PubMed-indexed records, major biomedical and genomic literature sources, and reference chaining through 15 September 2026. Searches focused on mammalian somatic mosaicism, repetitive DNA, tandem-repeat instability, centromeric and acrocentric sequence, ribosomal DNA, segmental duplications, retrotransposition, telomere and subtelomere biology, structural variation, long-read sequencing, telomere-to-telomere assemblies, pangenomics and single-cell DNA sequencing. Particular emphasis was placed on primary studies that demonstrated a somatic biological event, established a genomic detection limitation, or introduced approaches capable of resolving previously inaccessible sequence. Human evidence is considered alongside mechanistic studies in mouse and other mammalian systems where these clarify causality or evolutionary conservation. AI-assisted tools supported drafting, manuscript organisation and language refinement; all sources, scientific interpretations and final content were reviewed by the author, who retains responsibility for the manuscript.

The resulting framework deliberately distinguishes several levels of evidence. Somatic tandem-repeat instability and LINE-1 mobilisation are directly demonstrated phenomena. Somatic copy-number mosaicism is also established in normal differentiated tissues, although its prevalence varies by tissue and detection method (Gao et al. 2023; Piotrowski et al. 2008). Normal-tissue somatic dynamics of centromeric satellites, ribosomal DNA arrays and some classes of segmental duplication remain biologically plausible but incompletely measured. A further category consists of predictions generated by the proposed framework itself. Preserving these distinctions is essential because the scientific value of the Missing Somatic Genome Hypothesis lies in its falsifiability rather than in presenting inference as observation.

What constitutes the missing somatic genome?

A genomic compartment can be biologically present yet analytically under-observed. The distinction is particularly important in somatic genomics because two forms of scarcity act simultaneously. A genomic locus may be difficult to resolve because its sequence is repetitive, duplicated or structurally variable, while the mutation of interest may also be present in only a small proportion of cells. The resulting problem is therefore not simply one of sequencing depth. It is the intersection of complex genomic architecture with low cellular prevalence.

The principal compartments considered within this framework are summarised in Table 1. Their molecular biology differs substantially, but they share a common analytical feature: confidence in detecting genuine somatic variation depends unusually strongly on read length, reference architecture, haplotype resolution, variant class and analytical strategy. Table 1 does not imply equivalent mutation rates or functional importance across these genomic compartments. Instead, it defines territories in which conventional estimates of somatic variation warrant renewed examination.

Table 1
Principal genomic compartments contributing to the proposed missing somatic genome

Genomic compartment	Biological architecture	Potential somatic variation	Principal ascertainment limitation	Enabling approaches
Centromeric and pericentromeric satellites	Megabase-scale higher-order repeats and heterochromatic sequence associated with kinetochore formation	Repeat expansion or contraction, sequence conversion, array restructuring and rearrangement	Extreme repetitiveness, historical reference gaps and poor unique mapping	Ultra-long reads, telomere-to-telomere assemblies and haplotype-resolved assembly
Acrocentric short arms and ribosomal DNA arrays	Highly homologous sequence containing extensive segmental duplication and tandem rDNA arrays	Copy-number change, recombination, array restructuring and lineage-specific rDNA states	Homology between chromosomes and inability to assign repeated units uniquely	Long-read assembly, optical mapping and orthogonal copy-number validation
Tandem repeats and variable-number tandem repeats	Repeated sequence ranging from short microsatellites to long disease-associated arrays	Somatic expansion, contraction and repeat-length mosaicism	Read-length limitation, amplification bias and alignment ambiguity	Long-read sequencing, targeted repeat analysis and single-cell genotyping
Segmental duplications	Large blocks of highly similar duplicated sequence enriched for structural variation	Gene conversion, deletion, duplication, complex rearrangement and nucleotide substitution	Paralogue misassignment and collapsed reference representation	Pangenomes, phased long-read assemblies and duplication-aware variant calling
Mobile elements	LINE, Alu, SVA and related repetitive elements with active or potentially active source copies	New insertion, deletion and element-associated rearrangement	New copies resemble numerous existing genomic copies and source loci may themselves be unresolved	Long-read sequencing, specialised insertion callers and single-cell approaches
Telomeric and subtelomeric regions	Terminal tandem repeats adjacent to duplication-rich and structurally polymorphic sequence	Telomere-length mosaicism, subtelomeric rearrangement and repeat restructuring	Extreme repeat content, terminal location and extensive structural polymorphism	Ultra-long reads, telomere-enrichment strategies and optical mapping

The missing somatic genome should therefore be understood as a problem of **differential genomic observability**. A mutation is not biologically absent merely because an analytical system cannot assign it confidently to a genomic coordinate. This distinction becomes particularly important when difficult sequence possesses molecular properties that may themselves favour recombination, replication slippage, gene conversion, replication stress or structural reorganisation.

The genomic streetlight problem

Modern somatic genomics has been successful partly because it adopted stringent restrictions on where variants could be called confidently. This was scientifically necessary. Low-frequency somatic mutations must be distinguished from sequencing errors, DNA damage, mapping artefacts, library errors and inherited variants. Analyses therefore concentrated naturally on genomic territory in which high specificity could be maintained. The strategy enabled major advances in developmental lineage reconstruction, mutation-rate measurement and mutational-signature analysis.

The consequence, however, resembles a genomic streetlight effect. A short sequencing read generated from unique euchromatic sequence usually has a well-defined genomic address. A read generated from a highly identical segmental duplication, satellite array or recently inserted mobile element may have several possible addresses. If the relevant haplotype is structurally different from the reference, or if copies are collapsed or absent from the reference assembly, the problem becomes more fundamental because the analytical model does not completely represent the biological substrate being tested.

Assembly-based studies provide direct evidence for the magnitude of this problem in constitutional variation. Haplotype-resolved sequencing of diverse human genomes identified more than 100,000 structural variants, with a substantial majority of variants in that study not recovered by short-read sequencing (Ebert et al. 2021). Subsequent studies involving nearly complete genomes and more than one thousand long-read sequenced individuals have further expanded the detectable repertoire of structural variants and multiallelic tandem repeats, including complex rearrangements concentrated in repeat-rich and duplicated sequence (Logsdon et al. 2025; Schloissnig et al. 2025). These are predominantly germline observations, but they establish that genomic architecture strongly determines what becomes observable.

The same principle is likely to become more severe for somatic variation because cellular prevalence adds a second constraint. Long-read whole-genome sequencing of individual human cells has already shown that structural variants and tandem-repeat states can be recovered in regions inaccessible to conventional short-read analysis, while simultaneously illustrating the difficulty of obtaining complete and unbiased single-cell coverage (Hård et al. 2023). A genomic compartment that is difficult to represent when a variant is present constitutionally in every cell becomes considerably more difficult when the same variant is carried by only a small cellular subclone.

The missing somatic genome therefore arises from a multiplication of two uncertainties rather than from either alone. Complex sequence reduces positional certainty, while low variant allele fraction

reduces statistical signal. A genuine event may consequently disappear not because sequencing produced no information, but because available analytical frameworks cannot distinguish that information confidently from genomic ambiguity.

