

The Somatic Mosaicism across Human Tissues Network

<https://doi.org/10.1038/s41586-025-09096-7>

Received: 14 May 2024

Accepted: 2 May 2025

Published online: 2 July 2025

 Check for updates

Tim H. H. Coorens^{1,86,87}✉, Ji Won Oh^{2,3,87}✉, Yujin Angelina Choi², Nam Seop Lim^{2,3}, Boxun Zhao^{1,4,5,6}, Adam Voshall^{1,4,5,6}, Alexej Abyzov⁷, Lucinda Antonacci-Fulton⁸, Samuel Aparicio^{9,10,11}, Kristin G. Ardlie¹, Thomas J. Bell¹², James T. Bennett^{13,14}, Bradley E. Bernstein^{1,15,16}, Thomas G. Blanchard¹⁷, Alan P. Boyle^{18,19}, Jason D. Buenrostro^{1,20}, Kathleen H. Burns^{1,16,21}, Fei Chen^{1,20}, Rui Chen^{22,23}, Sangita Choudhury^{1,4,6}, Harsha V. Doddapaneni²², Evan E. Eichler^{24,25}, Gilad D. Evrony^{26,27,28}, Melissa A. Faith^{29,30}, Thomas G. Fazzio³¹, Robert S. Fulton³², Manuel Garber³³, Nils Gehlenborg³⁴, Soren Germer¹¹, Gad Getz^{1,16,35}, Richard A. Gibbs^{22,23}, Raquel G. Hernandez^{36,37}, Fulai Jin^{38,39}, Jan O. Korbel^{40,41}, Dan A. Landau^{11,42,43}, Heather A. Lawson⁴⁴, Niall J. Lennon¹, Heng Li^{34,45}, Yan Li³⁸, Po-Ru Loh^{1,46}, Gabor Marth⁴⁷, Michael J. McConnell⁴⁸, Ryan E. Mills^{18,19}, Stephen B. Montgomery^{49,50,51}, Pradeep Natarajan^{1,52,53}, Peter J. Park^{34,54}, Rahul Satija^{11,55}, Fritz J. Sedlazeck^{22,23,56}, Diane D. Shao^{4,6,57}, Hui Shen⁵⁸, Andrew B. Stergachis^{24,59,60}, Hunter R. Underhill^{61,62}, Alexander E. Urban^{50,63}, Melissa W. VonDran¹², Christopher A. Walsh^{1,4,5,6,64}, Ting Wang^{8,44,65}, Tao P. Wu^{23,66,67}, Chenghang Zong^{23,67}, Eunjung Alice Lee^{1,4,5,6}✉, Flora M. Vaccarino^{68,69,70}✉, The Somatic Mosaicism across Human Tissues Network*

From fertilization onwards, the cells of the human body acquire variations in their DNA sequence, known as somatic mutations. These postzygotic mutations arise from intrinsic errors in DNA replication and repair, as well as from exposure to mutagens. Somatic mutations have been implicated in some diseases, but a fundamental understanding of the frequency, type and patterns of mutations across healthy human tissues has been limited. This is primarily due to the small proportion of cells harbouring specific somatic variants within an individual, making them more challenging to detect than inherited variants. Here we describe the Somatic Mosaicism across Human Tissues Network, which aims to create a reference catalogue of somatic mutations and their clonal patterns across 19 different tissue sites from 150 non-diseased donors and develop new technologies and computational tools to detect somatic mutations and assess their phenotypic consequences, including clonal expansions. This strategy enables a comprehensive examination of the mutational landscape across the human body, and provides a comparison baseline for somatic mutation in diseases. This will lead to a deep understanding of somatic mutations and clonal expansions across the lifespan, as well as their roles in health, in ageing and, by comparison, in diseases.

Genetic diversity within the human population has been well described. The Human Genome Project resulted in the first near-complete mapping of the human DNA sequence¹, and was followed by large-scale projects, such as the 1,000 Genomes Project² and the Pangenome project³, that mapped the genetic diversity between individuals and populations. Now, there is growing recognition that extensive genetic variation exists within individuals among different tissues and cells. Two decades after completion of the first draft human genome, the Somatic Mosaicism across Human Tissues (SMAHT) Network plans to map the genetic diversity across different tissues and cells within individuals.

From fertilization onwards, the cells of the human body continuously experience damage to their genome, either from intrinsic causes or from exposure to mutagens^{4–9}. Although the vast majority of DNA

damage is repaired, and the genome is replicated with extremely high fidelity, cells steadily acquire somatic mutations throughout life. All cells within an individual harbour somatic mutation, but any given mutation is present in only a subset of the cells, or even in single cells. Hence, somatic mutations are often described as mosaic^{10,11}.

The detection of somatic mutations is challenging. In contrast to inherited variants, somatic mutations only exist in small and variable proportions of cells, ranging from embryonic mutations present in most cells to mutations present in just a single cell (Fig. 1a). This challenge is exacerbated by the introduction of artefacts and errors resembling low-frequency mutations during DNA library preparation and sequencing¹². Current short-read sequencing technologies limit detection of mutations in repetitive regions of the genome and are likely to be less suitable for detection of somatic structural variations.

A list of affiliations appears at the end of the paper. *Lists of authors and their affiliations appear at the end of the paper.

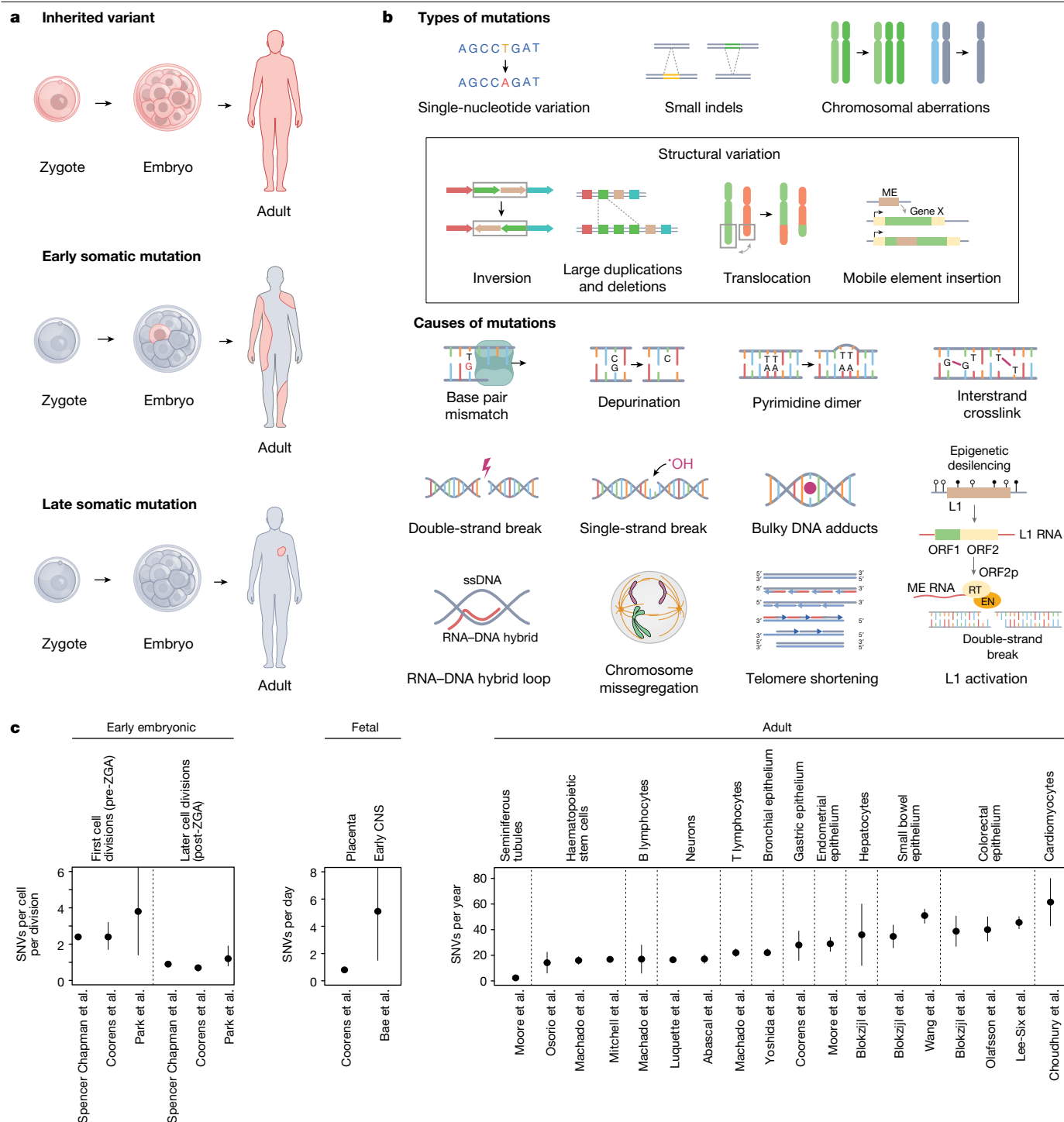


Fig. 1 | Somatic mutations, causes and patterns. a, Schematic comparison between inherited variants, an early somatic mutation and a late somatic mutation. **b**, Overview of causes and types of somatic mutations. EN, endonuclease; ME, mobile element; ORF, open reading frame; RT, reverse transcriptase; ssDNA, single-stranded DNA. **c**, Overview of the reported mutation rates of somatic SNV across developmental stages and tissues. Data of first cell divisions^{6,7,59} and later cell divisions^{6,7,59} are SNVs per cell

per division. Data from fetal development of the early central nervous system (CNS)⁹ and placenta⁶² are SNVs per cell per day. Adult data are SNVs per year and estimated for seminiferous tubules⁴⁸, haematopoietic stem cells^{26,52,144}, B lymphocytes⁵², neurons^{63,145}, T lymphocytes⁵², bronchial epithelium⁵³, gastric epithelium¹⁴⁶, endometrial epithelium⁷⁹, hepatocytes¹⁹, small bowel epithelium^{19,115}, colorectal epithelium^{19,24,29} and cardiomyocytes⁴⁹. ZGA, zygotic genome activation.

Although most somatic mutations are probably functionally neutral¹³, some can profoundly alter the phenotype of a cell and are implicated in a wide variety of diseases. Many insights have come from sequencing the genomes of cancers¹⁴, the best-known example of disease arising from somatic mutation, but mutagenesis in tumours is often accelerated, and normal mutational patterns are

distorted by genome instability. More recently, mapping the patterns of somatic mutations in normal tissues, exemplified by efforts of the Brain Somatic Mosaicism Network and other studies^{3-6,15-26}, has identified a role for somatic mutations in developmental syndromes, neurological diseases and inflammatory disorders²⁷⁻³⁹. Despite these efforts, there is currently no comprehensive reference

dataset of somatic mosaicism across many tissues of a large pool of donors.

In this Perspective article, we describe the SMaHT Network, initiated by the NIH Common Fund, which aims to generate a reference catalogue of somatic variation from 150 donors in 19 non-diseased tissue sites. To advance the field, the SMaHT Network will perform a comprehensive discovery and analysis of all types of somatic mutations at an unprecedented scale: the joint analysis of mosaicism across many tissues per donor; the robust discovery of structural variants (SVs) through long-read sequencing and donor-specific assemblies; and the widespread and robust application of ultrasensitive sequencing technologies, such as duplex sequencing, across sequencing centres. Furthermore, beyond applying established sequencing assays at scale, the SMaHT Network has a strong emphasis on tool and technological development to enable the next generation of somatic mutation studies. Before describing the network goals in detail, we briefly review the current knowledge about somatic mutations in health and disease, as well as the technical challenges in detection of mutations. A large part of the SMaHT Network will focus on the development of technologies and computational tools to improve the detection of all types of somatic variation.

Somatic mutations in healthy tissues

Throughout the human lifespan, from conception to death, cells acquire mutations in their DNA^{6–9,40} (Fig. 1b). These somatic mutations can be the consequence of erroneous repair of damaged DNA bases or DNA strand breaks, errors during replication, chromosome mis-segregation or the integration of mobile elements. Somatic mutations can be divided into different types⁴¹: substitutions, the vast majority of which are single-nucleotide variants (SNVs); small (less than 50 bp) insertions and deletions (indels); SVs, including segmental duplications, large deletions, translocations, inversions, mobile element insertions (MEIs) and complex SVs, including chromothripsis and chromoplexy; and other large chromosomal aberrations, such as whole-chromosome gains and losses. Duplications, deletions and whole-chromosome gains and losses are also referred to as copy number variants (CNVs) or mosaic chromosomal alterations. These classes differ profoundly in their underlying causes and patterns across tissues and their phenotypic effects on cells. In normal tissues, SNVs are by far the most common type of somatic variation, followed by indels. SVs and large chromosomal aberrations are observed less frequently²⁷, but typically affect more base pairs and thus may have larger functional effects. However, most previous studies on somatic mutations have relied on short-read DNA sequencing, which may fail to detect various types of SVs. Studies of germline differences have shown that SVs are far more abundant, but the majority are missed by short-read approaches⁴².

Different mutagenic processes cause distinct patterns of somatic mutation, depending on the types of DNA damage incurred and the pathways responsible for DNA lesion repair. Research over the past decade has deconvolved these patterns into mutational signatures and linked certain signatures to specific mutagens, such as ultraviolet light, tobacco smoke, chemotherapy or natural age-related accumulation of endogenous mutations⁴³. Mutational signatures are most commonly applied to SNVs⁴⁴, but they have been defined for other classes of somatic mutations, including indels⁴⁴, chromosomal alterations^{45,46} and SVs^{44,47}. In the context of SNVs, mutational signatures reflect the distribution of specific base changes within their trinucleotide contexts.

All normal tissues, including post-mitotic cells, exhibit SNV mutational signatures linked to clock-like endogenous processes (single base substitution signature 1 (SBS1) or SBS5) and, to a lesser extent, oxidative damage (SBS18)^{48–50}. Mutational signatures linked to mutagenic exposure can be confined to specific organs, such as UV damage (SBS7)

in the skin⁵¹ or skin-resident T lymphocytes⁵², damage from tobacco smoke (SBS4) in the bronchial epithelium of the lung⁵³ and exposure to a genotoxic strain of *Escherichia coli* (SBS88)²⁴ in the large intestine. These exposure differences drive some of the variation in the types of somatic mutations observed across different tissues of the human body^{7,48,54}. Furthermore, different mutational processes show different correlations with genomic features, such as replication timing, replication strand and transcription strand^{55–58}, reflecting genomic biases of DNA damage and repair.

The somatic mutation rate varies across human tissues and life stages (Fig. 1c). During the initial embryonic cell divisions, somatic SNVs accumulate at a high rate of approximately three per division, probably due to the high division rate and delayed activation of the zygotic genome^{6,7,59}. Afterwards, the mutation rate decreases (approximately one SNV per division) during development in utero, in both embryonic tissues, such as the fetal brain^{9,60,61}, and extraembryonic tissues, such as the placenta⁶². After birth, mutation rates further decline 5–10-fold and vary substantially across tissues, from 16–20 SNVs per year in post-mitotic cells such as neurons^{21,50,61,63} to 44 SNVs per year in colonic stem cells²⁴ (Fig. 1c). Germ cells have the lowest somatic mutation rate reported²³, in line with the parental age effect on de novo germline mutations⁴⁸. Although division rate may influence the endogenous somatic mutation rate, there are probably other factors that modulate both mutagenesis and repair of DNA damage^{64–66}.

Large somatic mutations such as SVs, chromosomal alterations and MEIs are detected much less frequently than SNVs and indels. Although somatic aneuploidy appears to be rare, sub-chromosomal structural variations affect 13–41% of neurons^{18,34,67,68}. Frequent CNVs, mostly duplications of likely developmental origin, have been detected in approximately 7% of brains from the Brain Somatic Mosaicism Network consortium³⁴, and mosaic chromosomal alterations have been observed in approximately 5% of blood samples in the UK Biobank⁶⁹. Single-neuron DNA sequencing of mobile element-enriched libraries or whole genomes has revealed MEI events that appear to occur during development and create mosaicism in the human brain^{5,70,71}. Bulk sequencing approaches have also detected a few examples of somatic MEIs in the brain⁷² and non-brain tissues including the heart⁷³, fibroblasts⁷³ and liver⁷⁴. Recent somatic MEI profiling in colorectal epithelial single-cell clones has indicated peak insertion rates during early embryogenesis⁷⁵. Considering the potential effect of these large mutations on the sequence, splicing or expression of genes^{76,77}, it is valuable to understand their prevalence across human tissues during development and ageing.

Although most somatic mutations do not discernibly affect the phenotype of a cell, some somatic mutations are under selection in different tissues. Such driver mutations may lead to a proliferative advantage or increased survival of the cell and its progeny, resulting in clonal expansions in tissues. Cancer is the canonical example of somatic evolution and often involves the stepwise accumulation of key somatic mutations and genomic instability^{1,78}. Mutations typically associated with cancer can be abundant across normal tissues with age. For comparison, in a typical individual of 60 years of age, approximately 90% of the endometrial epithelium harbours a driver mutation⁷⁹, whereas this is true of only about 1% of the colonic epithelium²⁴, despite the latter having a much higher somatic mutation rate^{24,79}. This difference is probably caused by the menstrual cycles of shedding and regrowth in the endometrium. Probably due to similar clonal expansion in development or ageing, about 6% of individuals harbour a 3–20-fold higher than average number of detectable SNVs in their brain³⁴. These varying proportions of clonally expanded cell populations probably reflect differences in tissue architecture, cell turnover, regeneration and selection pressures, but much is still unknown.

Although many driver mutations in normal tissues can be identical to those found in corresponding cancer types, their abundance and phenotypic consequence may differ profoundly as normal tissues may

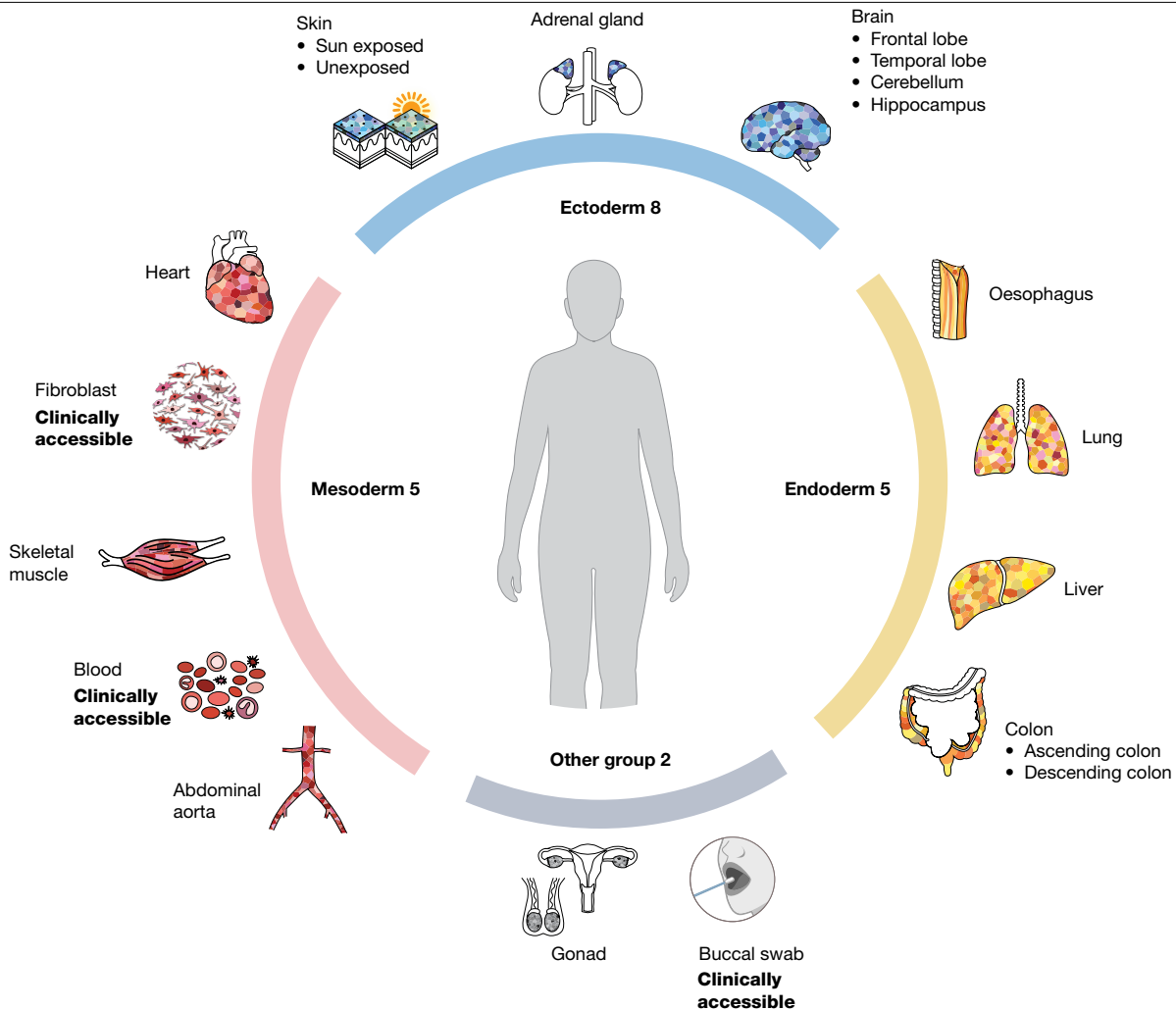


Fig. 2 | Tissue sampling. Overview of sampling from 19 primary tissue sites, spanning three developmental germ layers (endoderm, mesoderm and ectoderm) and germ cells. Although organs represent a mixture of cells derived from the germ layers (for example, skin epidermis (ectoderm) versus dermis

(mesoderm), and adrenal gland medulla (ectoderm) versus cortex (mesoderm)), we have indicated the major germ layer represented by each organ. Gonads represent germ cells and their supportive structures (mesoderm), whereas buccal swabs are variable mixtures of germ layers (mesoderm and ectoderm).

experience different selection pressures than cancer. For example, clones with *NOTCH1* mutations are exceedingly abundant in normal oesophageal epithelium, at even higher rates than oesophageal cancers⁸⁰. *NOTCH1*-mutant clones have a lower propensity of malignant transformation and even outcompete precancerous clones in the oesophagus^{81,82}. These observations suggest that characterizing the somatic mutation landscape in normal individuals will be important to understand the role of these mutations in pathological phenomena such as cancer.

Finally, somatic mutations can be used as intrinsic barcodes to create phylogenies and trace the ancestries of cells, such that it becomes possible to quantitatively study human development from somatic mutations ascertained in adult donors^{6,7,20,25,40,58–60,70,83}. This approach has been applied to studying embryogenesis, clonal expansions across the lifespan and the origins of childhood cancers⁸⁴. As the allele frequency of a mutation reflects the fraction of cells within a population that harbours it, this method can be used to quantitatively assess the contribution of embryonic progenitors to the adult body. Such studies have found that one of the two daughter cells of the zygote often has at least twice as many descendant cells as the other^{6,7,20,25,40,58–60,70,83,85,86}, probably due to cellular bottlenecks in embryogenesis, developmental cell death or migratory patterns, and confirming earlier observations in mice^{40,87}.