This distinction also complicates comparisons of somatic mutation burden. Mutation rates are generally calculated over a callable genome, and the boundaries of that callable genome are themselves technology dependent. If excluded sequence has mutation dynamics similar to included sequence, the biological consequences may be modest. If excluded sequence has qualitatively different modes of instability, conventional extrapolation may be incomplete. The magnitude of that difference remains inadequately quantified.

Centromeres and pericentromeric satellites

Centromeres represent perhaps the clearest historical mismatch between biological importance and genomic accessibility. Human centromeres are dominated by long alpha-satellite higher-order repeat arrays that remained incompletely represented until telomere-to-telomere assemblies became available. Complete genomic and epigenetic mapping demonstrated for the first time the sequence organisation, methylation patterns and kinetochore-associated domains of centromeres across an uninterrupted human reference genome (Altemose et al. 2022). Comparative analysis of complete human centromeres subsequently demonstrated striking differences in array size and sequence architecture, with nucleotide variation markedly elevated relative to unique flanking sequence and large portions of centromeric DNA resisting conventional alignment because of newly evolved higher-order repeat structures (Logsdon et al. 2024).

More recent population-scale work considerably strengthens the biological premise. Nearly complete assemblies from diverse individuals have revealed extensive differences in alpha-satellite array length, structural organisation and kinetochore-associated epigenetic architecture (Logsdon et al. 2025). Analysis of thousands of complete centromeres across global populations has now identified hundreds of major centromere haplotypes, thousands of higher-order repeat variants and substantial chromosome-to-chromosome variation in estimated mutation rate, with the kinetochore-associated region emerging as a particularly rapidly evolving domain (Gao et al. 2026).

Importantly, direct pedigree analysis provides evidence that mutation within previously inaccessible sequence is not merely an evolutionary abstraction. A four-generation, 28-member pedigree reconstructed using complementary sequencing technologies estimated dozens of tandem-repeat-associated de novo indels or structural variants per transmission and approximately four centromeric de novo mutations per transmission (Porubsky et al. 2025). These are germline transmission events rather than somatic mutations, but they demonstrate directly that centromeric and repeat-rich sequence contributes measurable mutation that is missed when mutation-rate studies are restricted to conventionally callable sequence.

None of these observations establishes that normal mammalian tissues acquire centromeric mutations at comparably high rates during life. That distinction is crucial. Their significance is that centromeric DNA is neither uniform nor genetically static and that the molecular machinery required to generate sequence, repeat-length and structural variation is demonstrably active in these regions.

A somatic centromeric change could theoretically involve repeat-array expansion or contraction, local sequence conversion, rearrangement at repeat boundaries, altered higher-order repeat composition or structural changes affecting the surrounding heterochromatic environment. Such events need not produce catastrophic chromosome segregation. More subtle sequence or array-length alterations could conceivably modify replication dynamics, centromeric chromatin, kinetochore positioning or susceptibility to subsequent chromosome instability without immediately generating an overt phenotype.

The evidence required to establish such effects in normal tissues remains limited. That absence is nevertheless difficult to interpret because centromeres have frequently been omitted from the searchable genome. Future studies will therefore need to distinguish between two fundamentally different explanations: either centromeric somatic variation is genuinely uncommon and tightly constrained, or it has remained systematically under-observed because the relevant substrate was analytically inaccessible.

Acrocentric chromosomes and ribosomal DNA arrays

The short arms of human acrocentric chromosomes contain extensive regions of homology, segmental duplications and tandem arrays encoding ribosomal RNA. Complete genomic assemblies have overturned the previous impression that these regions are merely repetitive appendages by demonstrating extensive structural complexity and ongoing recombination between heterologous acrocentric chromosomes (Guarracino et al. 2023).

Ribosomal DNA is particularly important because it provides an example of repetitive sequence with direct cellular function. Mammalian genomes contain numerous tandem copies of the genes encoding the major ribosomal RNAs, and rDNA copy number differs substantially among individuals. Human studies have associated rDNA dosage with variation in gene expression and mitochondrial abundance, while comparative mammalian work has linked rDNA copy number with body mass and growth-related phenotypes (Gibbons et al. 2014; Law et al. 2024).

Recent work provides a more direct view of within-individual biology. Analysis of healthy individuals across age groups found marked interindividual rDNA copy-number variation but no simple age-dependent loss of absolute rDNA copies in blood, while promoter methylation shifted with age, indicating that functional rDNA dosage can change without alteration of total array copy number (Geisen et al. 2025). More importantly for the present hypothesis, direct analysis of multiple normal tissues from the same individuals has demonstrated that absolute and hypomethylated active rDNA copy numbers

are generally constrained within an individual but can nevertheless show tissue-specific variation and occasional relaxation of this apparent homeostasis (Durackova et al. 2025).

These observations materially refine the missing-genome argument. They do not support an assumption of pervasive rDNA copy-number instability in every normal tissue. Instead, they suggest that rDNA arrays may be under active biological constraint while retaining a measurable capacity for intra-individual variation. Such a pattern is arguably more interesting than unrestricted instability because it raises the possibility of tissue-specific selection or dosage compensation within a repetitive genomic compartment.

The biological consequences could be substantial. rDNA arrays contribute directly to nucleolar organisation and ribosome production, and their activity is regulated not only by copy number but also by epigenetic state. A stable lineage-specific change in array structure, active-copy number or methylation could influence nucleolar activity, proteostasis, cellular growth or stress responses without generating a conventional protein-coding mutation.

The principal challenge remains measurement. Apparent differences in rDNA abundance between tissues can arise from genuine somatic change, constitutional haplotype differences, tissue composition, sequencing coverage or analytical bias. Demonstrating authentic mosaicism will therefore require donor-specific constitutional references, multiple tissues from the same organism, orthogonal copy-number measurement and, ideally, direct reconstruction of array architecture using long molecules. The critical biological question is no longer simply whether rDNA differs between individuals, but whether mammalian cell lineages acquire stable genomic or epigenomic rDNA states during life and whether those states participate in tissue-specific selection.

Tandem repeats as proof of principle

Tandem repeats provide the strongest proof that important classes of mammalian somatic variation extend beyond conventional substitutions and short indels. Somatic repeat instability is firmly established in repeat-expansion disorders, with Huntington disease providing the most extensively characterised example. An inherited expanded CAG tract within *HTT* can undergo further expansion after development, and this process displays marked cell-type specificity. Human studies have demonstrated differential CAG expansion among vulnerable neuronal populations, while genetic analyses have implicated mismatch-repair pathways as major modifiers of both somatic expansion and clinical phenotype (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium 2025; Mätlik et al. 2024).