Together, these initial studies on somatic mutations in normal tissues have shown the variability of rates, patterns and selection of mutations across tissues. It is unknown, however, how variable these patterns are between individuals and how different types of somatic mutations are correlated with inherited genetic background, environmental exposures or other behavioural characteristics. In addition, mutation discovery is severely hampered in poorly mapped regions of the genome, including acrocentric chromosomes, centromeric and repetitive regions, and, hence, the mutational patterns in these regions are largely unknown. Thus, identification of the differences in mutational patterns between tissues and individuals, particularly in the context of specific organs^{21,26,29,30,34,88,89}, may have profound clinical implications.

Somatic mutations and disease

Somatic mutations can profoundly alter the phenotype of a cell and have been implicated in human diseases. Besides cancer, various other diseases and conditions can be a result of somatic mutations, including cardiovascular anomalies, immunological and neurological disorders^{26,30–34,52,90,91}. Of note, early somatic mutations can cause clonal expansions and alterations in the differentiation programs of precursor cells that subsequently can lead to paediatric cancers

and organ overgrowth^{84,92,93}. Among the first described instances of somatic mutagenesis, PI3K–AKT–mTOR pathway mutations involving the brain were associated with brain malformations leading to intractable epilepsy^{33,94,95}. Other examples are *NRAS* mutations leading to congenital melanocytic nevi⁹⁶ and *UBA1* mutations in haematopoietic stem cells⁹⁷ leading to VEXAS syndrome, a rare and severe inflammatory disorder. Somatic expansions of short tandem repeats in the brain can cause cell death and neurodegeneration⁹⁸, and underpin Huntington disease^{99,100}. Large SVs, including CNVs and MEIs, have also been implicated in neurodevelopmental and neurodegenerative disorders^{72,101,102}.

The effects of somatic mutations can be highly specific to the timing and tissue of origin. For example, an activating *PIK3CA* mutation acquired during development can lead to widespread overgrowth across organs and vascular malformations⁹¹. However, *PIK3CA* mutations acquired after development can lead to cavernomas in the brain¹⁰³ and are also a common driver mutation observed in normal colonic²⁴ and endometrial epithelium⁷⁹.

Clonal expansions can also indirectly lead to or influence other diseases²⁶. An example is clonal haematopoiesis of indeterminate potential (CHIP), characterized by a clonal expansion within the haematopoietic stem cell compartment driven by somatic mutations. CHIP is highly prevalent in the context of normal ageing²⁶. Besides acting as a potential cancer precursor clone, CHIP has been linked to various non-cancer diseases, such as an increased risk of cardiovascular disease¹⁰⁴ and infections¹⁰⁵.

Conversely, diseases can also select clones with certain adaptive somatic mutations. Recently, inflammatory bowel disease has been shown to lead to the preferential remodelling of the colonic epithelium with clones harbouring IL-17 and Toll-like receptor pathway mutations^{29,106}. Likewise, chronic liver disease selects for clones of hepatocytes that escape the toxicity imposed by the disease, notably, by recurrent, independent mutations in *FOXO1*, *CIDEA* and *GPAM*, which are all involved in lipid metabolism⁸⁹.

Together, research over the past years has shown that somatic evolution is ubiquitous in normal tissues and is fundamental to our understanding of the causes, mechanisms and consequences of disease, and the normal process of ageing.

The SMaHT Network

The SMaHT Network, funded by the NIH Common Fund, was established with the goal of transforming our understanding of how somatic variation in human cells influences biological processes. The SMaHT Network will accomplish this through the following aims: (1) generate a comprehensive dataset of somatic variants across human tissues (Fig. 2); (2) develop tools and technologies to optimize detection and characterization of various types of somatic variants; and (3) create a somatic mutation database that is widely used by researchers and the wider public, and interoperable with similar datasets.

The Network comprises five Genome Characterization Centres (GCCs), 14 Tool and Technology Development projects (TTDs), an Organizational Centre (OC), a Data Analysis Centre (DAC) and a Tissue Procurement Centre (TPC), and includes over 250 researchers from 52 institutions. The GCCs are tasked with producing a core dataset of somatic mutations for the SMaHT Network from multiple tissues collected by TPC, whereas TTDs are tasked with developing novel experimental assays and computational tools. The DAC will integrate the data generated by GCCs and TTDs to build the somatic mutation catalogue, data portal and the analysis work bench for the Network. The OC will coordinate the Network activities and focus on outreach efforts and building liaison with other genomics consortia. The SMaHT Network has implemented a set of policies (<https://smaht.org/policies/>), including a policy to allow external researchers to apply for associate membership of the Network.

Box 1

Criteria for SMaHT donors

Inclusion criteria:

- Donor over 18 years of age
- Collection can be completed within 24 h of cross-clamp or cardiac cessation

Exclusion criteria:

- History of HIV, HCV or HBV
- History of intravenous drug use in the past 5 years
- Chemotherapy or radiation treatment in the past 24 months
- Known chromosomal or genetic disorder
- Current positive blood cultures (sepsis)
- Active and/or metastatic cancer
- Diagnosed with multisystem organ failure
- Received organ or allogeneic bone marrow transplant
- Received whole-blood transfusion in 48 h before cross-clamp or cardiac cessation

Exclusion criteria for brain donation specifically:

- Cause of death related to penetrating brain injury or head trauma
- Brain dead or ventilator dependent for greater than 24 h

The tissues to be profiled by the Network include those arising from the three germ layers and germlines within the human body, which will give the opportunity to delineate early somatic mutations that are common across all tissues, as well as later mutations that are unique to certain tissues (Fig. 2). The TPC is partnering with multiple organ procurement organizations (OPOs) in the USA for the screening, authorization and recovery of tissues from post-mortem organ and tissue donors. Tissues will be collected following transplant recovery and include the ascending and descending colon, oesophagus, lung and liver (predominantly endoderm); blood, heart, aorta and skeletal muscle (predominantly mesoderm); and the brain, adrenal gland, sun-exposed and non-sun-exposed skin (predominantly ectoderm). We also aim to collect buccal swabs to assess the extent of the somatic mutation landscape that can be gleaned from clinically accessible tissues in living donors. To study mutagenesis in germ cells, we also aim to collect ovaries and testes. Finally, to enable various experimental techniques requiring live cells, we will derive fibroblast cultures from the dermis (skin). All tissues are requested to be recovered from each donor approached for the SMaHT tissue collection. The number and type of samples collected from each donor will vary based on donor authorization and eligibility (Box 1), but the goal is to recover as many tissues from a single donor as possible. To study the mechanisms and consequences of somatic mosaicism across the lifespan, these post-mortem donors will span the human adult age ranges from 18 to over 85 years. The race and ethnicity of donors are assessed using a single-question framework.

To maximize the scientific and clinical impact of the dataset, the TPC will collect a large amount of donor metadata during donation and biospecimen collection, building on practices developed for the Genotype-Tissue Expression (GTEx)¹⁰⁷ and developmental GTEx projects¹⁰⁸. De-identified donor-level data will include demographic information, medical history, sample-based laboratory test results and death circumstances. Sample-level data will include tissue type and location, ischaemic time and tissue metrics from pathology review. Pathology images will be made publicly available. When possible, tissue sampling will align with the common coordinate framework structure of other large-scale projects. For all of these biospecimens, sufficient fresh-frozen material will be collected and banked to enable all core assays as well as implementation of novel emerging technologies. Fixed samples for pathology review will be collected from adjacent sites to

Table 1 | Comparison between somatic mutation discovery methods

Feature	Bulk sequencing	Duplex sequencing	Clonal expansion	Single-cell WGA (PTA)	LCM clones
Applicability to any tissue	Yes	Yes	No	Yes	No
Mutation types discovered	All	SNVs and indels	All	All	All
Applicability to long read	Yes	Yes	Yes	Inefficient	Inefficient
Fraction of genome sampled (%)	100	30–100, depending on fragmentation method	100	Approximately 90 per cell; 100 across cells	100
Detection of early and clonally expanded mutations	Most	Minority	Depending on clone number	Depending on cell number	Depending on clone number
Overall mutation spectrum	No	Yes	Yes	Yes	Yes
Likely amount of artefacts	Small (10^{-4})	Very small (less than 10^{-8})	Small (10^{-4})	Some	Small (10^{-4})
Information on cell lineages	No	No	Yes	Yes	Yes
Advantages	Sensitive detection of high-frequency mutations of all types	Obtaining overall mutation spectrum even at low (0.5–2×) duplex coverage	Accurate mutation discovery at a single-cell level	Mutation discovery at a single-cell level in any tissue	Accurate mutation discovery at a single-cell level
Limitations	Need for high coverage; missing low-frequency mutations	Only SNVs and indels; missing high-frequency mutations at low (0.5–2×) coverage	Applicable to culturable or reprogrammable cells	Less than 50% sensitivity because of dropout and amplification artefacts	Applicable to tissues with visible clonal substructures

A combination of different methods can achieve comprehensive mutation discovery and accurate analysis. LCM, laser capture microdissection; WGA, whole-genome amplification.

the fresh-frozen specimens utilizing a standardized collection schema developed for each tissue type.

To pursue a demographically robust and evenly sex-distributed pool of donors, the SMAHT Network includes an ethical, legal and social implications project¹⁰⁹ consistent with the recommendations of the American Society for Human Genetics to address under-representation in human genomics studies with meaningful engagement of under-represented communities¹⁰⁹. This ethical, legal and social implications substudy engages geographically, racially, ethnically and socioculturally diverse stakeholders, which include family decision-makers, tissue requesters, community advisory board members and multi-disciplinary specialty committee members throughout the entire duration of the SMAHT Network. Feedback from community stakeholders will be leveraged to inform communication and enrolment efforts as well as dissemination of study findings.

The SMAHT Network is uniquely positioned to collaborate with many other large consortia and programmes. These include the Human Pangenome Reference Consortium⁹, to leverage methods for constructing haplotype-phased genome assemblies; the Impact of Genomic Variation on Function Consortium¹¹⁰, to understand the functional consequences of genetic variation; the developmental GTEx project¹⁰⁸, to access datasets from tissues at early developmental stages; the Human Tumor Analysis Network and PreCancer Atlas, to further understand the progression from normal cells to tumour cells through somatic mutations; and PsychENCODE¹¹¹, to inform on the phenotypic consequences of brain somatic mosaicism. These collaborations will enrich the individual studies and, ultimately, through data integration and cross-network analyses, further enhance our understanding of the context and consequences of somatic mutations.

Producing the somatic mutation catalogue

To produce the first phase of the somatic mutation catalogue, the SMAHT Network will strike a balance between standard genomic assays, productionized and applied uniformly by the GCCs to all tissues, and bespoke assays developed by the TTD projects, focusing on novel

technological approaches. As part of the initial phase of the SMAHT project, benchmarking efforts are nearing completion, using both primary human tissues and cell lines. We have used this benchmarking to determine optimal sequencing coverage, compare the accuracy of variant calling algorithms, and evaluate the utility of long-read and short-read sequencing data generated on diverse sequencing platforms from multiple GCCs.

The GCCs will deploy three core assays across all tissue specimens that meet quality thresholds: deep short-read whole-genome sequencing (WGS; over 300× coverage), long-read WGS (over 30× coverage) sequencing and RNA sequencing (over 50 million reads). The deep short-read WGS will enable the discovery of high allele frequency somatic mutations across tissues acquired early in embryogenesis, as well as discovery of the large clonal expansions arising later in life. As these core assays will be performed on bulk tissues, composed of heterogeneous cell types, only mutations with a relatively high variant allele frequency (above 1–2%) will be accurately detectable at the proposed depth of sequencing. The long-read WGS will facilitate the detection of complex SVs, MEIs and variants in complex genetic loci that have been challenging to accurately study using short-read data, such as the MHC region, centromeres, telomeres, acrocentric DNA including ribosomal DNA and other tandem-repeat regions of the genome. Ultra-long-read sequencing will enable us to generate near telomere-to-telomere donor-specific reference genome assemblies for at least 50 donors and through reducing misalignment, enhance the discovery of diverse types of variants within an individual⁵⁰, including complex somatic SVs and other mutations in previously unmappable regions of the genome¹¹². Finally, RNA sequencing may allow us to assess transcriptional consequences of early mutations and late clonal expansions, as well as, by comparison with single-cell RNA sequencing atlases¹¹³, cell-type composition of heterogeneous tissues.

In addition to these core assays, GCCs will deploy three approaches specifically designed to profile low-frequency somatic mutation: duplex sequencing, single-cell WGS and transcript-based detection of mutations. These technologies, although published and well-tested,

Box 2

Experimental and analytical methods adopted by the SMaHT Network

Core assays

- Short-read DNA sequencing (Illumina)
- Short-read RNA sequencing (Illumina)
- Long-read DNA sequencing (PacBio and Oxford Nanopore Technologies)
- Long-read full-length transcript sequencing (PacBio)

Extended assays

- Single-cell DNA sequencing with PTA
- Duplex sequencing: concatenating original duplex for error correction, nanorate sequencing (NanoSeq), Tn5-duplex-seq (CompDuplex-seq and VISTA-seq) or Ultima Genomics ppmSeq

Scale-up approaches and technological developments

Single-molecule sequencing

- Hairpin duplex-enhanced fidelity sequencing (HiDEF-seq)

Single-cell duplex sequencing

- scNanoSeq and scUduplex-seq

Structural variants detection

- Single-cell full-length RNA sequencing for MEI detection
- Single-cell, mini-bulk mobile element-targeted sequencing (PTA-HAT-seq)
- Single-cell total RNA sequencing miniaturized sequencing (STORM-seq)

- Strand-seq

Spatial variant detection

- Slide-tags

Genotype-to-phenotype multi-omics

For epigenome:

- Single-cell 3C-seq protocols (for example, Dip-C)
- Assay for transposase-accessible chromatin (ATAC-seq)
- Genotyping of targeted loci with chromatin accessibility (GoT-ChA and GoT-EpiM)

- Fiber-seq (single-molecule chromatin accessibility, nucleosome occupancy, transcription factor occupancy, RNA polymerase occupancy, CpG methylation and genetic variant identification)
- Var-CUT&Tag (enrichment and epigenetic annotation of variants in regulatory elements)
- Strand-seq (nucleosome occupancy and structural variants in the same cell)
 - For transcriptome:
- ResolveOme method
- GoT-ChA-RNA
- Duplex-seq-based large-scale single-cell dual omics profiling
 - For proteome:
- GoT-ChA-Pro
- Computational tools*
- Donor-specific reference genome assemblies (paired Fiber-seq, ultra-long Oxford Nanopore Technologies and Hi-C data)
- Hybrid approach for single-cell WGA-RUFUS (a reference-free, kmer-based variant detection algorithm)

Infrastructure

- Centralized three germ layer tissue collection banking with donor metadata
- Automated analysis pipelines for quality control
- A variant catalogue containing a curated set of annotated somatic mutations
- Computational pipelines with multi-omics platforms
- Human pangenome visualization with somatic mutation detection pipelines
- A cloud-based infrastructure and a web portal

The SMaHT Network will apply a set of standard approaches as well as benchmark new methods to improve detection, analysis and functional annotation of somatic variants across tissues and individuals.

represent recent innovations and have not yet been systematically deployed across sequencing centres or applied at large scale.

As conventional DNA sequencing platforms have a non-trivial sequencing error rate (in the order of 1 in 1,000–10,000), a putative mutation needs to be detected in multiple independent reads to assure it is not artefactual. However, by sequencing both the forward and the reverse strands of each individual DNA duplex molecule, this error rate is drastically reduced. As the reduced error rate is much lower than the expected number of somatic mutations in most tissues, an average mutation burden and mutational profile can be obtained by shallow genome-wide duplex coverage ($0.5-2\times$)^{63,114}. Duplex sequencing of bulk tissue samples is well suited to finding average mutation burdens and spectra of SNVs and indels within cell populations, but the low depth generally precludes discovery of somatic CNVs and SVs, or the precise inference of variant allele frequency of specific mutations.

Even with a reduced sequencing error rate, bulk DNA sequencing will average out the mutational patterns of all cells and does not allow assessment of the variability of mutational patterns between cells or the reconstruction of cell lineages. Instead, sequencing the DNA of single cells or single-cell-derived clones will enable the most detailed discovery of somatic mutations. This can be achieved either by expanding single cells in vitro^{6,25,26,52,59} or laser capture microdissection to isolate naturally occurring clonal populations of cells^{7,24,79,88,115}.

Alternatively, direct single-cell DNA sequencing is applicable to all cell types, including non-dividing cells. However, whole-genome

amplification can cause allelic or locus dropout, uneven coverage across the genome and artefactual variants introduced during biochemical amplification. The direct library preparation (DLP+)^{116,117} method avoids whole-genome amplification and allows for the accurate detection of CNVs at the single-cell level and other mutations at the population level. The primary template-directed amplification (PTA)^{30,118} method offers a substantial improvement in data quality over previous single-cell amplification methods, resulting in more uniform genome coverage and fewer artefactual variants. A more recent version of PTA, the ResolveOme approach, profiles both the transcriptome and the genome from the same single cell. If validated, this approach will represent a major advance in allowing new mutation detection and cellular phenotyping at the same time. Profiling somatic mutations in single cells will enable us to characterize mutational patterns and associations between mutation types and to reconstruct phylogenetic trees of normal cells across tissues. In cases of polyploid cells, the variant allele frequencies of somatic mutations may deviate from the expected 0.5 and ploidy will need to be taken in consideration in downstream analyses.

Finally, at least some somatic mutations can be inferred from RNA^{119–121}. Methods that allow for the interrogation of the full-length transcriptome in single cells, such as Smart-seq3 (ref. 122) or STORM-seq¹²³, can facilitate the detection of somatic mutations, such as SNVs, indels and fusion genes within transcribed regions of the genome. This allows assessing cell-type specificity for clonal expansion of certain

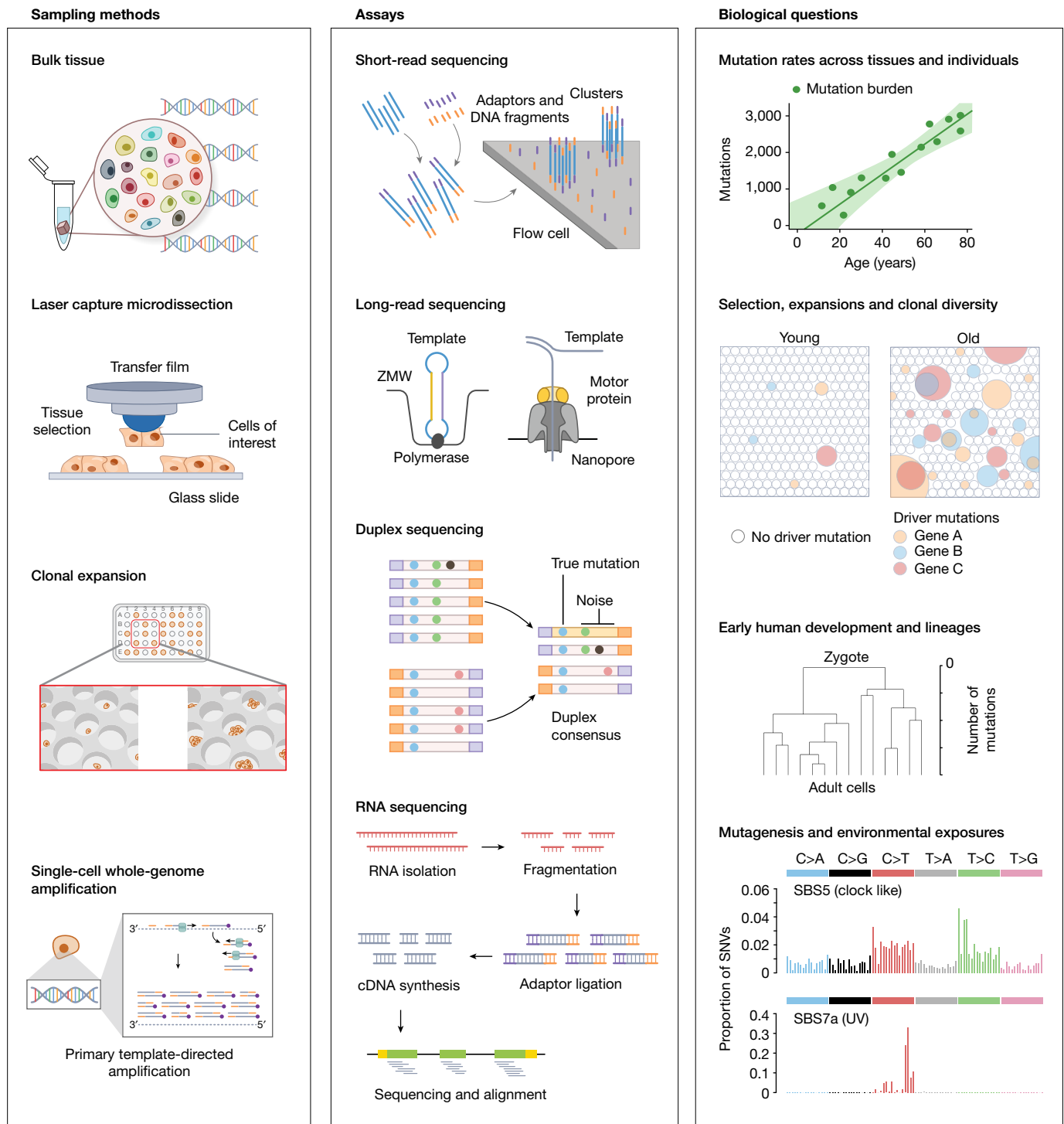


Fig. 3 | Methods, assays and questions. Overview of sampling methods and sequencing assays deployed in the SMaHT Network, as well as the biological questions, outcomes and inferred mutational patterns from downstream analyses of the catalogue of somatic mutations across normal tissues, including

mutation rates or burdens, selection, lineage tracing and mutational signatures (reference signatures were obtained from <https://cancer.sanger.ac.uk/signatures>)⁴⁴. ZMW, zero-mode waveguide.

genetic variants. Furthermore, STORM-seq enables quantification of transposable element expression at single-cell resolution, which has been shown to be challenging with other single-cell RNA sequencing methods¹²⁴. The single-cell data also provide references for a more precise deconvolution of cell types in bulk tissues.

Each of these methods for the detection of somatic mosaic variants presents its own advantages and disadvantages and thus they are complementary (Table 1). For example, although genome-wide

duplex sequencing has a lower sequencing error rate and excels at population-level inferences of patterns of short mutations acquired during the entire lifespan, the low depth precludes detection of the precise allele frequency of a specific variant. Bulk sequencing at medium-high coverage (300×) will only detect variants at a sufficiently high frequency (that is, 1–2%) in tissues, which are mostly acquired in early embryogenesis. Single-cell sequencing can in principle detect all variants present in a single cell and allow reconstruction of cell phylogenies,

but it requires significant costs and efforts to address genome amplification artefacts. RNA-based mutation discovery allows for direct integration of mutations with transcriptomic information but is naturally confined to expressed regions of the genome. Together, these genomic assays function as complementary techniques to detect somatic mutations and will enable the robust interrogation of mutational patterns across human tissues.