Mouse models established the mechanistic basis of this phenomenon well before modern long-read genomics. Loss of *Msh2* suppresses somatic instability of expanded CAG tracts in Huntington disease models, and subsequent in vivo CRISPR studies have identified additional DNA-repair pathways capable of modifying somatic repeat behaviour (Manley et al. 1999; Mouro Pinto et al. 2025). Together, the human and mouse evidence establishes that repetitive sequence can continue mutating within

differentiated tissues, that the process depends on cellular context, and that a somatic mutation can be a quantitative change in repeat length rather than a conventional binary sequence variant.

The importance of this principle expanded substantially in 2026. Analysis of more than 900,000 participants from the UK Biobank and the All of Us Research Program identified marked variation in repeat instability and demonstrated age-associated blood mosaicism at common repeat loci including *TCF4* and *ADGRE2* (Hujoel et al. 2026). This finding shifts somatic repeat instability conceptually beyond the boundaries of rare Mendelian expansion disorders and raises the possibility that repeat-length mosaicism may represent a more general feature of mammalian ageing and tissue biology.

Complete pedigree sequencing reinforces this broader interpretation. When difficult repeat-rich sequence was incorporated into de novo mutation analysis, tandem-repeat-associated indels and structural variants accounted for a substantial component of newly detected variation per transmission (Porubsky et al. 2025). Although germline and somatic mutational processes cannot be assumed to be equivalent, this result demonstrates quantitatively how mutation counts change when previously difficult repeat classes enter the callable genome.

The unanswered question is how far these observations extend across tissues and across the repetitive genome. Blood offers accessibility and enormous statistical power, but other organs differ in cell turnover, replication history, mismatch-repair activity and selective pressures. A locus that appears stable in blood may behave differently in neurons, epithelial stem cells or long-lived quiescent cells. Somatic repeat dynamics could ultimately function as disease mechanisms, modifiers of cellular fitness, neutral lineage markers or molecular records of cellular history depending on locus and tissue.

This example also demonstrates why the missing somatic genome cannot be reduced to an argument about sequencing technology. The biological mechanism preceded the technology capable of measuring it comprehensively. Improved sequencing now creates an opportunity to determine whether repeat instability represents a special case confined to selected loci or a broader property of mammalian genomic ageing.

Segmental duplications and hidden structural change

Segmental duplications occupy another analytically difficult but biologically important part of mammalian genomes. These high-identity duplicated blocks provide substrates for non-allelic homologous recombination, copy-number variation and interlocus gene conversion while simultaneously creating ambiguity for short-read mapping.

Long-read and pangenomic studies have revealed substantially greater structural diversity in these regions than conventional references suggested. Haplotype-resolved assemblies demonstrated that complex loci and structural variants are disproportionately recovered when long-range information is available (Ebert et al. 2021). Vollger et al. (2023) subsequently demonstrated increased single-nucleotide variation within segmental duplications and extensive interlocus gene conversion. More recent near-

complete assemblies identified tens of megabases of duplicated sequence absent even from the T2T reference and documented extensive structural polymorphism across diverse populations (Jeong et al. 2025; Logsdon et al. 2025).

These findings invite a direct somatic question. If duplicated regions undergo elevated sequence exchange and structural variation over germline and evolutionary timescales, how frequently do analogous processes occur within cell lineages during mammalian life? The answer remains uncertain because precisely the events of greatest interest are also those most vulnerable to misclassification. A somatic gene-conversion event may cause sequence within one paralogue to resemble another existing paralogue and therefore appear to be alignment error. A duplication occurring within an already duplicated block may change local copy number without generating a uniquely mappable breakpoint.

The broader existence of structural mosaicism in normal tissues is no longer in doubt. Tissue-comparison studies identified organ-restricted somatic copy-number changes in differentiated human tissues, and later multi-organ analyses confirmed that structural and chromosomal mosaicism can occur without overt malignancy (Gao et al. 2023; Piotrowski et al. 2008). What remains unclear is how much of that structural burden originates within, terminates within, or is mediated by repetitive and duplicated architecture that conventional pipelines resolve poorly.

Such events could be relevant to a disproportionate number of biologically active gene families because segmental duplications contain genes involved in immunity, neurodevelopment, sensory function and genomic defence. Nevertheless, this remains one of the more inferential components of the missing somatic genome. The evidence is compelling that segmental duplications are structurally dynamic and historically under-resolved; it is not yet compelling that extensive normal-tissue somatic mutation occurs within them. This distinction is retained explicitly in the evidence hierarchy summarised in Table 2.

Table 2
Present strength of evidence across major components of the missing somatic genome

Genomic compartment	Direct evidence for somatic variation	Mammalian evidence base	Present interpretation	Critical unresolved question
Tandem repeats	Strong, including tissue-specific repeat expansion and age-associated repeat-length mosaicism	Strong human and mouse evidence	Established somatic mutation class	How extensive is repeat-length mosaicism outside recognised unstable loci?
Mobile elements	Strong, although estimated burden varies substantially by tissue and method	Human and mouse evidence	Established but strongly context-dependent process	What is the true body-wide burden when source elements and low-frequency insertions are completely resolved?
Mosaic chromosomal and copy-number alterations	Strong across multiple normal tissues	Predominantly human	Established somatic structural mosaicism	How much breakpoint complexity resides within repetitive or unresolved sequence?
Centromeric satellites	Limited direct normal-tissue evidence; direct de novo mutation demonstrated in pedigrees	Strong germline, pedigree and evolutionary evidence	Biologically plausible but under-measured somatically	Do centromeric arrays acquire reproducible tissue-specific somatic structure?
Acrocentric and rDNA arrays	Emerging evidence for intra-individual tissue variation	Human and comparative mammalian evidence	Constrained but measurable variation	Do stable rDNA genomic or active-copy states participate in tissue-specific selection?
Segmental duplications	Limited normal-tissue somatic measurement	Strong evidence for germline variation, gene conversion and structural diversity	Strong ascertainment concern but limited direct somatic evidence	Are somatic gene conversion and complex copy-number events systematically underestimated?
Telomeric and subtelomeric DNA	Telomere-length heterogeneity is established, whereas sequence-level somatic structural mosaicism is less defined	Broad mammalian evidence	Mixed evidence	Can lineage-specific terminal structural changes be resolved reliably at cellular scale?

Table 2 emphasises an important restraint. The missing somatic genome should not be presented as a collection of already proven mechanisms. It is instead a framework connecting established examples of hidden somatic variation with genomic territories in which the measurement problem remains substantially greater than the direct biological evidence.

Mobile elements and a genome that is already moving

Mobile elements provide perhaps the clearest demonstration that repetitive genomic sequence can generate genuine somatic structural diversity. LINE-1, Alu and SVA elements occupy large fractions of the human genome, and new insertions are difficult to identify because a newly inserted sequence closely resembles numerous copies already present elsewhere.

Early experimental work demonstrated that human neural progenitor cells can support LINE-1 retrotransposition and identified differences in LINE-1 content between brain and non-neural tissues from the same individuals, establishing biological plausibility for tissue-restricted retrotransposition (Coufal et al. 2009). Somatic mobile-element insertions were subsequently demonstrated directly in human brain tissue (Baillie et al. 2011).