Areas of technological development

As new technologies to interrogate somatic mutations with high resolution or sensitivity are constantly emerging, a large part of the SMAHT Network is devoted to developing new tools and technologies (Box 2). The first area of innovation aims to increase the accuracy of mutation detection in single cells or molecules by further reducing background noise. For single-cell WGS, a limited cloning step to create small pools of cells can reduce allelic dropout and amplification artefacts. In parallel, the SMAHT Network aims to reduce the error rate of amplification and sequencing for single cells and molecules through various adaptations of duplex sequencing technologies^{63,125–127}. These approaches will allow for the interrogation of the landscape of somatic variation in single cells and complex multicellular tissues with high precision, which is crucial to study tissues without large-scale expansions.

Second, the SMAHT Network aims to increase the sensitivity of SV detection to single molecules or cells. As SVs extend beyond the length of a typical short read, long-read sequencing unlocks SV detection across the genome, especially for MEIs and other rearrangements in repetitive regions^{128,129}. However, many single-cell DNA amplification approaches result in short fragments. Therefore, we are applying long-read sequencing to clonal populations such as induced pluripotent stem cell lines, which have been used²⁵ in lieu of single cells for lineage reconstructions as they can be expanded and analysed by bulk sequencing, avoiding in vitro DNA amplification. In addition, MEIs can be cost-effectively assessed by target enrichment assays as new insertions share conserved sequences in each transposon subfamily. We are developing targeted detection of MEI insertions by utilizing Cas9-targeted long-read sequencing¹³⁰ and PTA-amplified micro-bulk or single cells⁷³. These efforts will unlock the study of SVs and MEIs in all tissues and across the lifespan, even in the absence of clonal expansions.

Third, the SMAHT Network will develop scalable platforms that can perform variant detection spatially in human tissues, through single-cell DNA and RNA sequencing with resolved spatial barcodes^{131,132}. This will allow us to study the prevalence and extent of clonal expansions across ages and tissues, especially in organs without a clearly organized tissue architecture.

An outstanding question is the effect of specific somatic mutations on the phenotype of the cells that harbour them. Although certain mutations are under positive selection and lead to clonal expansions, how these mutations alter cellular phenotypes is mostly unknown. The consequence of a mutation can be assessed by combining mutational readouts, either through genotyping of specific mutations^{133,134} or genome sequencing, in combination with functional readouts of cells, such as the transcriptome, proteome, epigenome, methylome and the chromatin accessibility landscape^{135–140}. Interpreting the phenotypic effects of somatic mutations will greatly benefit our understanding of the clinical consequences.

The efforts in tool and technological development within the SMAHT Network are focused on improving precision in somatic mutation detection and interpretation at scale, each addressing vital shortcomings of current assays, with a goal to productionize and deploy many of these within the Network at large. After the development phase, the precise extent and scope of the deployment of these assays across the SMAHT tissues and donors will depend on the cost, scalability and priorities of the Network.

Integration and analysis of data

The low variant allele frequency of mosaic variants brings unique challenges in bioinformatic analysis¹⁴¹, and we expect that novel computational methods and tools are needed to fully analyse the data and to increase the sensitivity and specificity of variant detection. Somatic mutation detection algorithms developed in cancer genomics are often inadequate for detecting variants with allele fractions less than 2–5% and simply increasing the depth of sequencing is not cost-effective. Thus, more sophisticated machine learning algorithms that efficiently incorporate various local features near candidate variants may prove useful^{136–138,142}.

Other challenges include optimal integration of long-read and short-read data, inference of lineage relationships based on bulk and single-cell data, and effective strategies for integrative and comparative analysis of samples across the tissues and across individuals. An important aspect of our analysis will be the use of donor-specific diploid genomes assembled using short Illumina, long PacBio and ultra-long Nanopore and Hi-C reads. Alignment to the donor-specific reference genome¹³⁵ will allow for more accurate variant identification, especially in repetitive regions, as well as for examination of allele-specific transcriptional and epigenetic modulations associated with genetic variants.

The SMAHT DAC will lead an effort to collect, curate and analyse the vast amount of multi-modal data generated on multiple platforms and to create a data resource for the scientific community. The DAC will ensure high data standards with various quality control steps and compile extensive metadata describing experimental and data processing protocols, following the FAIR (Findable, Accessible, Interoperable and Reusable) guidelines¹⁴³. Scalable and cost-effective analytical workflows will be implemented on a cloud platform with full provenance and docker images to enable reproducibility of the analysis output.

The data generated by the consortium will be made available to the wider scientific community via a user-friendly and secure web portal (<https://data.smaht.org>). This portal will feature: (1) a reference catalogue of somatic variants that can be searched (for example, by locus, tissue or phenotypic features such as age) and annotated with information from other genomics databases; (2) a workbench that enables users to apply the computational pipelines developed by the SMAHT Network to their own data; and (3) data visualization tools including a multi-scale browser that allows users to navigate the data from a genome-level view to the sequencing read-level view. Visual inspection of variants using such a browser will be particularly helpful in assessing their quality, and the annotations will enable rapid identification of variants that may be functionally relevant.

Conclusion

The SMAHT Network aims to produce a comprehensive reference catalogue of somatic mutations, across tissues and individuals, by harnessing the full potential of many different genomic assays, including short-read and long-read bulk WGS, duplex sequencing, ultra-long-read sequencing, single-cell DNA sequencing and RNA sequencing (Fig. 3). The Network will develop new tools and technologies to increase our ability to detect somatic mutations as well as infer their phenotypic consequences at greater resolution. All of these various data modalities will be integrated, analysed and released to the research community and wider public.

An extensive catalogue of somatic mutations will reveal mutational patterns, rates and signatures across tissues, allowing us to infer the biological and molecular processes that govern somatic mutagenesis and their adaptive and maladaptive consequences for development and disease (Fig. 3). Our assays can inform on mutations under selection in tissues, which result in clonal expansions and potentially tissue dysfunction. Single-cell analyses added to the bulk readouts will further

allow us to generate cellular phylogenies of human development, infer embryonic differentiation dynamics and improve our future assessment of de novo germline mutations.

Delineating the full extent of somatic mosaicism greatly exceeds the scope of the Human Genome Project. A typical cell may acquire hundreds to thousands of somatic mutations in a lifetime. There are trillions of cells in a human body and so the total number of somatic mutations acquired in a single individual may well exceed quadrillions (10^{15}), millions of times the size of the human genome. Beyond cataloguing somatic variation across tissues, the SMAHT Network provides the opportunity to understand the causes, patterns and consequences of somatic mutations in normal cells, and provide a crucial comparison baseline for disease research. The efforts of the SMAHT Network will substantially contribute to our insights into the role of somatic variation in health, ageing and disease.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Durbin, R. M. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
3. Wang, T. et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
4. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
5. Saini, N. & Gordenin, D. A. Somatic mutation load and spectra: a record of DNA damage and repair in healthy human cells. *Environ. Mol. Mutagen.* **59**, 672–686 (2018).
6. Park, S. et al. Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature* **597**, 393–397 (2021).
By amplifying cells into clones and subsequent sequencing, this study reconstructs embryonic dynamics through somatic mutation patterns.
7. Coorens, T. H. H. et al. Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**, 387–392 (2021).
This study reconstructs phylogenetic trees of human development through the detection of somatic mutations in many different tissues of the same donors.
8. Evrony, G. D. et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).
This study represents a first foray into single-cell DNA sequencing to uncover somatic mutations in single neurons.
9. Bae, T. et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* **359**, 550–555 (2018).
Using somatic SNVs discovered in human brain progenitor cell clones, this study obtains a human embryonic lineage tree and estimates mutation frequencies at pregastrulation and neurogenesis.
10. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14**, 307–320 (2013).
11. Martínez-Glez, V. et al. A six-attribute classification of genetic mosaicism. *Genet. Med.* **22**, 1743–1757 (2020).
12. Salik, J. J., Schmitt, M. W. & Loeb, L. A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.* **19**, 269–285 (2018).
13. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
14. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
15. Laurie, C. C. et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44**, 642–650 (2012).
16. Jacobs, K. B. et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44**, 651–658 (2012).
17. Abyzov, A. et al. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* **492**, 438–442 (2012).
Using induced pluripotent stem cell lines to perform clonal analysis, this study outlines somatic CNVs in skin fibroblasts from multiple families.
18. McConnell, M. J. et al. Mosaic copy number variation in human neurons. *Science* **342**, 632–637 (2013).
19. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
This study is one of the first to use single-cell-derived organoids across tissues in humans to demonstrate a variability in somatic mutation rate.
20. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
21. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
22. Milholland, B. et al. Differences between germline and somatic mutation rates in humans and mice. *Nat. Commun.* **8**, 15183 (2017).
23. Zhang, L. et al. Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl Acad. Sci. USA* **116**, 9014–9019 (2019).
24. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
25. Fasching, L. et al. Early developmental asymmetries in cell lineage trees in living individuals. *Science* **371**, 1245–1248 (2021).
Reconstructing lineage trees via somatic mutations in living humans, this study demonstrates that embryonic lineages are often asymmetric.

26. Mitchell, E. et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
27. Jourdon, A., Fasching, L., Scuderi, S., Abyzov, A. & Vaccarino, F. M. The role of somatic mosaicism in brain disease. *Curr. Opin. Genet. Dev.* **65**, 84–90 (2020).
28. Breuss, M. W. et al. Somatic mosaicism reveals clonal distributions of neocortical development. *Nature* **604**, 689–696 (2022).
29. Olafsson, S. et al. Somatic evolution in non-neoplastic IBD-affected colon. *Cell* **182**, 672–684.e11 (2020).
30. Miller, M. B. et al. Somatic genomic changes in single Alzheimer's disease neurons. *Nature* **604**, 714–722 (2022).
31. Heimlich, J. B. & Bick, A. G. Somatic mutations in cardiovascular disease. *Circ. Res.* **130**, 149–161 (2022).
32. Poduri, A. et al. Somatic activation of AKT3 causes hemispheric developmental brain malformations. *Neuron* **74**, 41–48 (2012).
33. Chung, C. et al. Comprehensive multi-omic profiling of somatic mutations in malformations of cortical development. *Nat. Genet.* **55**, 209–220 (2023).
34. Bae, T. et al. Analysis of somatic mutations in 131 human brains reveals aging-associated hypermutability. *Science* **377**, 511–517 (2022).
Using data collected by the Brain Somatic Mosaicism Network, this paper describes types and frequencies of early somatic mutations in the human brain.
35. Lim, E. T. et al. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat. Neurosci.* **20**, 1217–1224 (2017).
36. Baldassari, S. et al. Dissecting the genetic basis of focal cortical dysplasia: a large cohort study. *Acta Neuropathol.* **138**, 885–900 (2019).
37. D'Gama, A. M. & Walsh, C. A. Somatic mosaicism and neurodevelopmental disease. *Nat. Neurosci.* **21**, 1504–1514 (2018).
38. Evans, M. A. & Walsh, K. Clonal hematopoiesis, somatic mosaicism, and age-associated disease. *Physiol. Rev.* **103**, 649–716 (2023).
39. Park, J. S. et al. Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat. Commun.* **10**, 3090 (2019).
40. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
Using single-cell-derived organoids of mouse tissues, this paper demonstrates the use of SNVs for in vivo lineage tracing and is an early report of developmental asymmetry.
41. Pleasance, E. D. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
42. Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
43. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
44. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
45. Drews, R. M. et al. A pan-cancer compendium of chromosomal instability. *Nature* **606**, 976–983 (2022).
46. Macintyre, G. et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.* **50**, 1262–1270 (2018).
47. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
48. Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
By studying many tissues from the same donor, this study demonstrates the variability in mutation burden and signatures across human tissues.
49. Choudhury, S. et al. Somatic mutations in single human cardiomyocytes reveal age-associated DNA damage and widespread oxidative genotoxicity. *Nat. Aging* **2**, 714–725 (2022).
50. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
51. 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 demonstrates the abundance of clones harbouring cancer-driver mutations in a normal tissue using ultra-deep sequencing.
52. Machado, H. E. et al. Diverse mutational landscapes in human lymphocytes. *Nature* **608**, 724–732 (2022).
53. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
54. Li, R. et al. A body map of somatic mutagenesis in morphologically normal human tissues. *Nature* **597**, 398–403 (2021).
55. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
56. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
57. Vöhringer, H., Hoeck, A. V., Cuppen, E. & Gerstung, M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. *Nat. Commun.* **12**, 3628 (2021).
58. Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
59. Spencer Chapman, M. et al. Lineage tracing of human development through somatic mutations. *Nature* **595**, 85–90 (2021).
60. Bizzotto, S. et al. Landmarks of human embryonic development inscribed in somatic mutations. *Science* **371**, 1249–1253 (2021).
61. Bizzotto, S. & Walsh, C. A. Genetic mosaicism in the human brain: from lineage tracing to neuropsychiatric disorders. *Nat. Rev. Neurosci.* **23**, 275–286 (2022).
62. Coorens, T. H. H. et al. Inherent mosaicism and extensive mutation of human placentas. *Nature* **592**, 80–85 (2021).
63. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
This study develops ultra-low error rates for duplex sequencing that allows the interrogation of somatic mutations at the single-read level.

64. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
65. Bloom, J. C., Loehr, A. R., Schimenti, J. C. & Weiss, R. S. Germline genome protection: implications for gamete quality and germ cell tumorigenesis. *Andrology* **7**, 516–526 (2019).
66. Maklakov, A. A. & Immler, S. The expensive germline and the evolution of ageing. *Curr. Biol.* **26**, R577–R586 (2016).
67. Cai, X. et al. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep.* **8**, 1280–1289 (2014).
68. Sun, C. et al. Mapping the complex genetic landscape of human neurons. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.03.07.531594> (2023).
69. Loh, P. R. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
70. Evrony, G. D. et al. Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49–59 (2015).
- This paper represents early work using single-cell somatic mutation signals to reconstruct cell lineages.**
71. Erwin, J. A. et al. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat. Neurosci.* **19**, 1583–1591 (2016).
72. Zhu, X. et al. Machine learning reveals bilateral distribution of somatic L1 insertions in human neurons and glia. *Nat. Neurosci.* **24**, 186–196 (2021).
73. Zhao, B. et al. Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. *PLoS Genet.* **15**, e1008043 (2019).
74. Shukla, R. et al. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**, 101–111 (2013).
75. Nam, C. H. et al. Widespread somatic L1 retrotransposition in normal colorectal epithelium. *Nature* **617**, 540–547 (2023).
76. Shah, N. M. et al. Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. *Nat. Genet.* **55**, 631–639 (2023).
77. Gebrie, A. Transposable elements as essential elements in the control of gene expression. *Mob. DNA* **14**, 9 (2023).
78. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
79. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
80. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
81. Colom, B. et al. Mutant clones in normal epithelium outcompete and eliminate emerging tumours. *Nature* **598**, 510–514 (2021).
82. Abby, E. et al. Notch1 mutations drive clonal expansion in normal esophageal epithelium but impair tumor growth. *Nat. Genet.* **55**, 232–245 (2023).
83. Choi, S. H., Ku, E. J., Choi, Y. A. & Oh, J. W. Grave-to-cradle: human embryonic lineage tracing from the postmortem body. *Exp. Mol. Med.* **55**, 13–21 (2023).
84. Coorens, T. H. H. et al. Embryonal precursors of Wilms tumor. *Science* **366**, 1247–1251 (2019).
85. Kwon, S. G. et al. Asymmetric contribution of blastomere lineages of first division of the zygote to entire human body using post-zygotic variants. *Tissue Eng. Regen. Med.* **19**, 809–821 (2022).
86. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).
87. Zernicka-Goetz, M. First cell fate decisions and spatial patterning in the early mouse embryo. *Semin. Cell Dev. Biol.* **15**, 563–572 (2004).
88. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
89. Ng, S. W. K. et al. Convergent somatic mutations in metabolism genes in chronic liver disease. *Nature* **598**, 473–478 (2021).
90. Sherman, M. A. et al. Large mosaic copy number variations confer autism risk. *Nat. Neurosci.* **24**, 197–203 (2021).
91. Keppler-Noreuil, K. M. et al. PIK3CA-related overgrowth spectrum (PROS): diagnostic and testing eligibility criteria, differential diagnosis, and evaluation. *Am. J. Med. Genet. A* **167A**, 287–295 (2015).
92. Custers, L. et al. Somatic mutations and single-cell transcriptomes reveal the root of malignant rhabdoid tumours. *Nat. Commun.* **12**, 1407 (2021).
93. Pilet, J. et al. Preneoplastic liver colonization by 11p15.5 altered mosaic cells in young children with hepatoblastoma. *Nat. Commun.* **14**, 7122 (2023).
94. Checchi, R. et al. Detection of brain somatic mutations in focal cortical dysplasia during epilepsy presurgical workup. *Brain Commun.* **5**, fca174 (2023).
95. Lee, J. H. et al. De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. *Nat. Genet.* **44**, 941–945 (2012).
96. Kinsler, V. A. et al. Multiple congenital melanocytic nevi and neurocutaneous melanosis are caused by postzygotic mutations in codon 61 of NRAS. *J. Invest. Dermatol.* **133**, 2229–2236 (2013).
97. Beck, D. B. et al. Somatic mutations in UBA1 and severe adult-onset autoinflammatory disease. *N. Engl. J. Med.* **383**, 2628–2638 (2020).
98. Higham, C. F., Morales, F., Cobbold, C. A., Haydon, D. T. & Monckton, D. G. High levels of somatic DNA diversity at the dystrophic dystrophy type 1 locus are driven by ultra-frequent expansion and contraction mutations. *Hum. Mol. Genet.* **21**, 2450–2463 (2012).
99. De Rooij, K. E., De Koning Gans, P. A., Roos, R. A., Van Ommen, G. J. & Den Dunnen, J. T. Somatic expansion of the (CAG)_n repeat in Huntington disease brains. *Hum. Genet.* **95**, 270–274 (1995).
100. Swami, M. et al. Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum. Mol. Genet.* **18**, 3039–3047 (2009).
101. Maury, E. A. et al. Schizophrenia-associated somatic copy-number variants from 12,834 cases reveal recurrent NRXN1 and ABCB1 disruptions. *Cell Genom.* **3**, 100356 (2023).
102. Kim, J. et al. Prevalence and mechanisms of somatic deletions in single human neurons during normal aging and in DNA repair disorders. *Nat. Commun.* **13**, 5918 (2022).
103. Ren, A. A. et al. PIK3CA and CCM mutations fuel cavernomas through a cancer-like mechanism. *Nature* **594**, 271–276 (2021).
104. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
105. Zekavat, S. M. et al. Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat. Med.* **27**, 1012–1024 (2021).
106. Nanki, K. et al. Somatic inflammatory gene mutations in human ulcerative colitis epithelium. *Nature* **577**, 254–259 (2020).
107. Carithers, L. J. et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreserv. Biobank* **13**, 311–319 (2015).
108. Coorens, T. H. H. et al. The human and non-human primate developmental GTEx projects. *Nature* **637**, 557–564 (2025).
109. Lemke, A. A. et al. Addressing underrepresentation in genomics research through community engagement. *Am. J. Hum. Genet.* **109**, 1563–1571 (2022).
110. IGVF Consortium. Deciphering the impact of genomic variation on function. *Nature* **633**, 47–57 (2024).
111. Akbarian, S. et al. The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015).
112. Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
113. Regav, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
114. Cheng, A. P. et al. Error-corrected flow-based sequencing at whole-genome scale and its application to circulating cell-free DNA profiling. *Nat. Methods* **22**, 973–981 (2025).
115. Wang, Y. et al. APOBEC mutagenesis is a common process in normal human small intestine. *Nat. Genet.* **55**, 246–254 (2023).
116. Zahn, H. et al. Scalable whole-genome single-cell library preparation without preamplification. *Nat. Methods* **14**, 167–173 (2017).
117. Laks, E. et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179**, 1207–1221.e22 (2019).
118. Gonzalez-Pena, V. et al. Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc. Natl Acad. Sci. USA* **118**, e2024176118 (2021).
119. Yizhak, K. et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, eaaw0726 (2019).
120. 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).
121. Gao, T. et al. A Pan-tissue survey of mosaic chromosomal alterations in 948 individuals. *Nat. Genet.* <https://doi.org/10.1038/s41588-023-01537-1> (2023).
122. Hagemann-Jensen, M. et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).
123. Johnson, B. K. et al. Efficient profiling of total RNA in single cells with STORM-seq. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.03.14.484332> (2025).
124. Shao, W. & Wang, T. Transcript assembly improves expression quantification of transposable elements in single-cell RNA-seq data. *Genome Res.* **31**, 88–100 (2021).
125. Xing, D., Tan, L., Chang, C. H., Li, H. & Xie, X. S. Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. *Proc. Natl Acad. Sci. USA* **118**, e2013106118 (2021).
126. Bae, J. H. et al. Single duplex DNA sequencing with CODEC detects mutations with high sensitivity. *Nat. Genet.* **55**, 871–879 (2023).
127. Liu, M. H. et al. DNA mismatch and damage patterns revealed by single-molecule sequencing. *Nature* **630**, 752–761 (2024).
128. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
129. Zhou, W. et al. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res.* **48**, 1146–1163 (2020).
130. McDonald, T. L. et al. Cas9 targeted enrichment of mobile elements using nanopore sequencing. *Nat. Commun.* **12**, 3586 (2021).
131. Zhao, T. et al. Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature* **601**, 85–91 (2022).
132. Russell, A. J. C. et al. Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. *Nature* <https://doi.org/10.1038/s41586-023-06837-4> (2023).
133. Nam, A. S. et al. Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature* **571**, 355–360 (2019).
134. Izzo, F. et al. Mapping genotypes to chromatin accessibility profiles in single cells. *Nature* **629**, 1149–1157 (2024).
135. Feusier, J. et al. Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res.* **29**, 1567–1577 (2019).
136. Dou, Y. et al. Accurate detection of mosaic variants in sequencing data without matched controls. *Nat. Biotechnol.* **38**, 314–319 (2020).
137. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
138. Ding, J. et al. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* **28**, 167–175 (2011).
139. Gaiti, F. et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **569**, 576–580 (2019).
140. Grimes, K. et al. Cell-type-specific consequences of mosaic structural variants in hematopoietic stem and progenitor cells. *Nat. Genet.* **56**, 1134–1146 (2024).
141. Dou, Y., Gold, H. D., Luquette, L. J. & Park, P. J. Detecting somatic mutations in normal cells. *Trends Genet.* **34**, 545–557 (2018).
142. Yang, X. et al. Control-independent mosaic single nucleotide variant detection with DeepMosaic. *Nat. Biotechnol.* **41**, 870–877 (2023).
143. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
144. Osorio, F. G. et al. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell Rep.* **25**, 2308–2316.e4 (2018).
145. 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).
146. Coorens, T. H. H. et al. The somatic mutation landscape of normal gastric epithelium. *Nature* **640**, 418–426 (2025).

Acknowledgements This research is supported by the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under awards U24 MH133204, U24 NS132103, UG3 NS132024, UG3 NS132061, UG3 NS132084, UG3 NS132105, UG3 NS132127, UG3 NS132128, UG3 NS132132, UG3 NS132134, UG3 NS132135, UG3 NS132136, UG3 NS132138, UG3 NS132139, UG3 NS132144, UG3 NS132146, UM1 DA058219, UM1 DA058220, UM1 DA058229, UM1 DA058230, UM1 DA058235 and UM1 DA058236. E.E.E. and C.A.W. are investigators of the Howard Hughes Medical Institute.

Author contributions T.H.H.C., J.W.O., E.A.L. and F.M.V. wrote the manuscript, with input, review and feedback from all other authors. J.W.O., T.H.H.C., Y.A.C., N.S.L. and A.V. made the figures. A.C.L. was substantially involved in UM1 DA058219, UM1 DA058220, UM1 DA058229, UM1 DA058230, UM1 DA058235 and UM1 DA058236, consistent with the role as Program Officer; and is also the NIH Working Group Coordinator, and has involvement with the remaining awards, consistent with this role.

Competing interests F.C. is an academic founder of Curio Biosciences and Doppler Biosciences, and scientific advisor for Amber Bio; F.C.'s interests were reviewed and managed by the Broad Institute in accordance with their conflict-of-interest policies. G.G. receives research funds from IBM, Pharmacyclics/Abbvie, Bayer, Genentech, Calico, Ultima Genomics, Inocras, Google, Kite and Novartis; is an inventor on patent applications filed by the Broad Institute related to MSMutSig, POLYSOLVER, SignatureAnalyzer-GPU, MSeYE and MinimuMM-seq; is a founder, consultant and holds privately held equity in Scorpion Therapeutics and PreDICTA Biosciences; and was a consultant to Merck, all unrelated to the present work. E.E.E. is a scientific advisory board member of Variant Bio. C.Z. is a co-founder and equity holder of Pioneer Genomics and reports that Baylor College of Medicine filed a patent application related to the CompDuplex-seq or CompDup method. P.N. reports research grants from Allelica, Amgen, Apple, Boston Scientific, Cleerly, Genentech/Roche, Ionis, Novartis and Silence Therapeutics; personal fees from Allelica, Apple, AstraZeneca, Bain Capital, Blackstone Life Sciences, Bristol Myers Squibb, Creative Education Concepts, CRISPR Therapeutics, Eli Lilly & Co, Esperion Therapeutics, Foresite Capital, Foresite Labs, Genentech/Roche, GV, HeartFlow, Magnet Biomedicine, Merck, Novartis, Novo Nordisk, TenSixteen Bio and Tourmaline Bio; equity in Bolt, Candela, Mercury, MyOme, Parameter Health, Preciseli and TenSixteen Bio; and spousal employment at Vertex Pharmaceuticals, all unrelated to the present work. C.T. is the founder of C2T; a consultant for Bayer; a member of the scientific advisory board of PrognomiQ; and receives royalties from Exact Sciences. J.W.O. is the founder and CEO of Absolute DNA, with no direct relation to this study, and the interests are managed by University-Industry Foundation in Yonsei University Health System in accordance with their conflict-of-interest policies. E.A.L. is a member of the scientific advisory board for Inocras. All other authors declare no competing interests.

Disclaimer The views and opinions expressed in this article are those of the authors only and do not necessarily represent the views, official policy or position of the US Department of Health and Human Services or any of its affiliated institutions or agencies.

Additional information

Correspondence and requests for materials should be addressed to Tim H. H. Coorens, Ji Won Oh, Eunjung Alice Lee or Flora M. Vaccarino.

Peer review information *Nature* thanks Sarah Aitken, Young Seok Ju and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2025

¹Broad Institute of MIT and Harvard, Cambridge, MA, USA. ²Department of Anatomy, Yonsei University College of Medicine, Seoul, Korea. ³Graduate School of Medical Science, Brain Korea 21 Project, Yonsei University College of Medicine, Seoul, Korea. ⁴Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. ⁵Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA, USA. ⁶Department of Pediatrics, Harvard Medical School, Boston, MA, USA. ⁷Department of Quantitative Health Sciences, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA. ⁸McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA. ⁹Department of Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, British Columbia, Canada. ¹⁰Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada. ¹¹New York Genome Center, New York, NY, USA. ¹²National Disease Research Interchange, Philadelphia, PA, USA. ¹³Center for Developmental Biology and Regenerative Medicine, Seattle Children's Research Institute, Seattle, WA, USA. ¹⁴Division of Genetic Medicine, Department of Pediatrics, University of Washington, Seattle, WA, USA. ¹⁵Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. ¹⁶Department of Pathology, Harvard Medical School, Boston, MA, USA. ¹⁷Department of Pediatrics, University of Maryland School of Medicine, Baltimore, MD, USA. ¹⁸Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. ¹⁹Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA. ²⁰Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. ²¹Department of Pathology, Dana-Farber Cancer Institute, Boston, MA, USA. ²²Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. ²³Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. ²⁴Department of Genome Sciences, University of

Washington School of Medicine, Seattle, WA, USA. ²⁵Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. ²⁶Center for Human Genetics and Genomics, New York University Grossman School of Medicine, New York, NY, USA. ²⁷Department of Pediatrics, New York University Grossman School of Medicine, New York, NY, USA. ²⁸Department of Neuroscience and Physiology, New York University Grossman School of Medicine, New York, NY, USA. ²⁹Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ³⁰Center for Behavioral Health, Johns Hopkins All Children's Hospital, Saint Petersburg, FL, USA. ³¹Department of Molecular, Cell, and Cancer Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA. ³²Department of Genetics, Washington University, St. Louis, MO, USA. ³³Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA. ³⁴Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. ³⁵Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. ³⁶Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ³⁷Center for Pediatric Health Equity Research, Johns Hopkins All Children's Hospital, Saint Petersburg, FL, USA. ³⁸Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH, USA. ³⁹Department of Computer and Data Sciences, School of Engineering, Case Western Reserve University, Cleveland, OH, USA. ⁴⁰European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. ⁴¹Bridging Research Division on Mechanisms of Genomic Variation and Data Science, German Cancer Research Center, Heidelberg, Germany. ⁴²Division of Hematology and Medical Oncology, Department of Medicine and Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. ⁴³Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. ⁴⁴Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. ⁴⁵Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. ⁴⁶Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. ⁴⁷Department of Human Genetics, University of Utah, Salt Lake City, UT, USA. ⁴⁸Lieber Institute for Brain Development, Baltimore, MD, USA. ⁴⁹Department of Pathology, Stanford University, Stanford, CA, USA. ⁵⁰Department of Genetics, Stanford University, Stanford, CA, USA. ⁵¹Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. ⁵²Cardiovascular Research Center and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. ⁵³Department of Medicine, Harvard Medical School, Boston, MA, USA. ⁵⁴Ludwig Center at Harvard, Harvard Medical School, Boston, MA, USA. ⁵⁵Center for Genomics and Systems Biology, New York University, New York, NY, USA. ⁵⁶Department of Computer Science, Rice University, Houston, TX, USA. ⁵⁷Department of Neurology, Boston Children's Hospital, Boston, MA, USA. ⁵⁸Department of Epigenetics, Van Andel Institute, Grand Rapids, MI, USA. ⁵⁹Brotman Baty Institute for Precision Medicine, Seattle, WA, USA. ⁶⁰Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA. ⁶¹Division of Medical Genetics, Department of Pediatrics, University of Utah, Salt Lake City, UT, USA. ⁶²Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA. ⁶³Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. ⁶⁴Howard Hughes Medical Institute, Boston, MA, USA. ⁶⁵Edison Family Center for Genome Sciences & Systems Biology, Washington University School of Medicine, St. Louis, MO, USA. ⁶⁶Roy M. and Phyllis Gough Huffington Center on Aging, Baylor College of Medicine, Houston, TX, USA. ⁶⁷Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA. ⁶⁸Child Study Center, Yale University, New Haven, CT, USA. ⁶⁹Department of Neuroscience, Yale University, New Haven, CT, USA. ⁷⁰Yale Stem Cell Center, Yale University, New Haven, CT, USA. ⁶⁵Present address: European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL-EBI), Cambridge, UK. ⁶⁷These authors contributed equally: Tim H. H. Coorens, Ji Won Oh. ⁶⁵e-mail: tcoorens@broadinstitute.org; g1@yuhs.ac; elalice.lee@childrens.harvard.edu; flora.vaccarino@yale.edu

The Somatic Mosaicism across Human Tissues Network

National Institutes of Health

Richard S. Conroy⁷¹, Brionna Y. Hair⁷¹, Walter J. Koroshetz⁷², Roger Little⁷³, Amy C. Lossie⁷³, Jill A. Morris⁷², Dena C. Procaccini⁷¹ & Wendy Wang⁷⁴

Organizational Center (U24NS132103)

Ting Wang^{8,44,65}, Casey Andrews⁸, Lucinda Antonacci-Fulton⁸, Sarah Cody⁸, Milinn Kremitzki⁸, Heather A. Lawson⁴⁴, Daofeng Li⁴⁴, Tina Lindsay⁸ & Wenjin Zhang⁴⁴

Tissue Procurement Center (U24MH133204)

Thomas J. Bell¹², Thomas G. Blanchard¹⁷, Valerie J. Estela-Pro¹², Melissa A. Faith^{29,30}, Kayla Giancarlo⁷⁵, Melissa Grimm⁷⁵, Azra Hasan¹², Raquel G. Hernandez^{236,37}, M. Kathryn Leonard¹², Phoebe McDermott³⁰, Mary Pfeiffer⁷⁵, Isabel Sleeman¹² & Melissa W. Wondran¹²

Genome Characterization Center

University of Washington and Seattle Children's Research Institute – James T. Bennett (UM1DA058220)
James T. Bennett^{13,14}, Stephanie Bohaczuk²⁴, Colleen P. Davis^{24,76}, Evan E. Eichler^{24,25}, Christian D. Frasier^{24,76}, Kendra Hoekzema²⁴, Meng-Fan Huang^{24,76}, Caitlin Jacques^{24,76}, Dana M. Jensen¹³, J. Thomas Kolar^{24,76}, Youngjun Kwon²⁴, Kelsey Loy¹³, Yizi Mao²⁴, Min-Hwan Sohn⁶⁰, Katherine M. Munson²⁴, Shane Neph²⁴, Jeffrey Ou¹³, Nancy L. Parmalee¹³, Meranda M. Pham¹³, Jane Ranchalis⁶⁰, Luyao Ren^{24,25}, Adriana E. Sedeño-Cortés⁶⁰, Joshua D. Smith^{24,76}, Melanie Sorensen²⁴, Andrew B. Stergachis^{24,59,60}, Lila Sutherland¹³, Mitchell R. Vollger²⁴, Chia-Lin Wei^{24,76}, Jeffrey M. Weiss^{24,76}, Christina Zakarian^{24,76},

Natalie Y. T. Au¹³, Amy Leonardson¹³, Jiadong Lin²⁴, Taralynn Mack²⁴, Sean R. McGee^{24,76}, Anna Minkina⁶⁰, Patrick M. Nielsen^{24,76}, Chris Oliveira⁶⁰, Erica Ryke^{24,76}, Tristan Shaffer^{24,76} & Elliott G. Swanson^{24,60}

Baylor College of Medicine – Richard A. Gibbs (UM1DA058229)

Richard A. Gibbs^{22,23}, Elizabeth G. Atkinson²³, Sravya Bhamidipati²², Hsu Chao²², Rui Chen^{22,23}, Christopher M. Grochowski²², Harsha V. Doddapaneni²², Divya Kalra²², Ziad Khan²², Kavya Kottapalli²², Marie-Claude Gingras²², Walker Hale²², Heer Mehta²², Donna M. Muzny²², Muchun Niu²³, Luis F. Paulin²², Jeffrey Rogers^{22,23}, Evette Scott²², Fritz J. Sedlazeck^{22,23}, Kimberly Walker²², Tao Wu²³ & Chenghang Zong^{23,67}

Broad Institute of MIT and Harvard – Kristin G. Ardlie (UM1DA058235)