The quantitative extent of neuronal retrotransposition has, however, remained method dependent. Single-neuron genome analysis of cortical and caudate neurons estimated fewer than one detectable unique somatic LINE-1 insertion per neuron in those regions (Evrony et al. 2012), whereas targeted single-cell capture sequencing in hippocampal neurons produced substantially higher estimates and showed genomic features consistent with target-primed reverse transcription (Upton et al. 2015). This discrepancy is scientifically useful rather than inconvenient because it illustrates precisely how tissue identity, assay design, amplification and analytical sensitivity can reshape estimates of somatic variation.

Mobile-element-associated somatic variation also extends beyond canonical insertion. Single-cell analysis of healthy human brain identified LINE-1-associated somatic deletions and rearrangements, linking repetitive source sequences to local structural change through mechanisms that do not require successful retrotransposition (Erwin et al. 2016). Outside the nervous system, large-scale clonal analyses have demonstrated widespread somatic LINE-1 retrotransposition in normal colorectal epithelium (Nam et al. 2023). Mouse studies add further mechanistic depth, including lineage-associated LINE-1 activation during parvalbumin interneuron development (Bodea et al. 2024).

The importance of mobile elements for the present framework is therefore both biological and methodological. A new insertion may be entirely genuine yet difficult to map because it derives from repeated sequence and may originate from a source element whose own genomic structure is imperfectly represented. An inherited LINE-1 can also become the substrate for a somatic rearrangement rather than serving solely as a mobile element. Short-read methods frequently identify only insertion junctions or partial sequence, whereas long molecules can connect a structural event with its flanking genomic context and increasingly permit source-element reconstruction.

Mobile elements also caution against equating somatic change automatically with disease. Retrotransposition can be deleterious, neutral or potentially associated with regulatory diversification depending on developmental timing, cell lineage and insertion site. A complete map of the somatic genome must therefore separate detection of genomic change from assumptions about its functional consequence.

Telomeres and subtelomeres

Telomere shortening is among the most familiar somatic genomic changes in mammalian biology, yet telomere biology is frequently considered separately from somatic mutation. The distinction is useful experimentally but less absolute biologically. Telomeric DNA is an extreme tandem-repeat system whose length changes during cell division, while adjacent subtelomeric regions contain extensive duplication, recombination and structural polymorphism.

The missing somatic genome framework therefore asks whether terminal genomic change extends beyond average telomere length. Potential forms of variation include telomeric repeat restructuring, altered representation of telomere variant repeats, subtelomeric copy-number change and lineage-specific terminal rearrangements. Evidence for these processes in normal mammalian tissues remains considerably less mature than evidence for telomere attrition itself.

Long-read approaches have begun to change the scale at which terminal sequence can be examined. Telo-seq can resolve complete telomeres together with sufficient subtelomeric sequence to assign reads to chromosome arms and, in suitable settings, to individual alleles. Application of this approach demonstrated marked chromosome-arm-specific and allele-specific telomere-length heterogeneity, progressive shortening during cellular ageing and changes associated with telomerase or alternative lengthening states (Schmidt et al. 2024).

These observations are important for the present framework because they demonstrate that measurements collapsed into one bulk telomere estimate can conceal highly structured variation between chromosome ends. Nevertheless, sequence-level structural mosaicism must remain analytically distinct from telomere-length variation. Demonstrating that a tissue contains cells with different telomere lengths does not establish that those cells contain distinct terminal sequence structures. Rigorous testing will require assays capable of measuring length, sequence composition, subtelomeric structure and cellular lineage simultaneously.

Why this is a mammalian problem rather than a human sequencing problem

The missing somatic genome is conceptually broader than the human reference genome. Comparative mammalian work has already demonstrated that somatic mutation rates are constrained across species and display an inverse relationship with lifespan despite very large differences in body size and life

history (Cagan et al. 2022). Those estimates were derived principally from genomic regions in which substitutions could be called reliably.

Whether analogous evolutionary constraints apply to repetitive and structurally complex DNA remains unknown. Long-lived mammals may have evolved enhanced suppression of repeat instability, retrotransposition, structural rearrangement or recombination within duplicated sequence. Alternatively, difficult genomic compartments may obey evolutionary rules largely independent of point-mutation rates measured in unique sequence.

The comparison is particularly informative because repetitive architecture itself differs among mammalian species. Centromeric sequence, transposable-element repertoire, telomere biology, duplication architecture and repeat composition are not fixed properties of Mammalia. A mechanism important in human tissues may therefore be absent, exaggerated or replaced in another species. Cross-species analysis could distinguish general principles of somatic genome maintenance from consequences specific to one reference architecture.

Mouse models remain especially valuable because mammalian somatic processes can be manipulated causally. Work on CAG instability provides the clearest precedent: descriptive evidence of repeat expansion was converted into mechanistic knowledge through genetic alteration of mismatch-repair pathways (Manley et al. 1999; Mouro Pinto et al. 2025). Similar experimental logic could eventually be applied to centromeric stability, rDNA dynamics, retrotransposon repression and structurally duplicated loci.

Technologies that make the question testable

Three technical developments make systematic examination of the missing somatic genome increasingly realistic. The first is the transition from incomplete linear references to complete and pangenomic representations. T2T-CHM13 demonstrated that genomic regions once treated as effectively inaccessible could be assembled continuously, while complete-centromere and near-complete diploid studies showed that these regions possess extensive sequence, structural and epigenetic diversity (Altemose et al. 2022; Logsdon et al. 2024, 2025; Nurk et al. 2022). The human pangenome demonstrated further that completeness of one genome does not represent structural diversity across individuals (Liao et al. 2023). Somatic genomics adds another requirement: ideally, the relevant comparator should be the individual's own phased constitutional genome.

Without such a reference, inherited structural variation can be misclassified as somatic change, while genuine somatic mutations arising on structurally divergent haplotypes can disappear during alignment. The logical endpoint is therefore not merely a better universal reference but donor-specific reference reconstruction for studies seeking extremely high confidence in complex genomic compartments.

The second development is long-read sequencing. Long molecules can traverse repeated sequence into unique flanking regions, phase variants, resolve full-length mobile-element insertions and reconstruct complex rearrangements that short reads fragment into ambiguous signals. Population-scale

comparisons now demonstrate major gains in structural-variant and VNTR characterisation using long reads (Ebert et al. 2021; Schloissnig et al. 2025). Somatic long-read analysis has also matured substantially in cancer, with approaches such as SAVANA demonstrating haplotype-aware detection of structural variants and copy-number abnormalities (Elrick et al. 2025). The remaining difficulty in normal tissue is cellular prevalence. A tumour provides natural clonal amplification, whereas a variant present in a small fraction of normal cells may remain unsampled regardless of read length.

The third development is the convergence of low-error single-molecule and single-cell genomics. Single-molecule approaches can detect extremely rare substitutions without requiring clonal expansion, while advances in single-cell DNA sequencing increasingly permit direct reconstruction of mosaic architecture (Abascal et al. 2021; Shao et al. 2025). Importantly, long-read sequencing has already been applied to individual human cells, enabling recovery of structural variants and tandem-repeat states in genomic regions inaccessible to short reads (Hård et al. 2023). Such studies remain technically demanding and amplification-sensitive, but they demonstrate that cellular resolution and long-range genomic resolution are no longer mutually exclusive.

Neither approach currently resolves every repetitive genomic compartment with the accuracy, depth and throughput required for body-wide mosaicism mapping. The longer-term objective is therefore technological convergence involving long molecules, high per-base accuracy, haplotype resolution and cellular resolution within the same experimental framework.

These requirements also explain why current negative evidence must be interpreted cautiously. Failure to identify somatic variation in a complex locus may reflect biological stability, inadequate molecular length, insufficient sequencing depth, reference mismatch, cellular dilution or an inappropriate variant model. Demonstrating true absence is therefore more difficult than merely failing to call a variant.

Was somatic mosaicism defined by what could be sequenced?

The history of somatic genetics can be understood partly as a history of changing instruments. Cytogenetics revealed chromosomal mosaicism because chromosomes were visible. Locus-specific sequencing revealed pathogenic postzygotic variants because defined genes could be interrogated directly. High-throughput sequencing revealed millions of substitutions and indels because uniquely mappable DNA became measurable at scale. Single-cell and single-molecule technologies subsequently pushed the detectable cellular fraction progressively downward.

Long-read sequencing and complete assemblies now extend sensitivity along a different axis by expanding the fraction of the genome that can be interrogated reliably. This extension is no longer merely theoretical. Long-read haplotype assemblies have recovered large numbers of constitutional structural variants missed by short reads, long-read population studies have characterised extensive multiallelic VNTR variation, and single-cell long-read sequencing has demonstrated access to structural

variation and tandem repeats at cellular resolution (Ebert et al. 2021; Hård et al. 2023; Schloissnig et al. 2025).

The distinction is fundamental because somatic-variant observability depends on both cellular abundance and genomic accessibility. A substitution present in a very small fraction of cells may be recoverable by high-accuracy sequencing when it lies in unique sequence. Conversely, a structural alteration present in a much larger cellular fraction may remain difficult to identify when it occurs within a large satellite array or high-identity duplication.

This creates a two-dimensional model of somatic-genomic sensitivity in which variant allele fraction and genomic mappability represent independent constraints. Most methodological development has appropriately concentrated on the first. The Missing Somatic Genome Hypothesis argues that the second deserves equivalent biological consideration.

The implication extends to mutation-rate estimates. Statements describing the number of mutations accumulated per cell, division or year are necessarily estimates within a defined callable genomic territory. Such estimates remain valid within that territory. The unresolved issue is whether extrapolation from callable unique sequence adequately represents genomic compartments governed by different molecular mechanisms, including repeat slippage, gene conversion, retrotransposition and large-array restructuring. The change in de novo mutation counts produced by inclusion of tandem-repeat and centromeric sequence in complete pedigree genomes provides a germline example of how the measurable mutation spectrum changes when the callable genome expands (Porubsky et al. 2025).

The eventual answer may be reassuring. Previously inaccessible regions may contribute relatively little additional biologically consequential somatic mosaicism. Alternatively, they may reveal a layer of somatic variation qualitatively different from the mutation classes that currently dominate the field. Both outcomes are scientifically informative.

The Missing Somatic Genome Hypothesis

The framework can be expressed as a falsifiable hypothesis rather than as a presumption. The **Missing Somatic Genome Hypothesis** proposes that repetitive and structurally complex genomic compartments contain a quantitatively and functionally important burden of postzygotic variation that has been systematically underestimated by reference-dependent short-read somatic genomics.

The hypothesis contains three separable claims. Its quantitative component predicts that sufficiently complete analysis will reveal more somatic variation in difficult genomic compartments than expected merely from adding previously unmeasured genomic territory. Its qualitative component predicts that the spectrum of mutation will differ from that observed in unique sequence, with expansion, contraction, gene conversion, mobile-element insertion and complex rearrangement contributing disproportionately. Its functional component predicts that at least some of these changes influence cellular phenotype, lineage behaviour, clonal fitness, tissue function or disease.

None of these claims follows automatically from historical inaccessibility. The appropriate null model is that, after adjustment for sequence composition, genomic territory, technical error and detection sensitivity, repetitive and structurally complex compartments contribute little additional biologically meaningful somatic variation beyond processes already recognised by conventional somatic-genome studies.

The value of this formulation lies in its capacity to fail experimentally. The principal predictions and corresponding falsification strategies are summarised in Table 3. Failure of one component would not necessarily invalidate the others. Previously inaccessible regions might contain numerous additional somatic variants that are overwhelmingly neutral, or they might contain relatively few variants but include a small number with substantial functional consequences.

Table 3
Testable predictions and falsification criteria for the Missing Somatic Genome Hypothesis

Prediction	Experimental strategy	Observation supporting the hypothesis	Observation favouring the null model
Somatic variant burden is disproportionately enriched in difficult genomic compartments	Matched high-depth short-read and long-read sequencing using donor-specific phased assemblies	Validated mutation burden per callable base remains elevated in complex regions after technical correction	Mutation burden is similar to or lower than that in uniquely mappable sequence
Mutation spectra differ according to genomic architecture	Multi-tissue long-read analysis of repeat, duplication, mobile-element and satellite compartments	Reproducible enrichment of expansion, contraction, conversion or rearrangement processes	Mutation spectra resemble technical background or conventional unique-sequence processes
Difficult-region variants participate in lineage structure	Single-cell, microclonal or spatially separated long-read analysis	Shared validated variants reconstruct coherent developmental or clonal relationships	Candidate events fail replication or show no consistent lineage structure
Selected variants influence cellular phenotype	Isogenic cellular, organoid or mammalian models containing defined structural or repeat changes	Reproducible effects on transcription, chromatin, genome stability or cellular fitness	No detectable phenotype under biologically relevant conditions
Conventional short-read analysis systematically misses validated events	Blinded benchmarking against synthetic and biological truth sets	Long-read or assembly approaches recover reproducible events absent from standard short-read calls	Optimised short-read pipelines recover nearly all biologically meaningful events
Somatic maintenance of difficult sequence shows evolutionary constraint	Comparable analyses in mammals differing in lifespan and genome architecture	Reproducible species-level relationships between instability and life-history traits	Apparent differences disappear after correction for platform and reference quality

Table 3 also provides a safeguard against converting technological enthusiasm into biological certainty. A larger number of calls generated by long-read sequencing is not itself evidence of a missing somatic genome. Candidate events must survive orthogonal validation, demonstrate coherent biological distributions and, where functional importance is claimed, produce measurable consequences.

Measurement and validation

Long reads cannot by themselves solve the problem. Repetitive genomic compartments generate unusually difficult error models, and an analysis designed to discover previously invisible biology can

also generate previously invisible artefact.

DNA extraction may recover very long repetitive molecules unevenly. Amplification can distort tandem-repeat length. Cultured cells may acquire structural changes absent in vivo. Differences in tissue composition can mimic copy-number mosaicism. A reference haplotype that differs structurally from the sequenced chromosome may generate apparent rearrangements. Sequencing errors can become non-random in homopolymers and repetitive contexts. Bulk-tissue variation can also reflect early developmental lineage restriction rather than mutation acquired during adult life.

The evidentiary threshold should therefore increase with genomic ambiguity. Tandem-repeat changes should be confirmed using an independent repeat-sizing or long-read strategy. Large-array copy-number differences should ideally be supported by orthogonal approaches such as digital PCR or optical mapping. Structural events should have breakpoint-spanning molecular support, and candidate clonal mutations should demonstrate reproducible lineage structure in isolated cells, microdissections or spatially independent tissue units.

The mobile-element literature provides an instructive example of why validation strategy matters. Different single-cell approaches have yielded markedly different estimates of LINE-1 insertion burden in human neurons (Evrony et al. 2012; Upton et al. 2015). Rather than treating one estimate as universally correct, future complete-somatic-genome studies should regard such discrepancies as evidence that amplification method, tissue source, cell type, variant definition and analytical pipeline can materially affect inference.

Constitutional comparison is equally important. Blood is convenient but cannot always function as an unquestioned germline control because haematopoietic tissues themselves develop age-associated clonal mosaicism. Where feasible, donor-specific constitutional reconstruction should incorporate more than one tissue source and phased inherited variation to distinguish true postzygotic change from reference mismatch or constitutional structure.

The methodological principle is therefore proportionality: the more difficult the genomic compartment, the stronger the orthogonal evidence required before a somatic event is accepted.

Implications for ageing

Ageing is accompanied by increasing somatic mutation, altered DNA repair, epigenetic drift, clonal selection and changes in tissue architecture. Most mutation-centred models of ageing have understandably concentrated on substitutions, small indels and readily detectable structural alterations. Repetitive sequence introduces another possibility: ageing may also alter the capacity of cells to preserve genomic structures whose maintenance depends disproportionately on mismatch repair, replication-fork stability, chromatin organisation and suppression of mobile elements.

The 2026 biobank analysis provides an important signal in this direction by demonstrating that repeat-length mosaicism can increase with age outside the narrow setting of classical rare expansion disorders

(Hujoel et al. 2026). Somatic LINE-1 activity provides a second example of structural genomic diversification occurring in normal tissues, while age-associated chromosomal alterations demonstrate that clonally expanded structural mosaicism is not confined to malignancy (Gao et al. 2023; Nam et al. 2023).

The rDNA literature adds an important counterpoint. Absolute rDNA copy number in blood does not appear simply to decline with chronological age, whereas promoter methylation and the proportion of potentially active units change substantially (Geisen et al. 2025). This suggests that ageing-related variation in repetitive compartments may involve regulation and selective silencing as much as sequence loss or expansion.

Telomeres provide a further example of structured rather than uniform change. Long-read analysis demonstrates chromosome-arm-specific and allele-specific differences in telomere length and attrition, indicating that one bulk telomere measurement can conceal substantial heterogeneity among chromosome ends (Schmidt et al. 2024).

None of these observations establishes that repetitive or structurally complex genomic instability causes ageing. Such changes may instead represent molecular records of cumulative cellular history. A repeat that expands progressively could integrate DNA-repair exposure or proliferative history without itself altering cellular fitness. Conversely, selected changes could influence gene regulation, chromosomal stability or cell survival and thereby contribute actively to ageing phenotypes.

The relevant conceptual shift is therefore not to label difficult-region mutations as ageing drivers, but to include them among the genomic variables that ageing studies should be capable of measuring.

Implications for disease

Hidden somatic variation could influence disease through several mechanistically distinct routes. The clearest is direct pathogenic mosaicism, in which a postzygotic repeat expansion, insertion or structural rearrangement alters gene function in a vulnerable cellular population. A second mechanism is disease modification, in which additional somatic change to an inherited allele influences tissue vulnerability or clinical timing. Huntington disease provides the strongest demonstration of this principle because somatic CAG expansion is linked to cell-type vulnerability and modified by DNA-repair pathways (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium 2025; Mätlik et al. 2024).

A third possibility is genomic predisposition rather than direct causation. Changes in centromeric, telomeric or duplicated regions might increase the probability of subsequent mis-segregation, recombination or chromosomal restructuring without generating a phenotype independently. A fourth possibility is clonal selection, in which an unconventional structural or repetitive variant alters cellular fitness sufficiently to permit lineage expansion without producing overt neoplasia.

Cancer will remain an important experimental proving ground because clonal expansion increases the detectability of rare structural events, and long-read cancer-genome analysis already demonstrates gains

in reconstructing complex somatic rearrangements (Elrick et al. 2025). Nevertheless, the principal objective of the missing somatic genome framework is to establish what occurs in normal mammalian tissues before disease-associated enrichment is interpreted.

When a class of mutation becomes measurable only after cells have undergone neoplastic expansion, distinguishing a process created by malignant genomic instability from a pre-existing process amplified by clonal growth becomes difficult. Establishing the normal baseline is therefore essential.

An experimental programme for the complete somatic genome

Testing this framework requires a research programme that separates technical discovery from biological interpretation. The first priority is development of analytical truth sets. Synthetic mixtures containing defined repeat-length changes, mobile-element insertions and structural rearrangements at controlled cellular fractions could establish platform-specific detection limits and quantify false-positive rates within difficult sequence.

The next requirement is donor-specific constitutional genome reconstruction. For the most complex loci, comparison against a generic reference is insufficient because inherited structural variation can resemble somatic change. Haplotype-resolved constitutional assemblies would allow each tissue to be compared with the genome from which it actually arose rather than with an abstract population reference. The feasibility of reconstructing complex and near-complete diploid haplotypes is now established across increasingly diverse human genomes (Liao et al. 2023; Logsdon et al. 2025).

Multi-tissue analysis should then determine whether candidate changes are shared developmentally, restricted to one organ or associated with particular cell lineages. A variant present across tissues derived from distinct embryological lineages would indicate an early postzygotic event, whereas a variant confined to one clonal tissue unit could have arisen considerably later. Existing observations of tissue-restricted copy-number mosaicism and intra-individual rDNA variation demonstrate that this comparison is experimentally tractable (Durackova et al. 2025; Piotrowski et al. 2008).

Single-cell, microclonal and spatial sampling would subsequently provide the resolution necessary to distinguish genuine lineage structure from tissue-composition effects. Long-read sequencing of individual cells already demonstrates proof of principle that structural variants and tandem-repeat states can be recovered at cellular resolution, although amplification bias, incomplete coverage and error remain important limitations (Hård et al. 2023).

Mechanistic work must follow descriptive discovery. Mouse models, organoids and engineered mammalian cell systems can test whether defined changes in DNA repair, replication stress, transcription, chromatin organisation or ageing alter instability within particular genomic compartments. Functional experiments must then determine whether the resulting somatic differences change transcription, chromatin, genome stability, cellular fitness or tissue phenotype.

The order of these steps matters. Discovery alone cannot establish mechanism, and mechanism alone cannot establish population relevance. A convincing complete-somatic-genome programme must connect molecular detection, lineage structure, causal mechanism and functional consequence.

Towards a somatic pangenome

Population genomics is moving from the concept of one linear reference genome towards pangenomic representations capable of encoding multiple inherited haplotypes and alternative genomic structures. Somatic genomics may eventually require a comparable conceptual transition within the individual organism.

A somatic pangenome would not simply catalogue inherited variation among individuals. It would represent an individual's constitutional haplotypes together with the structurally coherent genomic states acquired by different cellular lineages. Such a representation might need to encode alternative repeat lengths, mobile-element insertions, structural haplotypes, copy-number states and other variants together with their cellular prevalence and tissue distribution.

This is computationally challenging because one organism can contain numerous low-frequency genomic states rather than a small number of stable haplotypes. Nevertheless, the conceptual direction is important. Mammalian genome biology has already progressed from the assumption of one canonical genome for a species to recognition of extensive inherited structural diversity among individuals. The difference between structural-variant discovery using short reads, haplotype-resolved assemblies and contemporary population-scale long reads demonstrates how strongly representation determines the visible landscape (Ebert et al. 2021; Liao et al. 2023; Schloissnig et al. 2025).

Somatic mosaicism requires a further transition: recognition that even within one individual there is no single immutable genome shared by every cell. The Missing Somatic Genome Hypothesis therefore connects naturally with the pangenomic era. A reference system capable of representing inherited sequence comprehensively while ignoring postzygotic structural diversity would remain incomplete for understanding tissue biology.

Limitations of the framework

The framework has several important limitations. First, the evidence base remains heavily weighted towards humans despite the mammalian scope of the argument. Mouse studies provide strong mechanistic evidence for selected processes, particularly tandem-repeat instability and retrotransposon biology, but body-wide comparative datasets across mammalian species remain limited.

Second, evidence is markedly uneven across genomic compartments. Somatic tandem-repeat instability, mobile-element mobilisation and conventional structural mosaicism are directly established. Widespread somatic restructuring of normal centromeric satellites, rDNA arrays or segmental duplications is not. Recent pedigree evidence for centromeric de novo mutation and multi-tissue

evidence for rDNA variation strengthen biological plausibility but should not be mistaken for proof of widespread somatic instability in those compartments (Durackova et al. 2025; Porubsky et al. 2025).

Third, the term “missing” is intrinsically technology dependent. Genomic regions that were effectively inaccessible a decade ago are increasingly measurable, and future analytical developments may render today's distinction obsolete. The concept therefore refers to historical and current ascertainment bias rather than to a permanent category of DNA.

Fourth, detecting additional variants does not establish biological importance. Complete sequencing may reveal large quantities of neutral somatic structural variation that exert little influence on cellular phenotype. Such an outcome would still be informative because it would define the degree of structural heterogeneity that mammalian tissues can tolerate.

Fifth, apparent abundance can be strongly assay dependent. The conflicting quantitative estimates of neuronal LINE-1 mosaicism provide a cautionary example that technological sensitivity and specificity can produce very different biological narratives from the same genomic compartment (Evrony et al. 2012; Upton et al. 2015).

Finally, repetitive and structurally complex sequence poses unusually severe validation challenges. New technology can increase both discovery and artefact. Strong evidence will therefore require replication, orthogonal confirmation, carefully matched constitutional material and explicit modelling of platform-specific error. These constraints are not weaknesses of the hypothesis; they define the standards required to test it rigorously.

Conclusion

Somatic genomics has transformed the mammalian organism from a genetically uniform entity into a dynamic population of related but non-identical cellular lineages. Yet the maps underlying that transformation were constructed disproportionately from genomic regions that reference assemblies and sequencing technologies could interrogate most confidently.

That observational boundary is now shifting. Telomere-to-telomere assemblies have resolved sequence previously absent from reference genomes. Diverse near-complete genomes reveal structural complexity that persists beyond a single finished reference. Pangenomes expose variation hidden by linear representation. Long reads traverse repetitive arrays and reconstruct mobile-element insertions, VNTRs and complex rearrangements. Single-cell long-read approaches demonstrate that structural and repeat variation can increasingly be interrogated at cellular resolution. At the same time, biological studies have established that tandem repeats can expand somatically, mobile elements can generate insertional and deletional mosaicism, normal tissues can harbour copy-number mosaicism, and rDNA and telomeric states can vary within an individual.

The unresolved question is therefore no longer whether difficult genomic regions are mutable. It is whether their historical undermeasurement has materially shaped the understanding of mammalian somatic evolution.

The Missing Somatic Genome Hypothesis provides a falsifiable framework for answering that question. It predicts that repetitive and structurally complex genomic compartments contain a layer of postzygotic variation that conventional reference-dependent short-read analyses have systematically underestimated and that at least part of this variation has biological consequences. The hypothesis may prove broadly correct, or it may ultimately apply only to selected loci, tissues and mutational processes. Either result would refine the boundaries of somatic genome biology.

A complete mammalian genome can no longer be defined solely by whether every chromosome has been assembled from telomere to telomere. Completeness must eventually also address whether genomic change has been measured across the cells, tissues and lifespan through which those chromosomes persist.

Declarations

Competing interests

The author declares no competing interests.

Ethics approval

This Review synthesises previously published literature and reports no new research involving human participants, animals or biological material. Ethical approval was therefore not required.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

No new datasets were generated or analysed for this Review. All evidence discussed is derived from the published literature cited in the manuscript.

Author contributions

The author conceived the Review, developed the conceptual framework, evaluated and interpreted the literature, and critically revised and approved the final manuscript.

Acknowledgements

None.

Funding

The author received no specific funding for this work.

References

1. Abascal F, Harvey LMR, Mitchell E et al (2021) Somatic mutation landscapes at single-molecule resolution. *Nature* 593:405–410. <https://doi.org/10.1038/s41586-021-03477-4>
2. Altemose N, Logsdon GA, Bzikadze AV et al (2022) Complete genomic and epigenetic maps of human centromeres. *Science* 376:eabl4178. <https://doi.org/10.1126/science.abl4178>
3. Baillie JK, Barnett MW, Upton KR et al (2011) Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479:534–537. <https://doi.org/10.1038/nature10531>
4. Bodea GO, Botto JM, Ferreiro ME et al (2024) LINE-1 retrotransposons contribute to mouse PV interneuron development. *Nat Neurosci* 27:1274–1284. <https://doi.org/10.1038/s41593-024-01650-2>
5. Cagan A, Baez-Ortega A, Brzozowska N et al (2022) Somatic mutation rates scale with lifespan across mammals. *Nature* 604:517–524. <https://doi.org/10.1038/s41586-022-04618-z>
6. Coorens THH, Oh JW, Choi YA et al (2025) The Somatic Mosaicism across Human Tissues Network. *Nature* 643:47–59. <https://doi.org/10.1038/s41586-025-09096-7>
7. Coufal NG, Garcia-Perez JL, Peng GE et al (2009) L1 retrotransposition in human neural progenitor cells. *Nature* 460:1127–1131. <https://doi.org/10.1038/nature08248>
8. Durackova J, Potabattula R, Rosenwald A, Haaf T (2025) Intraindividual rDNA copy number variation and methylation in humans. *Exp Cell Res* 453:114793. <https://doi.org/10.1016/j.yexcr.2025.114793>
9. Ebert P, Audano PA, Zhu Q et al (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372:eabf7117. <https://doi.org/10.1126/science.abf7117>
10. Elrick H, Sauer CM, Espejo Valle-Inclan J et al (2025) SAVANA: reliable analysis of somatic structural variants and copy number aberrations using long-read sequencing. *Nat Methods* 22:1436–1446. <https://doi.org/10.1038/s41592-025-02708-0>
11. Erwin JA, Paquola ACM, Singer T et al (2016) L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci* 19:1583–1591. <https://doi.org/10.1038/nn.4388>
12. Evrony GD, Cai X, Lee E et al (2012) Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* 151:483–496. <https://doi.org/10.1016/j.cell.2012.09.035>
13. Gao S, Oshima KK, Chuang SC et al (2026) A global view of human centromere variation and evolution. *Nature*. <https://doi.org/10.1038/s41586-026-10841-9>

14. Gao T, Kastriti ME, Ljungström V et al (2023) A pan-tissue survey of mosaic chromosomal alterations in 948 individuals. *Nat Genet* 55:1901–1911. <https://doi.org/10.1038/s41588-023-01537-1>
15. Geisen ABC, Santana Acevedo N, Oshima J et al (2025) rDNA copy number variation and methylation during normal and premature aging. *Aging Cell* 24:e14497. <https://doi.org/10.1111/accel.14497>
16. Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium (2025) Genetic modifiers of somatic expansion and clinical phenotypes in Huntington's disease highlight shared and tissue-specific effects. *Nat Genet* 57:1426–1436. <https://doi.org/10.1038/s41588-025-02191-5>
17. Gibbons JG, Branco AT, Yu S, Lemos B (2014) Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans. *Nat Commun* 5:4850. <https://doi.org/10.1038/ncomms5850>
18. Guarracino A, Buonaiuto S, Gomes de Lima L et al (2023) Recombination between heterologous human acrocentric chromosomes. *Nature* 617:335–343. <https://doi.org/10.1038/s41586-023-05976-y>
19. Hård J, Mold JE, Einfeldt J et al (2023) Long-read whole-genome analysis of human single cells. *Nat Commun* 14:5164. <https://doi.org/10.1038/s41467-023-40898-3>
20. Hujoel MLA, Handsaker RE, Tang D et al (2026) Insights into DNA repeat expansions among 900,000 biobank participants. *Nature* 650:920–929. <https://doi.org/10.1038/s41586-025-09886-z>
21. Jeong H, Dishuck PC, Yoo D et al (2025) Structural polymorphism and diversity of human segmental duplications. *Nat Genet* 57:390–401. <https://doi.org/10.1038/s41588-024-02051-8>
22. Law PP, Mikheeva LA, Rodriguez-Algarra F et al (2024) Ribosomal DNA copy number is associated with body mass in humans and other mammals. *Nat Commun* 15:5006. <https://doi.org/10.1038/s41467-024-49397-5>
23. Li R, Di L, Li J et al (2021) A body map of somatic mutagenesis in morphologically normal human tissues. *Nature* 597:398–403. <https://doi.org/10.1038/s41586-021-03836-1>
24. Liao WW, Asri M, Ebler J et al (2023) A draft human pangenome reference. *Nature* 617:312–324. <https://doi.org/10.1038/s41586-023-05896-x>
25. Logsdon GA, Rozanski AN, Ryabov F et al (2024) The variation and evolution of complete human centromeres. *Nature* 629:136–145. <https://doi.org/10.1038/s41586-024-07278-3>
26. Logsdon GA, Ebert P, Audano PA et al (2025) Complex genetic variation in nearly complete human genomes. *Nature* 644:430–441. <https://doi.org/10.1038/s41586-025-09140-6>
27. Manley K, Shirley TL, Flaherty L, Messer A (1999) Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. *Nat Genet* 23:471–473. <https://doi.org/10.1038/70598>
28. Mätlik K, Baffuto M, Kus L et al (2024) Cell-type-specific CAG repeat expansions and toxicity of mutant Huntingtin in human striatum and cerebellum. *Nat Genet* 56:383–394. <https://doi.org/10.1038/s41588-024-01653-6>

29. Mouro Pinto R, Murtha R, Azevedo A et al (2025) In vivo CRISPR–Cas9 genome editing in mice identifies genetic modifiers of somatic CAG repeat instability in Huntington’s disease. *Nat Genet* 57:314–322. <https://doi.org/10.1038/s41588-024-02054-5>
30. Nam CH, Youk J, Kim JY et al (2023) Widespread somatic L1 retrotransposition in normal colorectal epithelium. *Nature* 617:540–547. <https://doi.org/10.1038/s41586-023-06046-z>
31. Nurk S, Koren S, Rhie A et al (2022) The complete sequence of a human genome. *Science* 376:44–53. <https://doi.org/10.1126/science.abj6987>
32. Piotrowski A, Bruder CEG, Andersson R et al (2008) Somatic mosaicism for copy number variation in differentiated human tissues. *Hum Mutat* 29:1118–1124. <https://doi.org/10.1002/humu.20815>
33. Porubsky D, Dashnow H, Sasani TA et al (2025) Human de novo mutation rates from a four-generation pedigree reference. *Nature* 643:427–436. <https://doi.org/10.1038/s41586-025-08922-2>
34. Schloissnig S, Pani S, Ebler J et al (2025) Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* 644:442–452. <https://doi.org/10.1038/s41586-025-09290-7>
35. Schmidt TT, Tyler C, Rughani P et al (2024) High resolution long-read telomere sequencing reveals dynamic mechanisms in aging and cancer. *Nat Commun* 15:5149. <https://doi.org/10.1038/s41467-024-48917-7>
36. Shao DD, Kriz AJ, Snellings DA et al (2025) Advances in single-cell DNA sequencing enable insights into human somatic mosaicism. *Nat Rev Genet* 26:761–774. <https://doi.org/10.1038/s41576-025-00832-3>
37. Upton KR, Gerhardt DJ, Jesuadian JS et al (2015) Ubiquitous L1 mosaicism in hippocampal neurons. *Cell* 161:228–239. <https://doi.org/10.1016/j.cell.2015.03.026>
38. Vollger MR, Dishuck PC, Harvey WT et al (2023) Increased mutation and gene conversion within human segmental duplications. *Nature* 617:325–334. <https://doi.org/10.1038/s41586-023-05895-y>