Kristin G. Ardlie¹, Viktor Adalsteinsson¹, Lisa Anderson¹, Carrie Cibulskis¹, Tim H. H. Coorens^{1,86,87}, Laura Domènec¹, Kiran Garimella¹, Whitney Hornsby¹, Steve Huang¹, Satoshi Koyama¹, Niall J. Lennon¹, Stephen B. Montgomery^{19,50,51}, Tetsushi Nakao¹, Azeet Narayan¹, Pradeep Natarajan^{1,52,53}, Evin Padhi⁵⁰, Constantijn Scharlee¹, Md Mesbah Uddin^{1,52}, Liying Xue¹, Zhi Yu¹, Shadi Zaheri¹, Ruolin Liu¹, Ning Stella Li¹ & Hang Su¹

Washington University – Ting Wang (UM1DA058219)

Ting Wang^{8,44,65}, Derek Albracht⁸, Eddie Belter⁸, Emma Casey⁸, Justin Chen⁴⁴, Yuchen Cheng⁸, Shihua Dong⁸, Qichen Fu⁸, Robert Fulton⁸, John Garza⁸, H. Josh Jang⁵⁸, Sheng Chih Jin⁸, Benjamin K. Johnson⁵⁸, Nahyun Kong⁸, Daofeng Li⁴⁴, Zefan Li⁸, Tina Lindsay⁸, Shane Liu⁸, Juan F. Macias-Velasco⁸, Elvira Mehiniovic⁸, Benpeng Miao⁸, Theron Palmer⁵⁸, Purva Patel⁸, Mary F. Majewski⁵⁸, Daniel C. Rohrer^{58,77}, Andrew Ruttenberg⁸, Ayush Semwal⁵⁸, Hui Shen⁵⁸, Jiawei Shen⁸, Zitian Tang⁸, Chad Tomlinson⁸, Wenjin Zhang⁸, Zilan Xin⁸, Christina Missler⁷⁷ & Brianne E. Livermore⁷⁷

New York Genome Center – Soren Germer (UM1DA058236)

Soren Germer¹¹, Samuel Aparicio^{9,10,11}, Jade E. B. Carter¹¹, William Driscoll¹¹, Uday S. Evani¹¹, Heather Geiger¹¹, M. Tausif Hasan¹¹, Manisha Kher¹¹, Dan A. Landau^{11,42,43}, Rajeeva Musunuri¹¹, Giuseppe Narzisi¹¹, Nicolas Robine¹¹ & Alexi Runnels¹¹

Data Analysis Center (UM1DA058230)

Peter J. Park^{34,54}, Mingyun Bae^{4,5}, Michele Berselli³⁴, Ann Caplin³⁴, Hye-Jung Elizabeth Chun³⁴, Niklas L. Engel³⁴, William C. Feng³⁴, Yan Gao³⁴, Nils Gehlenborg³⁴, Dominik Glodzik³⁴, Yoo-Jin Jiny Ha³⁴, Hu Jin³⁴, Sehi L'Yi³⁴, Eunjung Alice Lee^{1,4,5,6}, Heng Li^{34,45}, Po-Ru Loh^{1,46}, Lovelace J. Luquette³⁴, Maximilian Marin^{34,45}, Julia Markowski³⁴, Dominika Maziec³⁴, Huang Neng^{34,45}, Sarah Nicholson³⁴, Junseok Park^{1,5}, Qin Qian⁴⁵, Han Qu³⁴, Douglas Rioux³⁴, William Ronchetti³⁴, Andrew Schroeder³⁴, Corinne E. Sexton³⁴, Yichen Si³⁴, Kar-Tong Tan⁴⁵, David Tang⁴⁶, Alexander D. Veit³⁴, Vinayak V. Viswanadham³⁴, Suenghyun Wang^{4,6}, Kyung Ah Woo³⁴, Xi Zeng^{4,5}, Yuwei Zhang³⁴, Yifan Zhao³⁴, Ying Zhou⁴⁵, Michail Andreopoulos³⁴, Shannon Ehmsen³⁴, Cesar Ferreyra-Mansilla³⁴, Clara TaeHee Kim³⁴, David Michaels³⁴ & Bianca Morris³⁴

Tool and Technology Development

University of Michigan – Ryan E. Mills (UG3NS132084)

Ryan E. Mills^{18,19}, Brandt A. Bessell¹⁸, Alan P. Boyle^{18,19}, Ingrid Flashpohler¹⁸, Steve Losh¹⁸, Michael J. McConnell¹⁸, Torrin L. McDonald¹⁸, Camille Mumm¹⁹, Jessica A. Switzenberg¹⁸, Jinhao Wang¹⁸ & Weichen Zhou¹⁸

Boston's Children Hospital – Sangita Choudhury (UG3NS132144)

Sangita Choudhury^{1,4,6}, Guanlan Dong^{1,4,6}, Nazia Hilal^{1,4,6}, Se-Young Jo^{4,6}, Eunjung Alice Lee^{1,4,5,6}, Shayna L. Mallett^{1,4,6}, Monica Devi Manam^{4,6}, Shulin Mao^{4,6}, Diane D. Shao^{4,6,57}, Christopher Walsh^{4,6,64} & Sijing Zhao^{4,6}

Stanford University – Alexander E. Urban (UG3NS132146)

Alexander E. Urban^{60,63}, Yiling Elaine Huang⁵⁰, Jan O. Korbet^{40,41}, Abhiram Natu⁶⁸, Reenal Pattni⁵⁰, Carolin Purman⁵⁰, Flora M. Vaccarino^{68,69,70}, Bo Zhou⁵⁰ & Xiaowei Zhu⁷⁸

Baylor College of Medicine – Chenghang Zong (UG3NS132132)

Chenghang Zong^{23,67}, Jiayi Luo²³, Muchun Niu²³, Yichi Niu²³, Rohan Thakur⁷⁹, David A. Weitz⁷⁹ & Yang Zhang²³

Baylor College of Medicine – Fritz J. Sedlazeck (UG3NS132105)

Fritz J. Sedlazeck^{22,23,56}, Yilei Fu²², Michal B. Izydorczyk²², Luis F. Paulin²², Tao P. Wu^{23,66,67}, Xiaomei Zhan²² & Xinchang Zheng²²

Mayo Clinic – Alexej Abyzov (UG3NS132128)

Alexej Abyzov⁷, Geon Hue Bae^{2,3}, Taejeong Bae⁷, Areum Cho^{2,3}, June Hyug Choi^{2,3}, Yujin Angelina Choi⁷, Hyungbin Chun^{2,3}, Mrunal Dehankar⁷, Yeongjun Jang⁷, Seok-Won Jeong^{2,3}, Min Ji³, Mee Sook Jun^{3,4}, Su Rim Kim^{2,3}, Seong Gyu Kwon^{2,3}, Soung-Hoon Lee^{2,3}, Nam Seop Lim^{2,3}, Nanda Maya Mali^{2,3}, Ji Won Oh^{2,3,87}, Arijit Panda⁷, Jung Min Park^{2,3}, JaeEun Shin^{2,3} & Milovan Suvakov⁷

University of Utah – Gabor T. Marth (UG3NS132134)

Gabor T. Marth⁴⁷, Brad Demarest⁴⁷, Stephanie Gardiner⁴⁷, Stephanie J. Georges⁴⁷, Hunter R. Underhill^{61,62} & Yingqi Zhang⁶¹

Weill Cornell Medicine – Dan A. Landau (UG3NS132139)

Dan A. Landau^{11,42,43}, Sammantha Avaylon¹¹, Alexandre Pellán Cheng^{11,42}, Wei-Yu Chi⁴², Mariela Cortés-López⁴², Andrew R. D'Avino^{11,42}, Husain Danish^{11,42}, Elliot Eton^{11,42}, Foteini Fotopoulou^{11,42}, Saravanan Ganesan^{2,11}, Yiyun Lin^{11,42}, Qing Luo^{11,42}, Levan Mekerishvili^{11,42}, Joe Pelt^{11,42}, Catherine Potenski^{11,42}, Tamara Prieto^{11,42}, Jake Qiu^{11,42}, Ivan Raimondi^{2,11}, Rahul Satija^{11,55}, Mandeep Singh^{11,42}, Dennis Yuan^{11,42} & John Zinno^{11,42}

New York University – Gilad D. Evrony (UG3NS132024)

Gilad D. Evrony^{26,27,28}, Benjamin Costa^{26,27,28}, Jonathan Evan Shoag^{80,81}, Jimin Tan⁸², Aristotelis Tsigiris^{83,84}, Nisrine T. Jabara^{26,27,28} & Marta Grońska-Pęski^{26,27,28}

Boston's Children Hospital – Christopher A. Walsh (UG3NS132138)

Christopher A. Walsh^{1,4,5,6,64}, Emre Caglayan^{4,6}, Hayley Cline^{4,6}, Niklas L. Engel³⁴, Shelbi E. Gill^{4,6}, Robert Sean Hill^{4,6}, Andrea J. Kriz^{4,6}, Julia Markowski¹, Alisa Mo^{4,6}, Peter J. Park^{34,54}, Daniel Snellings^{4,6} & Vinayak V. Viswanadham³⁴

Dana Farber Cancer Institute – Kathleen H. Burns (UG3NS132127)

Kathleen H. Burns^{116,21}, Justin S. Becker^{115,16}, Bradley E. Bernstein^{115,16}, Aidan H. Burn^{116,21}, Wen-Chih Cheng^{116,21}, Jennifer A. Karlow^{116,21}, Cheuk-Ting Law^{116,21}, Eunjung Alice Lee^{1,4,5,6}, Shayna L. Mallet^{1,4,5}, Carlos Mendez-Dorantes^{116,21}, Khue H. Nguyen^{1,4,5}, Adam Voshall^{1,4,5} & Boxun Zhao^{1,4,5}

Broad Institute of MIT and Harvard – Fei Chen (UG3NS132135)

Fei Chen¹²⁰, Jason D. Buenrostro¹²⁰, Tim H. H. Coorens^{186,87}, Gad Getz^{116,35}, Benno Orr¹ & Andrew J. C. Russell¹²⁰

Case Western University – Fulai Jin (UG3NS132061)

Fulai Jin^{38,39}, He Li^{38,39}, Yan Li^{38,39}, Leina Lu^{38,39} & Xiaofeng Zhu⁸⁵

University of Massachusetts Chan Medical School – Thomas G. Fazzio (UG3NS132136)

Thomas G. Fazzio³¹, Trishita Basak³¹, Manuel Garber³³, Azita Ghodssi³³, Katrina Newcomer³¹ & Yuqing Wang³³

⁷¹Office of Strategic Coordination, Division of Program Coordination, Planning, and Strategic Initiatives, National Institutes of Health, Bethesda, MD, USA. ⁷²National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA. ⁷³Division of Neuroscience and Behavior, National Institute on Drug Abuse, National Institutes of Health, Bethesda, MD, USA. ⁷⁴Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, USA. ⁷⁵ConnectLife, Williamsville, NY, USA. ⁷⁶The Northwest Genomics Center, University of Washington, Seattle, WA, USA. ⁷⁷Pathology and Biorepository Core, Van Andel Institute, Grand Rapids, MI, USA. ⁷⁸Department of Neuroscience, City University of Hong Kong, Hong Kong, China. ⁷⁹Department of Physics, Harvard University, Boston, MA, USA. ⁸⁰Department of Urology, University Hospitals Cleveland Medical Center, Case Western Reserve University School of Medicine, Cleveland, OH, USA. ⁸¹Case Comprehensive Cancer Center, Cleveland, OH, USA. ⁸²Institute for Systems Genetics, NYU Grossman School of Medicine, New York, NY, USA. ⁸³Department of Medicine, NYU Grossman School of Medicine, New York, NY, USA. ⁸⁴Department of Pathology, NYU Grossman School of Medicine, New York, NY, USA. ⁸⁵Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA.