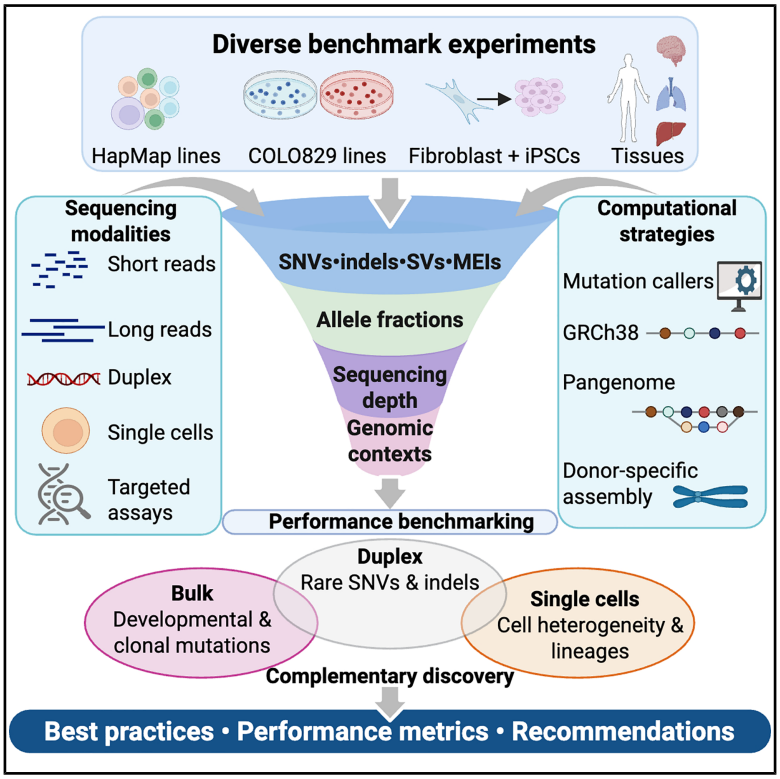
Cell

Resource

Comprehensive benchmarking of somatic mutation detection by the SMaHT Network

Graphical abstract



Highlights

- Bulk WGS assesses mutations occurring early in life or from clonal expansions
- Duplex sequencing discovers mutations independent of their clonal prevalence
- Single-cell WGS resolves mutation co-occurrence in a cell and inter-cell heterogeneity
- Personalized genome enables mutation discovery from telomere to telomere

Authors

The Somatic Mosaicism across Human Tissues Network

Correspondence

abyzov.alexej@mayo.edu

In brief

Benchmarking by the Somatic Mosaicism across Human Tissues Network demonstrates that advances in sequencing and analysis now enable systematic and comprehensive discovery of somatic mutations across diverse mutation types in non-cancerous tissues and across the genome from telomere to telomere.

The Somatic Mosaicism across Human Tissues Network, 2026, Cell *189*, 6594–6612
October 15, 2026 © 2026 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cell.2026.09.037

CellPress

CellPress
OPEN ACCESS

Cell

Resource

Comprehensive benchmarking of somatic mutation detection by the SMaHT Network

The Somatic Mosaicism across Human Tissues Network[1,2,*]
[1]Further details can be found in the supplemental information
[2]Lead contact: Alexej Abyzov (abyzov.alexej@mayo.edu)
*Correspondence: abyzov.alexej@mayo.edu (Alexej Abyzov)
https://doi.org/10.1016/j.cell.2026.09.037

SUMMARY

Somatic mosaicism is increasingly recognized as a fundamental feature of human biology, yet the detection of somatic mutations remains challenging. The Somatic Mosaicism across Human Tissues (SMaHT) Network conducted large-scale benchmarking experiments involving cell lines and donor tissues to evaluate sequencing technologies, experimental approaches, and computational methods for detecting different types of somatic mutations, generating community resources with >1,000× short-read and 100–400× long-read data for nine analyzed samples. We determined effective strategies for utilizing short- and long-read sequencing for mutation detection and demonstrated that using donor-specific and pangenome-inferred diploid assemblies improved calling, extending mutation catalogs to challenging genomic regions. We benchmarked six duplex technologies and showed that single-cell sequencing resolves cell-type-specific mutational patterns and heterogeneity. Our results indicate that bulk, single-cell, and duplex analyses are complementary and that leveraging all three provides comprehensive characterization of mosaicism within tissues. Together, these findings provide the scientific and clinical communities with a roadmap for accurate genome-wide somatic mutation analysis.

INTRODUCTION

Somatic mosaic mutations, i.e., post-zygotic genetic alterations, are increasingly recognized as important contributors to human health and disease beyond the context of cancer.[1,2] These mutations can arise during early development or in differentiated tissues, leading to cellular and genetic diversity that occurs naturally but also has been shown to play important roles in neurodevelopmental disorders, cardiovascular diseases, autoimmune conditions, and age-related health decline.[3–5] Unlike cancer, mutations arising from normal biological processes are relatively rare. Thus, detecting low variant allele fraction (VAF) somatic mutations remains a major technical and analytical challenge, particularly due to low mutation burden and, consequently, low signal-to-noise ratios in the data.

In bulk sample sequencing, the sensitivity to detect mutations is directly linked to sequencing depth, and, at standard genome coverage levels (e.g., 30–50×), low-VAF mutations are challenging to distinguish from sequencing errors or artifacts. Single-cell whole-genome amplification (scWGA) followed by sequencing can detect private mutations in individual cells but is limited by low throughput and often suffers from amplification noise, allelic dropout, and amplification biases, making it difficult to reliably detect mutations of all types genome wide.[6,7] High-fidelity single-molecule (also called "duplex") sequencing has emerged as a powerful strategy for reducing sequencing noise and enabling accurate mutation detection but is currently constrained to mostly quantifying average mutation burden and spectrum across cells in a sample or to deep profiling of specific targeted loci,[8] rather than genome-wide mutation discovery. Compounding these technical limitations unique to each sequencing modality is the absence of standardized analytical approaches for calling low-VAF mutations or annotating/interpreting their functional impact.

Only a handful of methods have been developed specifically to detect low-VAF somatic mosaic mutations, and these methods face trade-offs between sensitivity, specificity, runtime, and the ability to detect more complex mutation types, such as structural variants (SVs) or tandem repeat changes. Cancer consortia have developed resources for evaluating the discovery of somatic mutations,[9,10] but these efforts were primarily focused on detecting variants with higher VAF (>5%), often relying on paired analysis of tumor and matched normal tissues, limiting the applicability for detecting low-VAF mutations in normal tissues without a "matched normal" comparator sample. Other benchmark efforts by the Brain Somatic Mosaicism Network (BSMN),[11,12] the Genome in a Bottle (GIAB) consortium,[13] and Ha et al.[14] developed important insights and datasets for

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Box 1. Terminology utilized and contributing entities of the SMaHT Network

Somatic mutation: a mutation that is acquired in a person's lifetime, rather than inherited from their parents. Also known as post-zygotic mutations (PZMs) or mosaic mutations.
SNV: single-nucleotide variant, also sometimes referred to as single base substitution (SBS), are the mutations of one base into another, such as C > T.
Indel: short insertions or deletions of DNA sequence, by convention fewer than 50 base pairs in length.
SV: structural variant, mutations that involve large parts (over 50 bases) of the genome and may include large deletions, tandem duplications, inversions, and translocations between chromosomes.
MEI: mobile element insertion, a special class of SVs that involve the integration of retrotransposons into the genome.
sSNV, sIndel, sSV, and sMEI: the prefix "s" is often used to emphasize that mutations are somatic.
VAF: variant allele fraction, the ratio of sequencing reads supporting a specific mutation and the total number of reads covering the mutant site. It can be used to deduce the fraction of cells harboring a specific mutation.
Mutation burden/load: the total number of mutations (of a given class) in a tissue sample or cell.
DSA: donor-specific assembly, complete or near telomere-to-telomere haplotype assembly for a tissue donor.
Benchmarking or reference mutation set: a set of mutations that exists in a tissue, as supported by orthogonal evidence.
Single-molecule or duplex sequencing: dual-DNA-strand, ultra-low error rate sequencing approaches that randomly sample mutations in bulk DNA to estimate average mutation burdens and signatures across cells.
Primary template-directed amplification (PTA): a whole-genome amplification technique that in the context of this benchmarking exercise was applied to the analysis of genomes of individual cells.
Easy genomic region: intervals in the 1000 Genomes Project strict accessibility mask, representing confidently mappable, consistently callable loci.
Difficult genomic region: loci in PanMask pm151 (strict) but excluded from the 1000 Genomes Project strict mask, representing moderately mappable regions with elevated short-read error/ambiguity.
Extreme genomic region: the remainder of the genome outside both masks, enriched for highly repetitive or structurally complex sequence.
Contributing entities:
GCC: Genome characterization center—a center that is primarily focused on generating sequencing data.
TTD: Tools and technology development center—a center that is primarily focused on developing and testing experimental technologies and analytical tools for discovering somatic mutations.
TPC: Tissue procurement center—a center that is focused on collecting, processing, and distributing samples from post-mortem donors.
OC: Organization center—a center that organizes events and resources involving Network participants.
DAC: Data analysis center—a center that is focused on data aggregation, processing, analysis, and release.

evaluating somatic mutation detection. But these benchmark studies were focused on evaluating limited mutation types, i.e., single nucleotide variants (SNVs) and mobile element insertions (MEIs), with relatively high VAF (typically over 3%) or mutations in exonic regions. Overall, systematic benchmarking of duplex technologies and scWGA remains limited.

The Somatic Mosaicism across Human Tissues (SMaHT) Network aims to map somatic genetic diversity across different tissues and cells within individuals. In the initial phase of the SMaHT Network, benchmarking efforts were made to address the aforementioned gaps, expand and compare assessment capabilities, enable comprehensive detection of somatic mutations from using short- and long-read bulk whole-genome sequencing (WGS), inform the project's production phase, and support efforts from the broader scientific communities in the development and validation of tools that identify low VAF (<5%) SNVs, insertions or deletions (indels), MEIs, and SVs. Here, we describe key results of benchmark studies presented in a collection of papers from the network,[15–31] outline the datasets and resources generated from the benchmark experiments in the first phase of the SMaHT Network, which are publicly available for research communities, and provide conceptual and practical recommendations for the discovery of somatic mutations of various types across different experimental and sequencing contexts. To establish a common framework for consistency and clarity, we define terminology used in our studies and list contributing entities of the SMaHT Network (Box 1; Table S1).

RESULTS

Benchmark design

To address the challenges of benchmarking the detection of diverse somatic mutation types and advance method development, we designed and carried out four groups of benchmarking experiments: (1) "HapMap," using a mixture of 6 well-established lymphoblast cell lines; (2) "COLO829," well-characterized cancer and matched normal cell lines mixed at a known ratio; (3) a primary fibroblast sample with subclonal structure and matched clonal iPSC lines; and (4) six intact tissues or homogenate tissue samples from four phenotypically normal postmortem donors (Figures 1 and S1). These experiments were designed to comprehensively address the gaps in benchmark efforts that a single experiment alone could not address otherwise (Tables 1 and S2).

A common way of simulating somatic mosaicism is physical or *in silico* mixing of samples with well-characterized genomes such that known germline variants shared across a few, but not all mixed genomes, will simulate somatic mutations at different frequencies in the mix (Figure 1A). The SMaHT Network used six HapMap cell lines that are broadly consented for research use and genetically diverse with near telomere-to-telomere (T2T) genomes assembled by the Human Pangenome Reference Consortia (HPRC) (Figure 1A). The cell lines were combined at predefined ratios to simulate large numbers of somatic SNVs, indels, MEIs, and SVs at VAFs ranging from 0.25% to 16.5%, and the accuracy of the mixture was confirmed using droplet digital PCR (ddPCR). The availability of near-T2T genome assemblies enables us to benchmark mutation detection beyond the constraints of a standard reference genome. However, the simulated mutations are present on multiple haplotypes rather than just two as in real biological samples, and this discordance complicates interpretation of performance gains when using pangenome instead of a standard reference for the alignment, likely compromises mutation calling when using

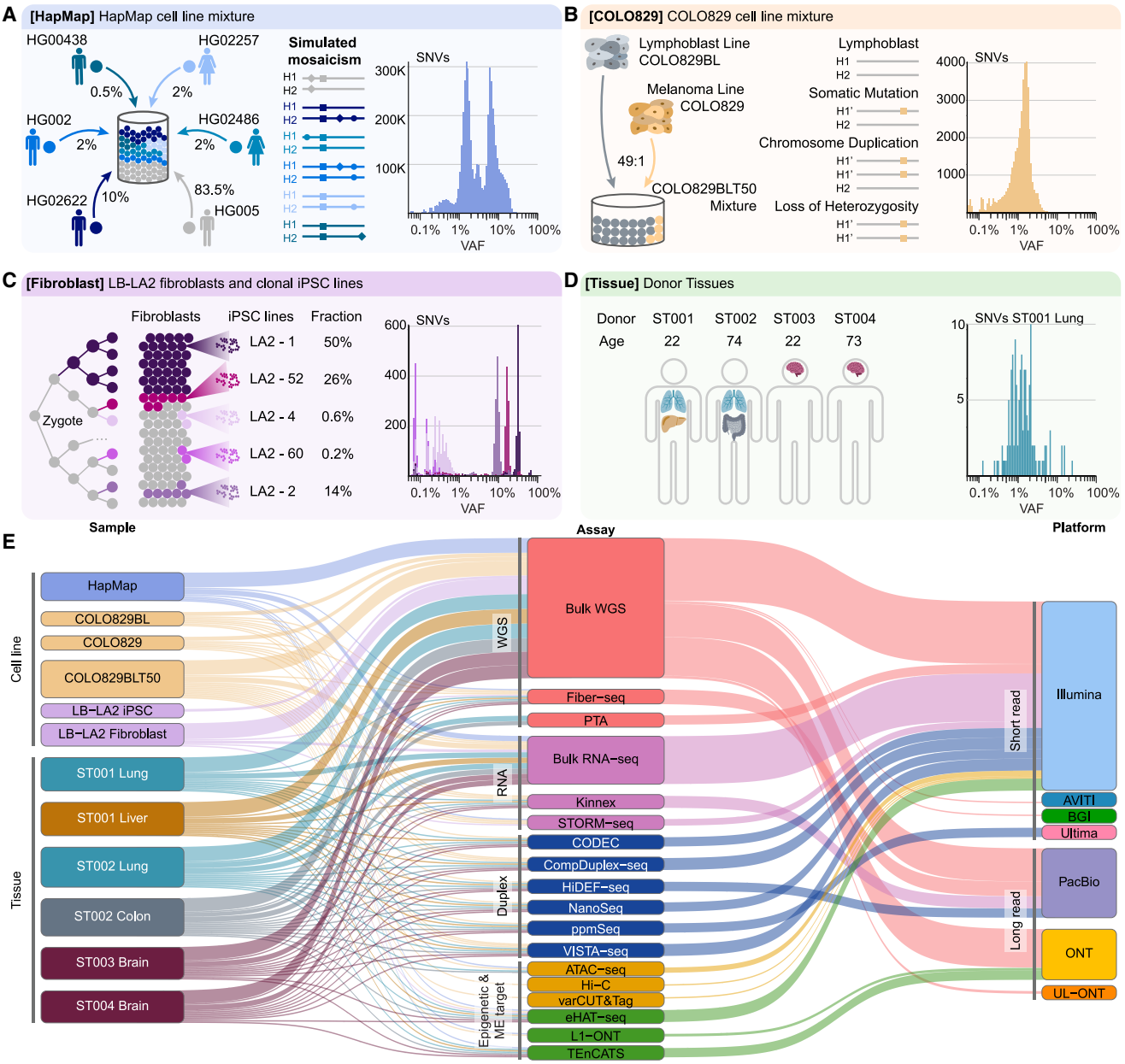


Figure 1. SMaHT benchmarking experiments

Four benchmarking experiments were carried out by the SMaHT Network.

(A) In the "HapMap" experiment, six well-characterized HapMap cell lines from donors with various ethnic backgrounds (CHB, Chinese; GWD, Gambian in Western Division, Mandinka; ASJ, Ashkenazim Jewish; ACB, African Caribbean in Barbados; CHS, Chinese Han in the South) were mixed at known proportions. Germline variants present in some (but not all) lines simulated somatic mutations (SNVs, indels, SVs, and MEIs) at VAFs ranging from 0.25% to 16.5%. The resulting sample has a mix of 12 haplotypes.

(B) In the "COLO829" experiment, well-characterized lymphoblastoid and melanoma lines from the same donor were mixed at a 49:1 ratio. Mutations in the melanoma line simulated SNVs and indels at VAF up to 5% in the mix.

(C) In the "fibroblast" experiment, a sample of primary fibroblasts, which has natural mutations accumulated during the lifetime of the donor, was analyzed. Five clonal iPSC lines were derived from the fibroblasts, allowing the inference of somatic SNVs and indels at VAF from 0.1% to 25% and revealing the clonal structure in the fibroblasts.

(D) In the "tissue" experiment, six frozen tissues obtained postmortem from four phenotypically normal donors were analyzed to benchmark the detection of somatic mutations in production-like samples. Intact tissue samples as well as homogenized samples were used depending on experimental assay types.

(E) The Sankey plot describes the cell line and tissue samples, experimental assays, and sequencing technologies applied in the benchmarking experiments. The width of connecting lines between columns reflects the number of samples from which data were generated from the corresponding experimental assays and sequencing technologies across multiple centers.

Table 1. Summary and comparison of benchmarking experiments and reference mutation sets

	HapMap mixture	COLO829BLT50	Fibroblasts	Donor tissues
Center for sample preparation	WashU-VAI	UW-SCRI	Yale	TPC
Mutation characteristics	germline based	cancer based	natural	natural
Haplotypes in a sample	12	2	2	2
SNVs	~5,000,000	49,377	11,227	868
Indels	~1,800,000	2,255	2,718	–
MEIs	3,066–10,607	1	–	21
SVs/CNVs	21,384	~100	a few	–
Reference mutation sets from clonal line	yes	yes	yes	no
Mutation frequencies	0.25%–16.5%; controlled	0.1%–5%; controlled	0.1%–25%; uncontrolled clonal structure	variable
Haplotype assemblies available	yes	yes	yes	yes
Sample condition	*in vitro* grown	*in vitro* grown	*in vitro* grown	postmortem snap frozen*
Terms for material and data distribution	unrestricted	unrestricted	protected	protected

The reference mutation sets from four donor tissues were inferred from the analysis of bulk WGS data. MEI counts ranged depending on whether they were classified based on repeat annotation (upper bound) or mechanistic signature of target primed reverse transcription (lower bound). Certain variant types were not analyzed in a few samples as denoted by "–." *For most analyses, homogenized tissue samples were used to ensure uniformity across analytical groups; however, single cells were also isolated from the intact tissue.

phasing information (e.g., with long reads), and limits the ability to benchmark somatic mutation calling in genomic regions where large stretches of DNA differ between individuals (e.g., satellite arrays). That led us to carry out further benchmark experiments in which all cells have the same two original mutated haplotypes.

To address this, we created another publicly available benchmarking resource using an established hypermutated melanoma cancer cell line (COLO829) and its matched normal lymphoblastoid cell line (COLO829BL) from the same individual (COLO829BLT50) (Figure 1B). For this benchmark, we generated a nearly complete T2T assembly and used deep short-read and long-read WGS of both lines to obtain a set of high-confidence somatic cancer mutations in the COLO829 line. The benchmark sample, COLO829BLT50, was created from mixing COLO829BL and COLO829 cells at a 49:1 ratio, and the accuracy of the mixture was confirmed using ddPCR. Aneuploidies in the COLO829 cell line enabled this mixture to simulate mutations with VAFs ranging from 0.1% to 5%. However, this approach was limited, as the somatic mutations in COLO829 are largely derived from a single mutational process (i.e., UV damage), lacking certain mutation types such as MEIs. Furthermore, this mixture sample is confounded by cell culture-derived clonal mutations in both the COLO829 and COLO829BL cell lines.

To provide an opportunity for benchmarking the discovery of genuine somatic mutations in non-tumor tissue and across a wide range of allele frequencies, we used a sample of primary human fibroblasts from a skin biopsy. The fibroblast sample was known to have a clonal structure with five clones represented by five single fibroblast cell-derived induced pluripotent stem cell (iPSC) lines.[32] We used these iPSC lines to derive a true-positive set of mutations present in the fibroblasts at VAF from 0.1% to 25% (Figure 1C). However, there were no somatic MEIs and only a few somatic SVs in the fibroblasts, making this approach suitable for only testing the discovery of somatic SNVs and indels.

All aforementioned benchmark experiments used *in vitro* cell lines, which are not likely to fully reflect all challenges when dealing with real human tissues. Therefore, the Network conducted benchmarking experiments using postmortem human tissue from four donors (Figure 1D). Mutation discovery in real human tissues provided the opportunity of assessing to what extent mutation heterogeneity across tissues could be captured using various sequencing approaches, including bulk WGS,[15–21] duplex sequencing,[22] and single-cell analyses by primary template-directed amplification (PTA).[23] As the genuine set of somatic mutations in the human tissues is unknown, we used orthogonal evidence from different analytical approaches and sequencing platforms to assess the accuracy of mutation detection.

For each experiment, a single laboratory prepared homogenized tissues (or suspended cells) and distributed samples to five Genome Characterization Centers (GCCs) and other groups in the Network, enabling us to evaluate the center-to-center variation in sequencing data (Table S1). For a few assays (including PTA), intact tissue samples processed at multiple sites were analyzed.

SMaHT benchmarking data generation

To comprehensively benchmark somatic mutation detection, we established a set of "core" and "extended" assays that were applied to all benchmark samples and to a subset of samples, respectively (Figure 1E). The "core" assays consisted of deep short- and long-read bulk WGS and RNA sequencing (RNA-seq) across multiple sequencing platforms, aimed to identify mutations at higher VAFs from earlier development and clonal expansions. We generated ultra-high coverage short-read bulk WGS data 1,712× for COLO829BLT50, 1,323× for HapMap mixture, 1,706× for fibroblasts, and 1,297–1,773× for tissue samples. For long-read PacBio HiFi WGS, we generated 390× for COLO829BLT50, 363× for HapMap mixture, 370× for

fibroblasts, and 95–345× for tissue samples. For long-read ONT WGS, we generated 218× for COLO829BLT50, 259× for HapMap mixture, 224× for fibroblasts, and 163–204× for tissue samples. In addition to the bulk WGS, we generated polyA-enriched and total RNA-seq (72 to 299 million read pairs) (Figure S1). The "extended" assays included, among others, duplex sequencing, single-cell WGS, ultra-long-read WGS, and targeted sequencing, and aimed to identify diverse types of low-VAF somatic mutations, discover mutations in genomic regions that were previously challenging to interrogate, and infer cellular lineages. In total, our benchmarking experiments produced over 90 Tb of raw sequencing data from multiple sequencing modalities (Tables S3 and S4).

All sequencing data and metadata were submitted to and systematically and uniformly processed by the Data Analysis Center (DAC) to ensure consistency and quality (STAR Methods; Figure S2; Table S5). At the DAC, each sample was checked for human and microbial contamination (Figure S3), sample mislabeling or duplicates, library qualities, and sequencing and alignment qualities in addition to individual standard sequencing QC performed at the data generating centers. In addition, the quality, contiguity, and completeness of the haplotype-resolved diploid genome assemblies were also checked using multiple approaches. Critical QC check results are tracked as part of metadata of the files accessible from the SMaHT Data Portal (https://data.smaht.org). The benchmarking data are registered under dbGaP (Table S6; key resources table).

As described below, we used these data to benchmark analytical approaches for detecting different classes of somatic mutations and summarized their key performance characteristics (Tables S7 and S8). We also provide practical recommendations for calling somatic mutations in Box 2.

Benchmarking experimental and computational methods for somatic SNV and indel discovery

We first sought to benchmark the performance of somatic SNV and indel detection using the COLO829BLT50 benchmarking material[15] (Figure 1B; Table 1). Across five Illumina WGS datasets generated from different sites (170–470×), 49% and 16.7% of the VAF <1% and 1%–3% COLO829-specific SNVs were not observed in at least one dataset, highlighting the experimental challenges associated with low-VAF (<3%) somatic mutation detection even under well-controlled conditions (Figure 2A). For robust detection of mutations with VAF less than 2%, we observed that bulk WGS with coverage greater than 300× was required.[15]

To assess the performance of various computational pipelines on somatic mutation detection, we conducted a Network-wide low-VAF somatic SNV/indel detection challenge, involving 7 institutions and 12 somatic variant detection workflows using five replicates of COLO829BLT50 WGS data with depths ranging from 170× to 470× and five computationally generated COLO829BLT50 WGS at 100× to 500×. An often-overlooked confounder in cell-line-based benchmarking is the presence of mutations acquired in cell culture via repeated expansion and harvesting of the cell line. We identified 5,568 of such cell-line mutations with high confidence (5,372 SNVs and 196 indels). By considering these true variants in the COLO829BLT50 mix, in addition to the variants present in the COLO829 tumor sample, we were able to more accurately evaluate the detection performance of the mutation calling methods. While the majority of the workflows performed well in detecting mutations above 2% VAF, all had a marked decline in performance for mutations at lower frequencies.[15] Furthermore, we observed that accuracy of mutation discovery strongly depended on the genomic regions (defined in Box 1) with significantly reduced sensitivity and increased false positive rate in difficult and extreme regions as compared with easy regions (Figures 2B and S4). We also found that beyond 300× sequencing depth, additional coverage yields diminishing returns in detecting additional mutations at VAF 2% or higher.[15]

To evaluate solutions to the challenges associated with the discovery of low-VAF mutations, we leveraged the fibroblast benchmarking material from donor LB using ultra-high coverage sequencing depth data, focusing on easy genomic regions.[16] Using the clonal iPSC lines derived from this fibroblast culture, we identified subclones with VAFs as low as 0.1% and 0.3%, enabling the establishment of low-VAF reference sets (Figure 2C). Our analyses revealed that ∼1,360× Illumina WGS enables accurate detection of ∼60% of SNVs at VAFs 0.5%–1%. By contrast, indels with VAFs <3% could be reliably detected only when located outside of homopolymeric regions. A similar analysis using CHM13 revealed approximately twofold higher sensitivity for low-VAF mutations, although this improvement came at the expense of reduced specificity (Figure S5). We conclude that CHM13 is promising for somatic mutation calling; however, native CHM13 population resources, rather than resources lifted over from GRCh38, will be required to accurately assess its overall benefit for somatic mutation detection (STAR Methods).

To assess the effects of various sampling approaches and sample conditions in real tissues on genome-wide somatic SNV discovery, we leveraged the post-mortem donor tissue benchmarking material to detect somatic mutations.[16] As compared with live fibroblasts, in post-mortem tissues we observed decreased accuracy of mutation discovery that was particularly prominent in non-easy genomic regions (e.g., in GIAB strict mask outside easy regions), where confidence of mapping short reads is reduced. This likely reflects the overall lower quality of and more damage to DNA due to the post-mortem interval before sample collection.[16] We discovered that filtering calls by read mapping quality and fragment length supporting mutations largely eliminates excessive false positives with little reduction in the true calls (Figure 2D), while read mapping in specific genomic regions highlighted data that is prone to a higher rate of false positive calls (Figure 2E). Finally, we demonstrated that combining 200–300× short-read WGS datasets from multiple tissues from the same donor to reach higher coverage is a viable strategy to enable the discovery of low VAF (<3%) mutations shared across tissues, which would otherwise be undetectable when each tissue is analyzed individually (Figure 2F).

Benchmarking alternative assembly approaches for somatic mutation discovery

Linear haploid reference genomes can differ substantially from the actual diploid 6 Gbp genome of an individual, causing

Box 2. Key recommendations for mutation calling from the SMaHT Network

SNVs and indels: mapping errors are the major source of false-positive calls.

- Mutation detection should be sequencing-depth aware and stratified by genomic mappability (e.g., easy and difficult-to-map regions)
- For VAF ~2% and coverage <300×, MosaicForecast provides the most balanced performance
- Higher coverage benefits likelihood-based callers like Mutect2 and alignment free callers like Rufus
- To arrive to a high-confidence call set, filter calls by:
  - Median read mapping quality (e.g., over 56)
  - Mean fragment length (e.g., over 300 bp)
  - Mean position in read (e.g., at least 20 bp)
  - Mean base quality (e.g., over 28)
- Fraction of reads with zero mapping quality in GIAB strict mask outside easy regions indicates data that is prone to yield false positive calls (particularly outside easy regions)
- Use large panel of normal genomes (such as the one derived by BSMN) to remove recurrent artifacts
- Use assembly-based approaches (e.g., Lancet or Mutect2) for indel discovery

SVs: long reads enable detection of approximately twice as many somatic SVs as short reads.

- Insertions are detected almost exclusively by long-read sequencing
- Use parameters optimized for mosaic and low VAF detection
- Stratify performance by genome context, repeat content, SV type and frequency
- For VAF >4%, standard long-read SV callers (SVDSS, Severus, Delly, or Sniffles2) perform well with ≥60× coverage
- For VAF 1%–4%, use SVDSS (HiFi) or Severus (HiFi/ONT) with ≥2 supporting reads and ≥90× coverage
- For VAF 0.5%–1%, use Severus with ≥2 supporting reads and ~180× HiFi coverage to balance sensitivity and precision

MEIs: xTea_mosaic (short reads, 200×) and PALMER (long reads, >60×) show robust performance, with the integrative strategy providing additional gains.

- For single-platform analyses with expected VAF >3%, 200× short-read WGS or 60× long-read WGS (with higher depth yielding additional performance) as a baseline offers an optimal cost-performance balance.
- For expected VAF <3%, long-read WGS is recommended for higher sensitivity and better performance in complex regions; targeted MEI assays (e.g., TEnCATS, eHAT-seq) provide cost-effective alternatives to WGS.
- For multi-platform analyses, performance improves by integrating evidence at the raw read/alignment level; precision can be further improved with 151-mer mappability filtering (short reads) and haplotype phasing (long reads).
- Phasing- and/or DSA-based MEI detection improves precision by at least 2.4-fold, particularly within segmental duplications.

Assembly: mutation calling in extreme regions and in loci poorly represented in GRCh38 reference necessitates use of a pangenome-inferred or donor-specific assembly, respectively.

- Evaluation of somatic variation within the most sequence and structurally complex genomic loci largely necessitates use of a donor-specific reference assembly
- Use Flagger and NucFlag to flag genomic loci inappropriately assembled
- Best practice is not to align directly to a population pangenome graph; instead infer donor-matching haplotypes and build a personalized diploid reference
- Pangenome-inferred diploid reference enables more confident mapping of reads in difficult and extreme regions with the highest ambiguous mappability in standard reference

Single-molecule/duplex: all techniques result in comparable SNV results.

- Comparison of advantages, disadvantages, and applicability of different techniques is provided in Table S10.

Single-cell analyses: extracting high-quality nuclei is key for successful amplification and mutation calling.

- Conduct 4-loci PCR to flag cells with uneven amplification, i.e., those with <4 clear expected amplicon bands
- Typical cell with good amplification sequenced at of 20–30× whole genome coverage will have:
  - Bell-shaped VAF distribution at 50% with ≤23% standard deviation and allele drop rate <8%
  - Median of the absolute values of all pairwise differences (MAPD) for 250 kb bins less than 0.5
  - Genomic fraction at 1× depth < 3%
  - Less than 15% of genome is CNV variable
- Once data is generated, conduct different QC for SNV/indel and for CNV analyses
- To increase precision of the call set use specialized callers that take into account allelic imbalance across genome
- Coverage over 15× per cell improves sensitivity of mutation discovery but does not improve accuracy
- Use highest stringency calls to estimate mutation burden and spectra in each cell
- Use more permissive calling for reconstructing developmental cell phylogenies
- High confidence CNVs and SVs are supported by coverage-based evidence (i.e., read depth or b-allele frequency) and fragment-based evidence (split read or discordant read pairs)

inaccurate read alignments and large missing genomic segments,[19] both of which can contribute to false positive and false negative somatic mutation calls (Figure 3A). We sought to evaluate whether donor-specific assemblies (DSAs) improve the performance of somatic variant detection,[33,34] using assemblies either inferred from a human pangenome reference or constructed directly for a donor. Direct alignment to the draft human pangenome reference is included for methodological completeness (Figure S6).[35,36] We first developed a somatic benchmarking mutation call set for the HapMap

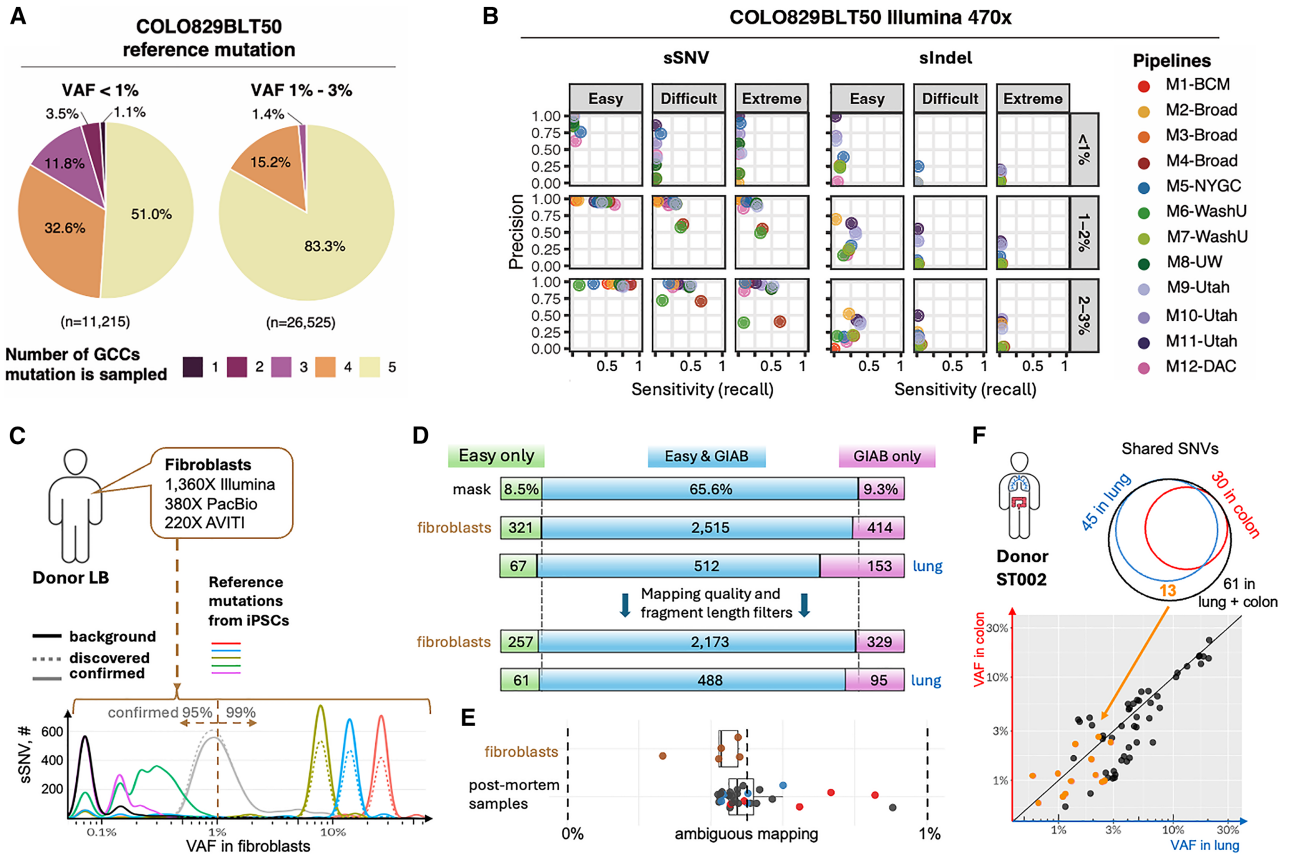


Figure 2. Multi-platform discovery and benchmarking of mosaic SNVs across the whole genome

(A) Distribution of COLO829BLT50 reference mutations by the number of GCCs in which each mutation was sampled across all datasets. VAFs were defined using the two highest depth datasets (470×). The number of reference mutations in each VAF group is denoted with *n*.

(B) Precision–sensitivity (recall) performance for discovering SNVs and indels of each calling pipeline in easy, difficult, and extreme genomic regions, aggregated into three VAF bins (<1%, 1%–2%, and 2%–3%). Dots correspond to pipelines.

(C) Utilizing fibroblasts and iPSCs from the second skin biopsy of left arm (LA2) from LB donor allowed deriving best practices for discovery somatic mutations by bulk WGS at high (>5%) and low (<1%) VAFs using 1,360× coverage Illumina data. Mutations discovered in bulk fibroblasts (gray line) but not present in the iPSC reference mutation set had a high confirmation rate by long-read sequencing.

(D) Excess of mutation calls outside of easy regions (e.g., in GIAB strict only) in samples from post-mortem donors (as compared with genomic mask and fibroblasts) was mostly mitigated by applying fragment length and mapping quality filters.

(E) Ambiguous mapping in GIAB strict-only regions (i.e., outside easy regions) was indicative of datasets with higher false positives (red for colon sample and blue for lung sample from donor ST002; black dots for samples from other donors; boxes for box plots).

(F) Applying best practices using data from production-like samples from the ST002 donor demonstrated that combining coverage from tissues enables discovery of mutations that are shared by but present at low cell fractions in the tissues.

mixture using a graph-based framework built from HapMap individual assemblies.[17] We then assessed the precision and recall of somatic mutation discovery using the HapMap mixture short-read sequencing data mapped to either GRCh38 or a pangenome-inferred diploid assembly (Figure 3B). Mapping to a pangenome-inferred diploid assembly improved somatic mutation calling accuracy compared with GRCh38, especially in extremely challenging genomic regions. In 500× short-read data, SNV precision increased from 0.79 to 0.89 and recall from 0.48 to 0.59 (Figure 3C).[18] For easy loci, accurate detection of ∼70% of SNVs was possible for VAFs >0.5%–1%, but for extreme loci this was limited to VAF >3%–4%. Gains extended to indels and SVs across all genomic strata: genome-wide, precision/recall improved from 0.74/0.19 to 0.77/0.26 for indels and from 0.66/0.20 to 0.66/0.32 for SVs, with the largest gains in difficult and extreme regions (in extreme regions, from 0.46/0.09 to 0.52/0.13 for indels and from 0.53/0.16 to 0.55/0.28 for SVs). For both variant classes, the inferred assemblies improve performance, but the inferred DSA cannot benchmark somatic mutation calling in some of the most challenging genomic regions (Table S9). Some, such as satellite arrays, contain large stretches of DNA that differ between individual haplotypes and, as a result, are often clipped from pangenome graphs owing to their complexity, so they remain as reference sequences in the inferred assembly. Others are retained in the graph but are too topologically complex for the diploid path to be inferred reliably, so the assembly is referenced there as well.

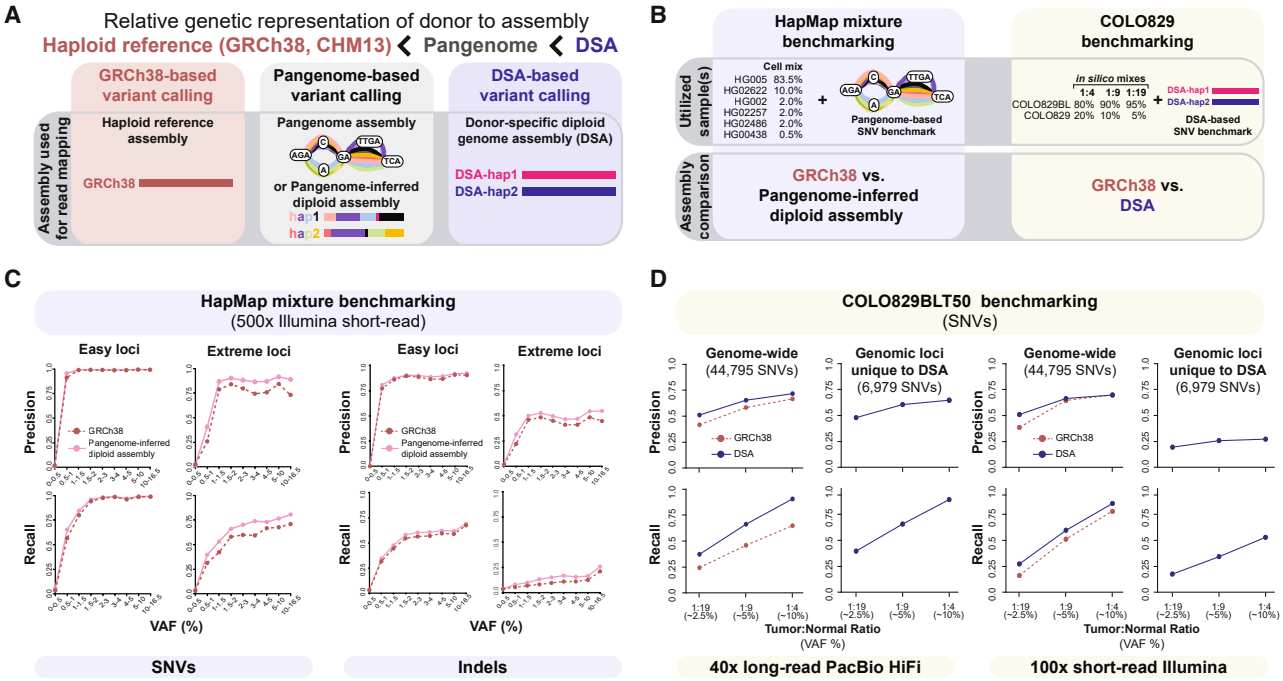


Figure 3. Benchmarking alternative assembly approaches for somatic mutation discovery

(A) Conceptual diagram showing the similarity of different genome references to the genome of the studied individual. By definition, DSA is the most similar, as it is generated directly from the genome of the individual being studied. A pangenome is a reference that includes a population of individuals, but not the individual being studied, and thus is less similar.

(B) A framework for evaluating the utility of pangenome- and DSA-based approaches for improving somatic variant detection.

(C) Precision and recall for somatic variant discovery using 500× Illumina short-read sequencing of the HapMap mixture using GRCh38- and pangenome-based approaches. Findings are separated based on the genomic mappability of the locus in which that somatic variant exists.

(D) Precision and recall for somatic variant discovery using 40× PacBio HiFi long-read sequencing data and 100× Illumina short-read sequencing data, each from a separate *in silico* mixture of the COLO829BL and COLO829 lines for GRCh38- and DSA-based approaches. Findings are separated into genome-wide as well as genomic loci that are unique to the COLO829BL DSA compared with GRCh38.

To more comprehensively benchmark somatic mutation discovery genome-wide, including within regions that lack synteny with GRCh38, we leveraged the COLO829 benchmark, for which we constructed a paired DSA for COLO829BL.[19] We then assessed the performance of somatic mutation discovery using *in silico* mixtures of the COLO829 and COLO829BL PacBio HiFi long-read and Illumina short-read sequencing data mapped to either GRCh38, CHM13, or the COLO829BL DSA (Figures 3B and S7). This revealed that using the DSA as the reference improved both the precision and recall of somatic mutation calling accuracy compared with either GRCh38 or CHM13 and enabled the accurate discovery of somatic mutations within genomic regions absent from reference genomes (Figures 3D and S7). This includes mutations within the kinetochore-binding region of the centromere, for which the paired single-molecule C-phosphate-G (CpG) methylation data can be used to assess associated functional alterations in centromere structure.[19] Notably, we found that CHM13 performed worse than GRCh38 for mutation detection, which we found was at least partially driven by unique and complex regions in CHM13 that are hypervariable between individuals (Figure S7). These regions are not yet well accounted for in current catalogs of germline variants, resulting in an excess of false positive somatic calls. Together, these findings demonstrate the utility of pangenome- and DSA-based approaches for improving somatic mutation discovery, especially within traditionally challenging genomic regions that are known to exhibit some of the highest rates of somatic variation.[19] Furthermore, these findings highlight the importance of using an appropriate panel of normal germline variation when performing somatic mutation discovery on a reference genome other than the donor's. We anticipate that development of additional computational tools that use diploid genome assemblies and graph genomes will further improve the ability of these emerging reference approaches for accurately identifying the full catalog of somatic mutations within an individual.

Benchmarking experimental and computational methods for somatic SV and MEI discovery

We next sought to benchmark the performance of somatic SV detection using the HapMap benchmark data.[20] First, we developed a high-confidence somatic SV reference set by harmonizing SVs across the six HapMap diploid assemblies (Figures 1A and 4A), resulting in ∼21,000 pseudo-somatic insertions and deletions, which we validated using ddPCR and other orthogonal methods. Using this reference set, we benchmarked 12 somatic SV detection methods across Illumina, PacBio HiFi, and ONT

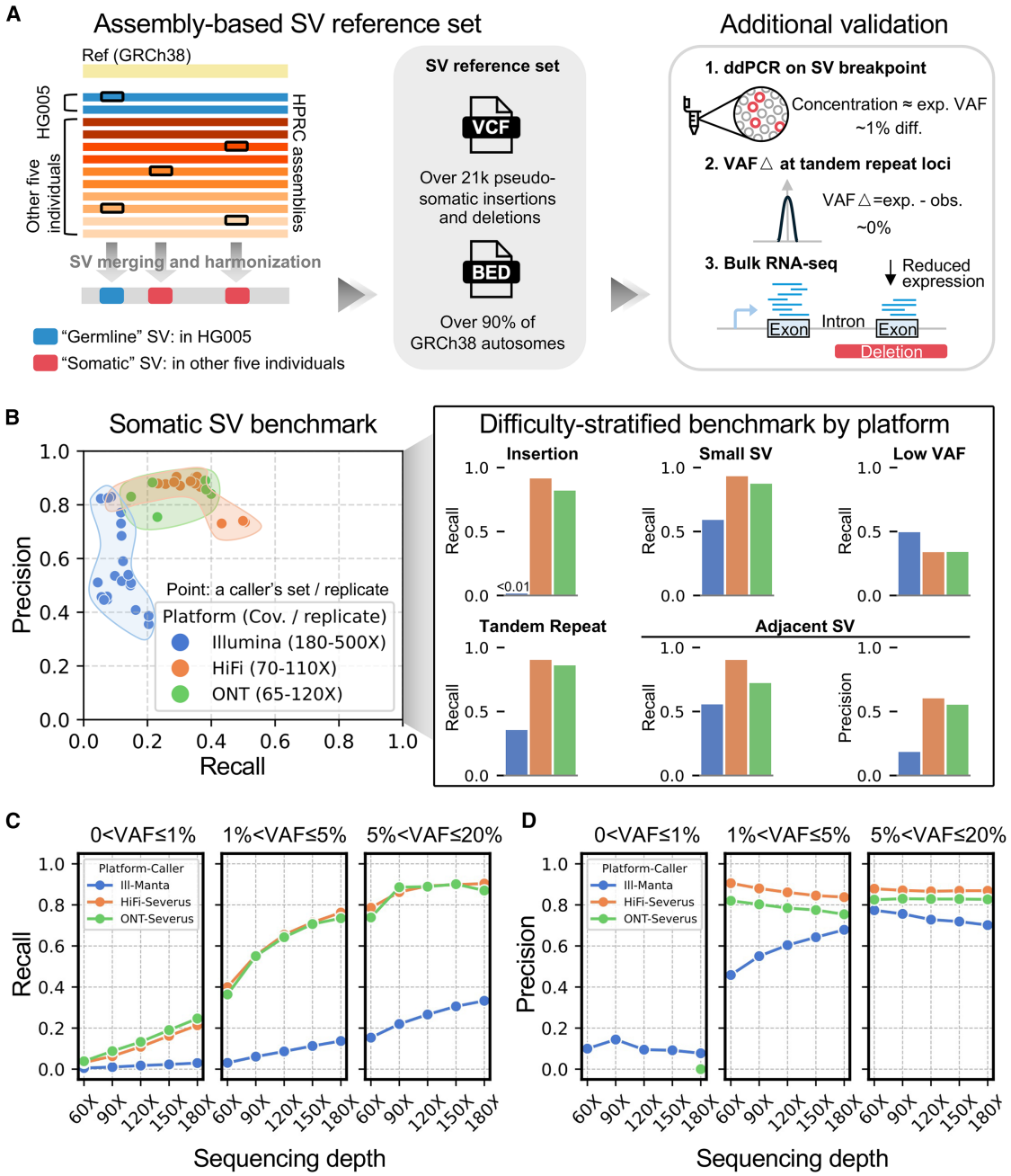


Figure 4. Assembly-based SV reference set for the HapMap mixture benchmark and cross-platform benchmarking of somatic SV calling

(A) The SV reference set was constructed by harmonizing diploid assembly-based SV calls from six individuals, validated via ddPCR, VAF checks at tandem repeat loci, and bulk RNA-seq.

(B) Cross-platform performance of somatic SV callers on each replicate, plus a difficulty-stratified benchmark covering insertions, small SVs (50–250 bp), low-VAF (<10%), tandem repeat-associated SVs, and adjacent SVs (breakpoints ≤ 500 bp apart). Illumina, HiFi, and ONT are shown in blue, orange, and green (legend on the left).

(C and D) Platform-specific detection limits at matched sequencing coverages stratified by VAF bins: ≤1%, 1%–5%, and 5%–20%. Missing data points in (D) reflect depth/VAF conditions where the caller detects zero SVs with a reported VAF ≤ 1%, precluding precision calculation.

sequencing data. Overall recall and precision of each strategy varied substantially across platforms, prompting a difficulty-stratified benchmark to pinpoint the causes of underperformance (Figure 4B). This difficulty-stratified analysis revealed limitations with short-read sequencing for detecting insertions, small SVs, and tandem repeats, and limitations with both short- and long-read sequencing for detecting low-VAF SVs and adjacent SVs. Under matched sequencing depths (60–180×), we observed

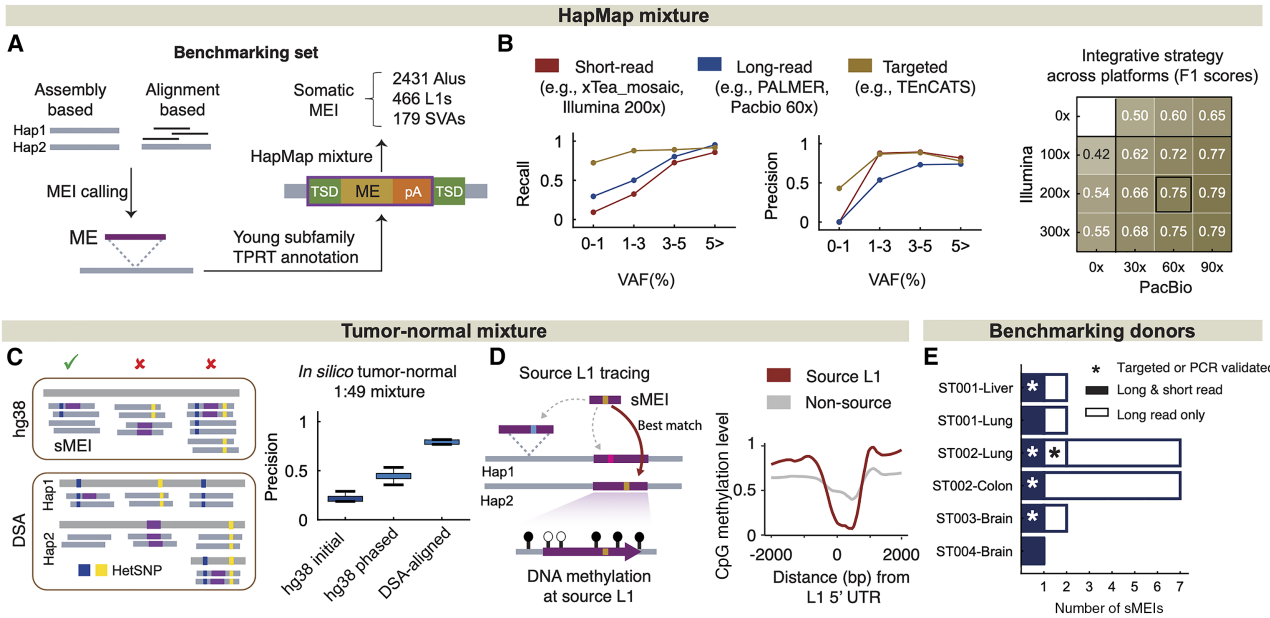


Figure 5. Benchmarking and optimized discovery of somatic MEIs

(A) Establishment of a benchmarking set in HapMap mixture. An sMEI truth set was generated as the foundation for evaluating performance, leveraging multiple technologies and incorporating information from repeat element subfamilies as well as target-primed reverse transcription (TPRT) features.

(B) Benchmarking sMEI detection in the HapMap mixture. Performance (recall and precision) was assessed across multiple sequencing platforms, including Illumina, PacBio, ONT, and transposon-targeted sequencing, as well as across different ranges of variant allele frequencies (VAFs). The improvement of a multi-platform integration strategy on detection performance was evaluated based on different sequencing depths.

(C) Enhanced sMEI detection through haplotype phasing and DSA in an *in silico* tumor-normal mixture using the CASTLE project data.[40] Precision improvements are demonstrated using two strategies at 2% mixture or 1% VAF level.

(D) Long-read data facilitate precise mapping of source L1 elements by leveraging internal sequence variation and assessment of DNA methylation profiling of the 5′ UTR region.

(E) Barplot of the number of somatic L1 insertion candidates identified in the benchmarking donors after implementing the optimized strategies from (A)–(C).

distinct platform-specific ceilings: short-read sequencing (i.e., Illumina) plateaued at ∼30% recall for SVs with a 5%–20% VAF, even at 180× coverage, whereas long-read sequencing (both PacBio HiFi and ONT) achieved ∼40% recall for 1%–5% VAF and ∼80% for 5%–20% VAF, while preserving precision >80%, at only 60× coverage (Figures 4C and 4D).

MEIs represent a clinically relevant subclass of SVs that are known to cause somatic variation.[37–39] As such, we sought to benchmark the performance of somatic MEI detection using a combination of computational methods and data from "core" and "extended" experimental assays.[21] First, using the HapMap mixture, we established a somatic benchmarking MEI reference set annotated with target-primed reverse transcription (TPRT) features critical for distinguishing RNA-mediated retrotransposition from DNA-level genomic rearrangements (Table 1; Figure 5A). Our evaluation across sequencing platforms, depths, VAFs, and genomic regions highlighted the advantages of long-read WGS and ME-targeted sequencing for detecting MEIs at low VAFs (< 3%) and in challenging genomic contexts (Figure 5B). Short-read WGS performance saturated at ∼200×, whereas 60× long-read WGS achieved comparable performance. Notably, long-read methods demonstrated 1.6-fold higher recall than short-read methods for MEIs at VAF <1%, while also resolving complete insertion structures and providing CpG methylation profiles, a key indicator of epigenetic ME suppression. ME-targeted assays achieved 72% recall for MEIs even with VAF <1%, offering a cost-effective alternative to WGS-based approaches for detecting very low-VAF events. Given the distinct error profiles between short- and long-read call sets, cross-platform integration combined with rescuing raw read-level signals further improved detection, yielding a 1.6-fold higher F1 score of 0.75, compared with simple intersection from 200× Illumina and 60× PacBio data, sequencing depths achievable for most SMaHT production samples (Figure 5B).

To further assess the performance of mutation detection using haplotype phasing and direct read alignment to DSA, we generated an *in silico* tumor-normal mixture from the CASTLE project,[40] where 376 somatic L1 insertions were identified in the tumor sample. Even at a VAF of 1% (i.e., 2% mixture), precision improved 2.1-fold with haplotype phasing and 3.7-fold with the DSA method compared with long-read-based detection alone (Figure 5C). We also developed a source-tracing pipeline leveraging L1 internal sequence variation, which successfully identified haplotype-resolved source elements for three times more somatic L1 insertions (67% vs. 22% of insertions > 1 Kbp) compared with conventional transduction-based tracing. These source L1s showed hypo-CpG methylation at promoter (5′ UTR) regions, consistent with these representing active L1

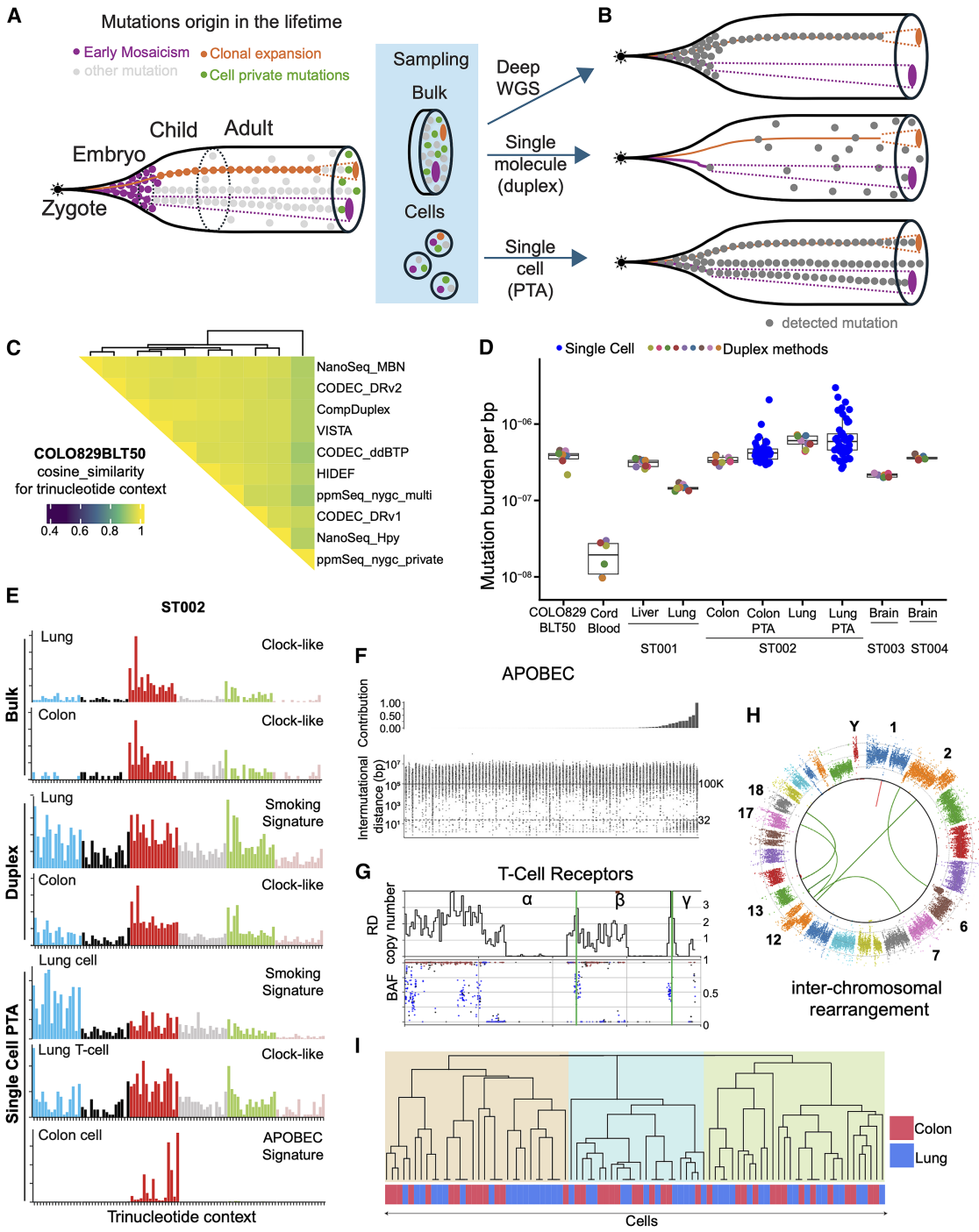


Figure 6. Benchmarking mutation discoveries using duplex and single-cell analyses

(A) Mutations in human cells occur across the life span. Early developmental and clonally expanded mutations will be frequent in sampled bulk tissues (orange and purple circles). They can also be sampled in single cells along with cell private and other mutations that occurred during lifetime.

(B) Bulk WGS is limited to assessing mutations occurring early in life or from clonal expansions. Duplex sequencing approaches discover mutations independent of their clonal prevalence. Single-cell analyses discover mutations in the analyzed cell regardless of their occurrence in time and clonal prevalence.

(C) Different duplex approaches are concordant in mutation spectra in the COLO829BLT50 sample.

(D) All duplex sequencing approaches yield consistent mutation burden estimates (box plots show variability) revealing heterogeneity between tissues.[22] PTA analyses result in concordant estimates of average mutation burdens while revealing heterogeneity between cells.

(E) Mutation spectra of two tissues profiled by deep bulk WGS, duplex sequencing, and PTA analyses. Bulk profiling is dominated by early developmental mutations, and spectra in two tissues are therefore indistinguishable. Duplex sequencing, given its unbiased profiling capability, captures tissue-specific

(legend continued on next page)

elements (Figure 5D). Finally, applying these optimized strategies to benchmark donor tissues, we identified 21 somatic L1 insertions (all VAFs < 1%) across all tissues profiled—liver, lung, colon, and brain. All events were fully resolved at the sequence and structural level, including the discovery of somatic U5 snRNA-L1 chimeric retrotransposition in normal human tissues (Figure 5E).

Benchmarking duplex and single-cell sequencing methods

The aforementioned benchmarks outline the challenges associated with identifying somatic mutations present in <1% of cells using bulk short- or long-read sequencing methods, indicating that bulk sequencing approaches are best powered to find mutations that occur early in development or that are present in expanded clonal lineages (Figure 6A). For instance, the fibroblast and the tissue donor benchmarking experiments showed that even at very high sequencing depths (~1,400×) only mutations above 0.5% VAF can be detected with sufficient sensitivity; such VAF would normally correspond in a human to the pre-gastrulation embryonic period. By contrast, duplex and single-cell sequencing approaches have the potential to probe the somatic mutation landscape across the entire lifespan irrespective of clonality (Figure 6B). Duplex sequencing tracks both strands of each original DNA molecule and calls a mutation only when complementary changes are seen on both strands, suppressing sequencing and amplification errors. This technology detects somatic mutations cost-effectively and provides unbiased single-molecule resolution, enabling simultaneous profiling of clonal and rare somatic SNVs and indels and their average burdens and spectra. By contrast, single-cell sequencing involves whole-genome amplification and sequencing of individual cells from a sample, enabling the assessment of the compendium of somatic mutations present within single cells, which can be used to uncover a cell's mutational burden and spectra and ancestral relationships between cells.

Aligned with the SMaHT objectives of discovering mutations across the entire genome, we conducted a comprehensive benchmarking of six different whole-genome duplex sequencing technologies (and their variants): CODEC, CompDuplex, HiDEF-seq, NanoSeq, ppmSeq, and VISTA-seq—applying them to the engineered COLO829BLT50, six tissues from post-mortem donors, and umbilical cord blood.[22] While the technologies differed in their library construction, sequencing platforms, depth of sequencing, cost-efficiency, genomic footprint, sensitivity, and ability to detect indels, the comparison revealed that the inference of mutation burden and signatures of both SNVs and indels was highly concordant across all methods (Figures 6C and 6D). Specifically, in the donor ST002, we observed a higher mutation burden in the lung than in the colon, which likely can be attributed to smoking of the donor and that is reflected in the substitution mutation spectrum and smoking-associated signature (Figure 6E) (donor information is available at dbGAP collection). Moreover, duplex sequencing yields a largely distinct set and profile of mutations as compared with deep bulk WGS. Mutational signature analysis demonstrated that duplex sequencing uncovers mutational processes at low VAFs (0.01%–1%) that are largely undetectable by deep bulk WGS, consistent with their later emergence during development and/or lifetime. Overall, duplex sequencing reveals a much more extensive and unique repertoire of <1% VAF clonal mutations, although deep duplex sequencing will be required to estimate their VAFs directly from data. Currently, the choice of duplex technology will depend upon specific desired applications, including targeting specific regions or conducting genome-wide assessment, the ability to detect indels or single-strand damages, and input material requirements (Table S10).

For donor ST002, we also performed PTA followed by short-read WGS (~30× coverage) of 102 single nuclei from the lung and colon.[23] The analysis of the PTA data reproduced the differences in mutation burden between lung and colon found by duplex sequencing (Figure 6D) and revealed that a fraction of cells from the lung are strongly impacted by signatures of tobacco exposure, while several cells from the lung and colon revealed the signature of APOBEC mutagenesis. Such cells also exhibited an increase in closely spaced (within 32 bp) mutations previously termed didyma[41] (Figure 6F). The observed mutation heterogeneity between cells can be partially explained by the mixture of different cell types in a tissue, as, for example, in both lung and colon samples we found the presence of T cells (five in total) as evidenced by genomic rearrangement at T cell receptor loci (Figure 6G). Moreover, several cells harbored large copy number variations (CNVs), chromosomal aneuploidies, inter-chromosomal rearrangements (resembling those found in cancers), and, consistent with expectations for an aged male, frequent losses of chromosome Y (Figure 6H). Finally, despite regional dropouts and amplification noise, PTA allowed detection of mutations shared between single cells, enabling the reconstruction from amplified cells of a phylogenetic tree of cellular lineages rooted at the zygote and validated by bulk sequencing (Figure 6I). Together, duplex and single-cell analyses revealed a rich landscape of somatic mutations that was otherwise obscured by bulk sequencing, while PTA positioned itself as a universal approach for single-cell genomics with unbiased recovery of mutations across cells, tissues, and developmental timeline.

mutational patterns. PTA further resolves mutation heterogeneity at the single-cell level, including developmental signatures, APOBEC signatures, and exposure to tobacco smoking.

(F) High contribution of APOBEC mutagenesis was pronounced in a number of cells. Such cells also exhibited an increase in closely spaced (within 32 bp) mutations previously termed didyma.[41]

(G) Five T cells were identified from discovered rearrangement and deletions at the known T cell receptor loci (heterozygous SNPs are in blue; homozygous SNPs are in brown).

(H) Some cells had large CNVs that were related to inter-chromosomal rearrangements (rearrangements in one cell are shown).

(I) Sharing of mutations between cells allowed reconstructing cell ancestry lineages. The root of the tree likely represents a zygote, while three highlighted developmental clades are apparent.[23,25]

DISCUSSION

We present a comprehensive and practical roadmap for the detection of somatic mosaic mutations (Box 2). This effort assessed the detection of various types of somatic mutations—including SNVs, indels, MEIs, and SVs/CNVs—across a broad range of VAFs, utilizing multiple technologies such as bulk deep whole-genome, single-cell, duplex, and targeted sequencing. The benchmark experiments and the analyses highlighted here revealed several key takeaways: (1) while bulk sequencing at 200–300× has limited sensitivity to detect mutations with VAF <2%, combining coverage across several tissues to reach a depth of ~1,400×, allows accurate discovery of SNVs at VAFs of 0.5%–1%, making possible the discovery of low frequency mutations shared across tissues (Figure 2F); (2) the use of personal genome assemblies and pangenome-inferred diploid assemblies enabled more accurate (both in terms of sensitivity and specificity) detection of somatic mutations, especially in repetitive genomic regions, than standard reference linear genomes (Figure 3); (3) read length matters more than depth alone for identifying somatic SVs and MEIs in repetitive regions (Figures 4 and 5); (4) various whole-genome duplex techniques, while having advantages and disadvantages, uncover concordant tissue-specific mutation burden and spectra that are obscured by bulk sequencing; and (5) PTA is a comprehensive single-cell genomics approach that yields an expansive view of diverse somatic mutation types from development through aging across diverse tissues (Figure 6).

We observed that bulk-, duplex-, and PTA-based single-cell WGS provide complementary and orthogonal means of validating findings from one another, where each approach offers distinct variant discovery niches with unique error profiles that could be corrected. We note that, although duplex sequencing and PTA-based single-cell analyses offer unique insights into tissue mosaicism, they are not yet complete substitutes for the bulk WGS approach, which delivers uniform genome coverage with <1% VAF detection sensitivity at sufficient coverage. Duplex sequencing is currently best suited for genome-wide sampling of low-VAF SNVs and indels, enabling assessment of average mutational burdens and profiles. As the duplex methods continue to advance, they can be predicted to have better utility for routine detection of SVs or MEIs. PTA, on the other hand, enables high-resolution analysis of SNVs, indels, and CNVs at the single-cell level but is a medium-throughput assay and comes with significant cost when scaling to hundreds of cells per sample. This drawback can also be anticipated to decrease with the advance of robotic technologies.[42]

The choice of an experimental approach depends on the specific scientific or clinical question being addressed, as well as budgetary considerations. Based on our analyses, we identified several common scenarios that researchers and clinicians are likely to encounter and provide recommendations for the most suitable strategies (Table S11). A common case is the detection of somatic mutations of interest—whether pathogenic or functionally relevant—including SNVs, SVs, or MEIs with VAFs greater than 1%. This scenario is best addressed using bulk WGS, with long-read sequencing particularly advantageous for detecting SVs and MEIs. Duplex sequencing would require substantially higher coverage to achieve comparable sensitivity, is not applicable to SVs, and has only limited applicability to MEIs. Similarly, detecting mutations at these VAFs using single-cell sequencing would require amplification and analysis of dozens of cells, making it cost-inefficient. For these reasons, duplex and single-cell approaches are generally less favorable for this application.

For detecting mutations at lower VAFs, targeted sequencing is typically the most cost-efficient solution, while single-cell sequencing remains the only viable approach for detecting low-VAF SVs. When the goal is not to identify specific mutations but rather to characterize the overall mutational landscape of a sample, duplex sequencing is the most favorable approach. Single-cell sequencing can provide an even more detailed view of mutational heterogeneity across cells, but at a substantially higher cost.

Detecting and monitoring clonal expansions can be achieved using both bulk and single-cell sequencing. Single-cell sequencing provides a definitive solution, as it unambiguously resolves whether multiple mutations co-occur within cells of the same clone or across distinct clones. Bulk sequencing, while less definitive in this regard—since mutations from different clones may appear at similar VAFs—is more cost-efficient because it does not require sequencing large numbers of individual cells.

Based on our experience, the analysis of mutations in highly repetitive genomic regions (such as tandem repeats, satellite repeats, and segmental duplications) is feasible with long-read sequencing, particularly when diploid DSAs are available. While the application of duplex and single-cell sequencing to these regions is theoretically possible, it remains an open technical challenge. For reconstructing cellular ancestry and building phylogenetic trees, single-cell sequencing is currently the only available approach.

Finally, elucidating the functional consequences of somatic mutations requires a functional readout, such as RNA-seq, CpG methylation profiling, or an assessment of chromatin configuration or histone biochemical marks. Several such analyses have been benchmarked as part of SMaHT companion papers. For example, it was demonstrated that single-molecule chromatin fiber sequencing (Fiber-seq) can be used to accurately identify the functional impact of genetic mutations with single-molecule precision and can also quantify somatic epimutations genome-wide, including in primary human tissue.[43] In addition, application of Fiber-seq to the COLO829 benchmarking material uncovered somatic epimutations altering the regulation of critical oncogenes.[19] Furthermore, it was demonstrated that deaminase-assisted Fiber-seq (DAF-seq) is an accurate and sensitive approach for quantifying the functional impact of low-VAF somatic mutations, as well as resolving the complete genomic and chromatin epigenomic map of single cells,[26] providing a path for future studies into the functional consequence of somatic variation. Concurrent profiling of genome sequence and functional output from the same cell provide a promising path forward, and approaches extending PTA to enable joint genome, transcriptome, and functional assays have recently emerged and show considerable promise.[25,26] For instance, combined genome-transcriptome amplification from the same cell permits the identification of cell types carrying specific mutations and assessment of potential functional consequences on the transcriptome. Duplex sequencing

coupled with functional readouts could, in principle, serve a similar role.[27]

The conclusions and described datasets represent a valuable community resource for benchmarking current and emerging sequencing technologies, studying somatic mosaicism, and advancing the development of analytical methods for somatic mutation discovery and analysis. The recommendations for mutation analysis spanning multiple experimental modalities and technologies equip the scientific and clinical communities with the knowledge needed to design and implement new studies relating somatic mosaicism to health, diseases, and basic biology. Although several challenges in somatic mutation discovery remain to be addressed, the findings presented here provide a widely applicable roadmap for accurate genome-wide somatic mutation analysis.

Limitations of the study

The findings from our study also illumined two key considerations when interpreting benchmarking experiments and choosing benchmarking materials. First, extensive culturing of cell lines may lead to culture-acquired mutations at high VAF, complicating the definition of a comprehensive true-positive mutation set in the test sample consisting of a mixture of cell lines. Second, specificity estimates derived from experiments with *in vitro* samples, although useful for method development, do not necessarily extrapolate to post-mortem tissues; hence, it is essential to provide orthogonal evidence to support the validity of mutation calls. Furthermore, we identified several challenges in mutation discovery that haven't completely been solved (Table S12). Our benchmarking efforts further enable us to propose a more streamlined and rational design based on creating controlled mixtures of multiple iPSC lines (Table S2).

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Alexej Abyzov (abyzov.alexej@mayo.edu).

Materials availability

Benchmarking samples are available to qualified researchers per request to the Network.

Data and code availability

This study was conducted as part of the NIH Common Fund's Somatic Mosaicism across Human Tissues (SMaHT) Network. Samples, data, and relevant software availability are described in the key resources table. More information about the SMaHT Network and data is available online at https://smaht.org and at https://data.smaht.org. Data generated were deposited into dbGaP: phs004193 and phs004425. New code developed during the described study can be found at the following locations: https://github.com/parklab/SMaHT_SNV_COLO829BLT50_HAPMAP (for COLO829 benchmark), https://zenodo.org/records/14009132 (for fibroblast benchmark), https://github.com/StergachisLab/T2T-COLO829BLT-Manuscript (for COLO829 DSA benchmark), https://github.com/twlab/SMaHT_Pangenome_Benchmark (for pangenome benchmark), https://github.com/BCM-HGSC/SMaHT_MIMS (for SV benchmark), https://github.com/ealee-lab/SMaHT-MEI-Benchmarking (for MEI benchmark), https://github.com/zonglab/CompDuplex (for duplex benchmark), and 10.5281/zenodo.17413251 (for PTA benchmark).

CONSORTIA

The members of the SMaHT Network are Alexej Abyzov, Abhiram Natu, Adam Voshall, Adriana E. Sedeño-Cortés, Aidan H. Burn, Alan P. Boyle, Alexander D. Veit, Alexander E. Urban, Alexandre Pellan Cheng, Alexi Runnels, Alisa Mo, Amoolya Srinivasa, Amy Leonardson, Anders Rasmussen, Andrea J. Kriz, Andrew B. Stergachis, Andrew J. C. Russell, Andrew J. Schroeder, Andrew R. D'Avino, Andrew Ruttenberg, Ann H. Caplin, Anna Minkina, Areum Cho, Arijit Panda, Aristotelis Tsirigos, Arnaz Maryam Tariq, Ayush Semwal, Azeet Narayan, Azra Hasan, Benjamin Costa, Benjamin K. Johnson, Benno Orr, Benpeng Miao, Bianca Morris, Bo Zhou, Boxun Zhao, Brad Demarest, Bradley E. Bernstein, Brandt A. Bessell, Brianne Carroll, Brianne E. Livermore, Caitlin N. Jacques, Camille Mumm, Carlos Mendez-Dorantes, Carolin Purman, Carrie Cibulskis, Casey Andrews, Catherine Potenski, Cesar R. Ferreyra-Mansilla, Chad Tomlinson, Charlotte Zhang, Chenghang Zong, Cheuk-Ting Law, Chia-Lin Wei, Chris Oliveira, Christian D. Frazar, Christina Missler, Christina Zakarian, Christopher A. Walsh, Christopher M. Grochowski, Christopher Walsh, Clara TaeHee Kim, Colleen P. Davis, Constantijn Scharlee, Corinne E. Sexton, Dan A. Landau, Dana M. Jensen, Daniel C. Rohrer, Daniel Snellings, Daofeng Li, David A. Weitz, David Michaels, David Tang, Dennis Yuan, Derek Albracht, Diane D. Shao, Divya Kalra, Dmitrii Olisov, Doga C. Gulhan, Dominik Glodzik, Dominika Maziec, Donna M. Muzny, Douglas Rioux, Eddie Belter, Elizabeth G. Atkinson, Elliot Eton, Elliott G. Swanson, Elvisa Mehinovic, Emma Casey, Emre Caglayan, Eric Ho, Erica Ryke, Erik Stricker, Eunjung Alice Lee, Evan E. Eichler, Evette Scott, Evin Padhi, Fei Chen, Flora M. Vaccarino, Foteini Fotopoulou, Fritz J. Sedlazeck, Fulai Jin, Gabor T. Marth, Gad Getz, Geon Hue Bae, Gilad D. Evrony, Giuseppe Narzisi, Guanlan Dong, H. Josh Jang, Han Qu, Hang Su, Harsha V. Doddapaneni, Hayley Cline, He Li, Heather A. Lawson, Heather Geiger, Heer Hemant Mehta, Heng Li, Hsu Chao, Hu Jin, Huang Neng, Hui Shen, Hunter R. Underhill, Husain Danish, Hye-Jung Elizabeth Chun, Hyungbin Chun, Ingrid Flashpohler, Isabel Sleeman, Ismail Yaman, Ivan Raimondi, J. Thomas Kolar, Jade E. B. Carter, JaeEun Shin, Jake Qiu, James T. Bennett, Jan O. Korbel, Jane Ranchalis, Jani Huuhtanen, Jason D. Buenrostro, Jeffrey M. Weiss, Jeffrey Ou, Jeffrey Rogers, Jennifer A. Karlow, Jessica A. Switzenberg, Ji Won Oh, Jiadong Lin, Jiawei Shen, Jiayi Luo, Jimin Tan, Jinhao Wang, Joe Pelt, John Garza, John Zinno, Jonathan Evan Shoag, Joseph Brew, Joshua D. Smith, Juan F. Macias-Velasco, Julia Markowski, June Hyug Choi, Junseok Park, Justin Chen, Justin S. Becker, Katherine M. Munson, Kathleen H. Burns, Katrina Newcomer, Kavya Chowdary Kottapalli, Kayla Giancarlo, Kelsey Loy, Kendra Hoekzema, Khue H. Nguyen, Kimberly Walker, Kiran Garimella, Kristin G. Ardlie, Kyung Ah Woo, Laura Domènech, Leina Lu, Levan Mekerishvili, Lila Sutherlin, Lisa Anderson, Livia Tomasini, Liying Xue, Lovelace J. Luquette, Lucinda Antonacci-Fulton, Luis F. Paulin, Luyao Ren, M. Kathryn Leonard, M. Tausif Hasan, Mandeep Singh, Manisha Kher, Manuel Garber, Marie-Claude Gingras, Mariela Cortés-López, Marta Grońska-Pęski, Mary F. Majewski, Mary Pfeiffer, Maximillian G. Marin, Md Mesbah Uddin, Mee Sook Jun, Melanie Sorensen, Melissa A. Faith, Melissa Grimm, Melissa W. VonDran, Meng-Fan Huang, Meranda M. Pham, Michael J. McConnell, Michail Andreopoulos, Michal B. Izydorczyk, Michele Berselli, Milinn Kremitzki, Milovan Suvakov, Min Kyu Kim, Min-Hwan Sohn, Mingyun Bae, Mitchell R. Vollger, Monica Devi Manam, Mrunal K. Dehankar, Muchun Niu, Nahyun Kong, Nam Seop Lim, Nancy L. Parmalee, Nanda Maya Mali, Natalia Brzozowska, Natalie Y. T. Au, Nazia Hilal, Niall J. Lennon, Nicolas Robine, Niklas L. Engel, Nils Gehlenborg, Ning Li, Nisrine T. Jabara, Patrick M. Nielsen, Patrick M. Van Hoose, Peter J. Park, Phoebe McDermott, Po-Ru Loh, Pradeep Natarajan, Purva Patel, Qichen Fu, Qing Luo, Rahul Satija, Rajeeva Musunuri, Ramon Massoni, Raquel G. Hernandez, Reenal Pattni, Richard A. Gibbs, Robert Fulton, Robert Sean Hill, Rohan Thakur, Rohini K. Gadde, Rui Chen, Rui Luo, Ruolin Liu, Ryan E. Mills, Saeyeon Jung, Sammantha Avaylon, Samuel Aparicio, Sang-Ho Yoon, Sangita Choudhury, Sarah Cody, Sarah Nicholson, Saravanan Ganesan, Satoshi Koyama, Se-Young Jo, Sean R. McGee, Sehi L'Yi, Seok-Won Jeoung, Seunghyun Wang, Shadi Zaheri, Shane Liu, Shane Neph, Shannon Ehmsen, Shayna L. Mallett, Shelbi E. Gill, Sheng Chih Jin, Shihua Dong, Shulin Mao, Sijing Zhao, Soren Germer, Soung-Hoon Lee, Sravya Venkata Bhamidipati, Srivinas Rajagoplan, Stephanie Bohaczuk, Stephanie Gardiner, Stephanie J. Georges, Stephen B. Montgomery, Steve Huang, Steve Losh, Su Rim Kim, Suhas S. Rao, Taejeong Bae, Tamara Prieto, Tao P. Wu,

Tao Wu, Taralynn Mack, Tetsushi Nakao, Theo Botella, Theron Palmer, Thomas G. Blanchard, Thomas G. Fazzio, Thomas J. Bell, Tim H. H. Coorens, Timothy R. Chu, Tina Lindsay, Ting Wang, Trishita Basak, Tristan Shaffer, Uday S. Evani, Utku S. Öztürk, Valerie J. Estela-Pro, Viktor Adalsteinsson, Vinayak V. Viswanadham, Walker Hale, Wei-Yu Chi, Weichen Zhou, Wen-Chih Cheng, Wenjin Zhang, Wenxuan Hong, Whitney Hornsby, William C. Feng, William Ronchetti, William T. Harvey, Xi Zeng, Xiaofeng Zhu, Xiaomei Zhan, Xiaowei Zhu, Xinchang Zheng, Yan Gao, Yan Li, Yang Zhang, Yeongjun Jang, Yichi Niu, Yifan Wang, Yifan Zhao, Yilei Fu, Yiling Elaine Huang, Ying Zhou , Yingqi Zhang, Yiyun Lin, Yizi Mao, Yoo-Jin Jiny Ha, Youngjun Kwon, Yuchen Cheng, Yuelin Liu, Yujin Angelina Choi, Yuwei Zhang, Zefan Li, Zhi Yu, Ziad Khan, Zilan Xin, and Zitian Tang.

ACKNOWLEDGMENTS

We are extremely grateful to the SMaHT donors and donor families, who have generously provided such precious gifts to support this important work. This research is supported by the NIH Common Fund through the Office of Strategic Coordination/Office of the NIH Director under awards U24MH133204, U24NS132103, UG3NS132024, UG3NS132061, UG3NS132084, UG3NS 132105, UG3NS132127, UG3NS132128, UG3NS132132, UG3NS132134, UG3NS132135, UG3NS132136, UG3NS132138, UG3NS132139, UG3NS 132144, UG3NS132146, UM1DA058219, UM1DA058220, UM1DA058229, UM1DA058230, UM1DA058235, and UM1DA058236.

AUTHOR CONTRIBUTIONS

Writing group, A.A., H.-J.E.C., T.H.H.C., H.V.D., S.C.J., M.K.L., J.M., A. Runnels, F.J.S., A.B.S., F.M.V., S. Zaheri, and W. Zhou; Figure 1 is by J.M., H.-J.E.C., and W. Hong; Figure 2 is by Y.-J.J.H. and Y.J.; Figure 3 is by Q.F., S.C.J., N.K., Y. Kwon, J.F.M.-V., B. Miao, A. Minkina, M.-H.S., A.B.S., Z.T., and Z.X.; Figure 4 is by Yuwei Zhang and F.J.S.; Figure 5 is by S.W., M. Bae, J.W., B. Zhao, R.E.M., W. Zhou, and E.A.L.; Figure 6 is by T.H.H.C., L.J.L., A. Natu, M. Suvakov, and Yang Zhang; Method and supplement writing contributors, M.A., T.J.B., M. Berselli, T.G.B., H. Chao, R. Chen, S.C., H.-J.E.C., M.K.D., H.V.D., G.D.E., C.D.F., Q.F., R.F., S. Gardiner, S.J.G., C.M.G., Y.-J.J.H., Eric Ho, Hui Shen, N.T.J., C.N.J., Y.J., S.C.J., B.K.J., Z.K., N.K., Y. Kwon, E.A.L., M.K.L., L.J.L., J.F.M.-V., M.F.M., H.H.M., B. Miao, D.M.M., A. Natu, M.N., J.O., N.L.P., A. Runnels, F.J.S., M.-H.S., A.B.S., M. Suvakov, Z.T., L.T., M.W.V., J.M.W., V.V.V., S.W., Y. Wang, S. Zaheri, Yang Zhang, Yuwei Zhang, C. Zong, and W. Zhou.

DECLARATION OF INTERESTS

A.B.S. is a co-inventor on a patent relating to the Fiber-seq and DAF-seq methods. M.N. has co-filed a patent application related to the CompDuplex-seq method through Baylor College of Medicine and is a co-founder and equity holder of Pioneer Genomics Inc. P.J.P. is a member of the Scientific Advisory Board for Bioskryb Genomics and a holder of software license for xTea (transposable element detection). C. Zong is a founder and equity holder of Pioneer Genomics Inc. A patent application related to CompDuplex technology has been filed by Baylor College of Medicine. E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc. J.E.S. is an EngagedMD consultant. G.D.E. is an inventor on a patent application for the HiDEF-seq technology. J.T.B. is a consultant for Mosaica Medicines. A.J.C.R. is listed as an inventor on a patent related to the Slide-tags technology (Single-nucleus high-resolution multimodal spatial genomics, PCT/US2023/085534 [2023]). N.G. is a co-founder and equity owner of Datavisyn. In the past 3 years, R.S. has received compensation from Bristol Myers Squibb, ImmunAI, Resolve Biosciences, Nanostring, 10x Genomics, Parse Biosciences and Neptune Bio. and is a co-founder and equity holder of Neptune Bio. 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. P.N. reports research grants from Allelica, Amgen, Apple, Boston Scientific, Cleerly, Genentech / Roche, Ionis, Novartis, and Silence Therapeutics; personal fees from AIRNA, 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, Incyte, Magnet Biomedicine, Merck, Novartis, Novo Nordisk, TenSixteen Bio, and Tourmaline Bio; equity in Bolt, Candela, Mercury, MyOme, Parameter Health, Preciseli, and TenSixteen Bio; royalties from Recora for intensive cardiac rehabilitation; and spousal employment at Vertex Pharmaceuticals, all unrelated to the present work. J.O.K. holds the patent "Comprehensive Detection of single cell genetic structural variation, Korbel J, Meiers S, Ghareghani M, Porubsky D, Marschall T, Sanders A (2020). Application reference code PCT/EP2020/060245." R. Liu is a co-inventor on a patent application (–†–°–¢/US2021/062966) covering CODEC. S. Aparicio is a co-founder of Genome Therapeutics Ltd. S.B.M. is a scientific advisor to MyOme, BridgeBio, PhiTech and Valinor Therapeutics. D.A.L. is on the SAB of Mission Bio, Veracyte, Ultima, and BioSkryb; has received prior research funding from 10x Genomics, Ultima Genomics, Oxford Nanopore Technologies, and Illumina; and has multiple patents on duplex and single-cell sequencing. D.K.'s immediate family member is employed at Oxford Nanopore Technologies. C.A.W. is a member of the SAB of Bioskryb Genomics, Mosaica Medicines, and CAMP4 Therapeutics; a consultant to Maze Therapeutics (equity); has received honoraria from Regeneration and Bristol Meyers Squibb; and is a co-inventor on patents for somatic mutation discovery. R.A.G. and Baylor College of Medicine have equity in Codified Genomics, Inc.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE MANUSCRIPT PREPARATION PROCESS

Portions of the text were refined using ChatGPT and subsequently reviewed to ensure that the original meaning and intent were preserved.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Sample acquisition, preparation and distribution
  - Data generation
  - Sequencing at UW, contributed by Jeffrey Ou, Christian D. Frazar, Caitlin N. Jacques, and Jeffrey M. Weiss
  - Sequencing at BCM, contributed by Chris Grochowski, Hsu Chao, Muchun Niu, Ziad Khan, Heer Mehta, Donna Muzny, Rui Chen, Chenghang Zong, and Harsha Doddapaneni
  - Sequencing at NYGC, contributed by Alexi Runnels
  - Sequencing at Broad, contributed by Shadi Zaheri
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data Processing at the Data Analysis Center (DAC), contributed by Michele Berselli and Hye-Jung Elizabeth Chun
  - Comparing SNV calling on CHM13 and GRCh38 references, contributed by Mrunal Dehankar, Yeongjun Jang, and Yifan Wang
  - Methods for MEI analyses, contributed by Seunghyun Wang and Weichen Zhou
  - Methods for SV analyses, contributed by Yuwei Zhang and Fritz Sedlazeck
  - Methods for assembly, contributed by Min-Hwan Sohn, Youngjun Kwon, and Andrew B. Stergachis
  - Methods for duplex, contributed by Yang Zhang, Vinayak V. Viswanadham, and Michail Andreopoulos
  - Methods for PTA analyses, contributed by Joe Luquette and Milovan Suvakov
  - Comparing benchmarking mutations set for SVs and MEIs, contributed by Nahyun Kong and Zitian Tang
  - Evaluating pangenome-based somatic variant detection approach, contributed by Qichen Fu, Benpeng Miao, and Juan F. Macias-Velasco

- Evaluating DSA-based somatic variant detection approach, contributed by Min-Hwan Sohn and Andrew B. Stergachis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2026.09.037.

Received: October 7, 2025
Revised: August 26, 2026
Accepted: September 18, 2026
Published: October 15, 2026

REFERENCES

1. Coorens, T.H.H., Oh, J.W., Choi, Y.A., Lim, N.S., Zhao, B., Voshall, A., Abyzov, A., Antonacci-Fulton, L., Aparicio, S., Ardlie, K.G., et al. (2025). The Somatic Mosaicism across Human Tissues Network. Nature *643*, 47–59. https://doi.org/10.1038/s41586-025-09096-7.
2. Zlotnikoff, M. (1945). A HUMAN MOSAIC. J. Hered. *36*, 163–168. https://doi.org/10.1093/oxfordjournals.jhered.a105487.
3. Vijg, J., and Dong, X. (2020). Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. Cell *182*, 12–23. https://doi.org/10.1016/j.cell.2020.06.024.
4. Ogawa, H., Horitani, K., Izumiya, Y., and Sano, S. (2022). Somatic mosaicism in biology and disease. Annu. Rev. Physiol. *84*, 113–133. https://doi.org/10.1146/annurev-physiol-061121-040048.
5. Corrigan, R.R., Mashburn-Warren, L.M., Yoon, H., and Bedrosian, T.A. (2025). Somatic mosaicism in brain disorders. Annu. Rev. Pathol. *20*, 13–32. https://doi.org/10.1146/annurev-pathmechdis-111523-023528.
6. Middelkamp, S., Manders, F., Peci, F., van Roosmalen, M.J., González, D.M., Bertrums, E.J.M., van der Werf, I., Derks, L.L.M., Groenen, N.M., Verheul, M., et al. (2023). Comprehensive single-cell genome analysis at nucleotide resolution using the PTA Analysis Toolbox. Cell Genom. *3*, 100389. https://doi.org/10.1016/j.xgen.2023.100389.
7. Luquette, L.J., Miller, M.B., Zhou, Z., Bohrson, C.L., Zhao, Y., Jin, H., Gulhan, D., Ganz, J., Bizzotto, S., Kirkham, S., et al. (2022). Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. Nat. Genet. *54*, 1564–1571. https://doi.org/10.1038/s41588-022-01180-2.
8. Lawson, A.R.J., Abascal, F., Nicola, P.A., Lensing, S.V., Roberts, A.L., Kalantzis, G., Baez-Ortega, A., Brzozowska, N., El-Sayed Moustafa, J.S., Vaitkute, D., et al. (2025). Somatic mutation and selection at population scale. Nature *647*, 411–420. https://doi.org/10.1038/s41586-025-09584-w.
9. Xiao, W., Ren, L., Chen, Z., Fang, L.T., Zhao, Y., Lack, J., Guan, M., Zhu, B., Jaeger, E., Kerrigan, L., et al. (2021). Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. Nat. Biotechnol. *39*, 1141–1150. https://doi.org/10.1038/s41587-021-00994-5.
10. Espejo Valle-Inclan, J., Besselink, N.J.M., de Bruijn, E., Cameron, D.L., Ebler, J., Kutzera, J., van Lieshout, S., Marschall, T., Nelen, M., Priestley, P., et al. (2022). A multi-platform reference for somatic structural variation detection. Cell Genom. *2*, 100139. https://doi.org/10.1016/j.xgen.2022.100139.
11. Wang, Y., Bae, T., Thorpe, J., Sherman, M.A., Jones, A.G., Cho, S., Daily, K., Dou, Y., Ganz, J., Galor, A., et al. (2021). Comprehensive identification of somatic nucleotide variants in human brain tissue. Genome Biol. *22*, 92. https://doi.org/10.1186/s13059-021-02285-3.
12. Zhu, X., Zhou, B., Pattni, R., Gleason, K., Tan, C., Kalinowski, A., Sloan, S., Fiston-Lavier, A.-S., Mariani, J., Petrov, D., et al. (2021). Machine learning reveals bilateral distribution of somatic L1 insertions in human neurons and glia. Nat. Neurosci. *24*, 186–196. https://doi.org/10.1038/s41593-020-00767-4.
13. Daniels, C.A., Abdulkadir, A., Cleveland, M.H., McDaniel, J.H., Jáspez, D., Rubio-Rodríguez, L.A., Muñoz-Barrera, A., Lorenzo-Salazar, J.M., Flores, C., Yoo, B., et al. (2026). Characterization of subclonal variants in HG002 Genome in a Bottle reference material as a resource for benchmarking variant callers. Cell Genomics *6*. https://doi.org/10.1016/j.xgen.2025.101104.
14. Ha, Y.-J., Kang, S., Kim, J., Kim, J., Jo, S.-Y., and Kim, S. (2023). Comprehensive benchmarking and guidelines of mosaic variant calling strategies. Nat. Methods *20*, 2058–2067. https://doi.org/10.1038/s41592-023-02043-2.
15. Jiny Ha, Y.-J., Maziec, D., Markowski, J., Georges, S.J., Parmalee, N.L., Berselli, M., Coorens, T.H.H., Dong, S., Gardiner, S., Kalra, D., et al. (2026). Detecting low allele fraction mosaic mutations: A multi-institution benchmarking with cell line mixtures. Cell Genomics *6*, 101390. https://doi.org/10.1101/2025.10.13.681545.
16. Jang, Y., Gardiner, S., Wang, Y., Natu, A., Dehankar, M.K., Tomasini, L., Panda, A., Cibulskis, C., Anderson, L., Coorens, T.H.H., et al. (2026). Accurate detection of sub-1% frequency somatic mutations by whole genome sequencing. American Journal of Human Genetics *113*. https://doi.org/10.1016/j.ajhg.2026.09.010.
17. Kong, N., Tang, Z., Ruttenberg, A., Macias-Velasco, J.F., Li, Z., Casey, E., Zhang, W., Miao, B., Xin, Z., Fu, Q., et al. (2026). A pangenomic method for establishing a somatic variant detection resource in HapMap mixtures. Cell Genomics *6*. https://doi.org/10.1016/j.xgen.2026.101358.
18. Fu, Q., Xin, Z., Miao, B., Zhang, W., Kong, N., Tang, Z., Ruttenberg, A., Albracht, D., Belter, E.A., Garza, J.E., et al. (2026). Leveraging human pangenome for improved somatic variant detection. Preprint at bioRxiv, 2026.01.04.697580. https://doi.org/10.64898/2026.01.04.697580.
19. Sohn, M.-H., Dubocanin, D., Vollger, M.R., Kwon, Y., Minkina, A., Munson, K.M., Hart, S.F.M., Ranchalis, J.E., Parmalee, N.L., Sedeño-Cortés, A.E., et al. (2026). A telomere-to-telomere map of somatic mutation burden and functional impact in a cancer. Cell *189*. https://doi.org/10.1016/j.cell.2026.09.033.
20. Zhang, Y., English, A.C., Paulin, L.F., Grochowski, C.M., Maheshwari, S., Mack, T., SMAHT SV working group, Park, P.J., and Sedlazeck, F.J. (2025). Comprehensive benchmarking of somatic structural variant detection at ultra-low allele fractions. Preprint at bioRxiv. https://doi.org/10.1101/2025.09.18.677206.
21. Wang, S., Bae, M., Wang, J., Zhao, B., Nguyen, K., Mallett, S., Switzenberg, J.A., Losh, S.J., Sexton, C.E., Miao, B., et al. (2025). Multi-platform framework for mapping somatic retrotransposition in human tissues. Preprint at bioRxiv, 2025.10.07.680917. https://doi.org/10.1101/2025.10.07.680917.
22. Zhang, Y., Viswanadham, V.V., Andreopoulos, M., Glodzik, D., Liu, R., Luquette, L.J., Jo, S.-Y., Narayan, A., Niu, M., Anderson, L., et al. (2025). Benchmarking of duplex sequencing approaches to reveal somatic mutation landscapes. Preprint at bioRxiv, 2025.12.12.692823. https://doi.org/10.64898/2025.12.12.692823.
23. Luquette, L.J., Coorens, T.H.H., Natu, A., Suvakov, M., Caplin, A., Jun, M.S., Mo, A., Pelt, J., Anderson, L., Berselli, M., et al. (2025). A comprehensive view of somatic mosaicism by single-cell DNA analysis. Preprint at bioRxiv, 2025.10.31.685648. https://doi.org/10.1101/2025.10.31.685648.
24. Zhang, Y., Chao, H., Niu, M., Grochowski, C.M., Kottapalli, K., Bhamidipati, S.V., Muzny, D.M., Gibbs, R.A., Zong, C., and Doddapaneni, H.V. (2025). Expanding the genome in a Bottle truth set: Detection and validation of novel low-frequency variants using high-accuracy NanoSeq. Preprint at bioRxiv, 2025.12.05.692678. https://doi.org/10.64898/2025.12.05.692678.
25. Natu, A., Dehankar, M.K., Pattni, R., Suvakov, M., Olisov, D., Tomasini, L., Jang, Y., Huang, Y., Benito-Garragori, E., Hasenfeld, P., et al. (2025). Single cell whole genome and transcriptome sequencing links somatic mutations to cell identity and ancestry. Preprint at bioRxiv, 2025.10.28.685157. https://doi.org/10.1101/2025.10.28.685157.

26. Swanson, E.G., Mao, Y., Mallory, B.J., Vollger, M.R., Bohaczuk, S.C., Oliveira, C.B., Lyon, D.B., Ranchalis, J., Parmalee, N.L., Cohen, B.A., et al. (2025). Mapping single-cell diploid chromatin fiber architectures using DAF-seq. Nat. Biotechnol. https://doi.org/10.1038/s41587-025-02914-3.

27. Niu, M., and Chuck Zong, C. (2024). CompDuplex: Accurate detection of somatic mutations by duplex-seq with comprehensive genome coverage v1. https://doi.org/10.17504/protocols.io.kxygx3x4og8j/v1.

28. Mack, T., Lin, J., Ren, L., Sohn, M.-H., Minkina, A., Kwon, Y., Yoo, D., Sui, Y., Munson, K.M., Hoekzema, K., et al. (2026). Donor-specific assemblies enhance somatic structural variant detection in complex genomic regions. Preprint at bioRxiv, 2026.02.20.707061. https://doi.org/10.64898/2026.02.20.707061.

29. Stricker, E., Jaryani, F., Izydorczyk, M., Poon, C.-L., Sanio, P., Alexander, A., Deb, S.K., Sedlazeck, F., Rogers, J., and Atkinson, E.G. (2025). MosaicSim: A novel mosaic variant simulator reveals diminishing returns of ultra-high coverage for mosaic variant detection. Preprint at bioRxiv, 2025.12.03.692191. https://doi.org/10.64898/2025.12.03.692191.

30. Engel, N.L., Brors, B., Feuerbach, L., and Park, P.J. (2025). Nanopore whole-genome sequencing reveals conserved chromosome-specific telomere architecture across tissues and populations. Preprint at bioRxiv. https://doi.org/10.64898/2025.12.10.693405.

31. Zhou, W., Mumm, C., Gan, Y., Switzenberg, J.A., Wang, J., De Oliveira, P., Kathuria, K., Losh, S.J., Bessell, B., McDonald, T.L., et al. (2025). A personalized multi-platform assessment of somatic mosaicism in the human frontal cortex. Preprint at bioRxiv, 2024.12.18.629274. https://doi.org/10.1101/2024.12.18.629274.

32. Fasching, L., Jang, Y., Tomasi, S., Schreiner, J., Tomasini, L., Brady, M.V., Bae, T., Sarangi, V., Vasmatzis, N., Wang, Y., et al. (2021). Early developmental asymmetries in cell lineage trees in living individuals. Science *371*, 1245–1248. https://doi.org/10.1126/science.abe0981.

33. Cheng, H., Asri, M., Lucas, J., Koren, S., and Li, H. (2024). Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. Nat. Methods *21*, 967–970. https://doi.org/10.1038/s41592-024-02269-8.

34. Rautiainen, M., Nurk, S., Walenz, B.P., Logsdon, G.A., Porubsky, D., Rhie, A., Eichler, E.E., Phillippy, A.M., and Koren, S. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat. Biotechnol. *41*, 1474–1482. https://doi.org/10.1038/s41587-023-01662-6.

35. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. Nature *617*, 312–324. https://doi.org/10.1038/s41586-023-05896-x.

36. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy, A.B., Asri, M., Carson, C., Chaisson, M.J.P., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature *604*, 437–446. https://doi.org/10.1038/s41586-022-04601-8.

37. Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A., et al. (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483–496. https://doi.org/10.1016/j.cell.2012.09.035.

38. Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley, P., Lehmann, H.S., Park, P.J., et al. (2015). Cell lineage analysis in human brain using endogenous retroelements. Neuron *85*, 49–59. https://doi.org/10.1016/j.neuron.2014.12.028.

39. Nam, C.H., Youk, J., Kim, J.Y., Lim, J., Park, J.W., Oh, S.A., Lee, H.J., Park, J.W., Won, H., Lee, Y., et al. (2023). Widespread somatic L1 retrotransposition in normal colorectal epithelium. Nature *617*, 540–547. https://doi.org/10.1038/s41586-023-06046-z.

40. Park, J., Cook, D.E., Chang, P.-C., Kolesnikov, A., Brambrink, L., Mier, J.C., Gardner, J., McNulty, B., Sacco, S., Keskus, A., et al. (2026). Accurate somatic small variant discovery for multiple sequencing technologies with DeepSomatic. Nature Biotechnology. https://doi.org/10.1038/s41587-025-02839-x.

41. Durfee, C., Bergstrom, E.N., Díaz-Gay, M., Zhou, Y., Temiz, N.A., Ibrahim, M.A., Nandi, S.P., Wang, Y., Liu, X., Steele, C.D., et al. (2025). Tobacco smoke carcinogens exacerbate APOBEC mutagenesis and carcinogenesis. Preprint at bioRxiv, 2025.01.18.633716. https://doi.org/10.1101/2025.01.18.633716.

42. Prieto, T., Yuan, D.J., Zinno, J., Hughes, C., Midler, N., Kao, S., Huuhtanen, J., Raviram, R., Fotopoulou, F., Ruthen, N., et al. (2025). Large-scale single-cell phylogenetic mapping of clonal evolution in the human aging esophagus. Preprint at bioRxiv, 2025.10.11.681805. https://doi.org/10.1101/2025.10.11.681805.

43. Vollger, M.R., Swanson, E.G., Neph, S.J., Mallory, B.J., Dubocanin, D., Ho, C.-H., Cheng, Y.-H.H., Ranchalis, J., Munson, K.M., Sedeno-Cortes, A.E., et al. (2026). Somatic epimutations cap genetic determinism in the human diploid chromatin epigenome. Cell *189*. https://doi.org/10.1016/j.cell.2026.09.034.

44. Bae, T., Fasching, L., Wang, Y., Shin, J.H., Suvakov, M., Jang, Y., Norton, S., Dias, C., Mariani, J., Jourdon, A., et al. (2022). Analysis of somatic mutations in 131 human brains reveals aging-associated hypermutability. Science *377*, 511–517. https://doi.org/10.1126/science.abm6222.

45. Sarangi, V., Jang, Y., Suvakov, M., Bae, T., Fasching, L., Sekar, S., Tomasini, L., Mariani, J., Vaccarino, F.M., and Abyzov, A. (2022). All2: A tool for selecting mosaic mutations from comprehensive multi-cell comparisons. PLOS Comput. Biol. *18*, e1009487. https://doi.org/10.1371/journal.pcbi.1009487.

46. Freed, D., Aldana, R., Weber, J.A., and Edwards, J.S. (2017). The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. Preprint at bioRxiv. https://doi.org/10.1101/115717.

47. Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods *15*, 591–594. https://doi.org/10.1038/s41592-018-0051-x.

48. Krishnamachari, K., Lu, D., Swift-Scott, A., Yeraliyev, A., Lee, K., Huang, W., Leng, S.N., and Skanderup, A.J. (2022). Accurate somatic variant detection using weakly supervised deep learning. Nat. Commun. *13*, 4248. https://doi.org/10.1038/s41467-022-31765-8.

49. Van Der Auwera, G.O., and Connor, B.D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra.

50. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219. https://doi.org/10.1038/nbt.2514.

51. Ostrander, B.E.P., Butterfield, R.J., Pedersen, B.S., Farrell, A.J., Layer, R.M., Ward, A., Miller, C., DiSera, T., Filloux, F.M., Candee, M.S., et al. (2018). Whole-genome analysis for effective clinical diagnosis and gene discovery in early infantile epileptic encephalopathy. NPJ Genom. Med. *3*, 22. https://doi.org/10.1038/s41525-018-0061-8.

52. Richter, F., Morton, S.U., Kim, S.W., Kitaygorodsky, A., Wasson, L.K., Chen, K.M., Zhou, J., Qi, H., Patel, N., DePalma, S.R., et al. (2020). Genomic analyses implicate noncoding de novo variants in congenital heart disease. Nat. Genet. *52*, 769–777. https://doi.org/10.1038/s41588-020-0652-z.

53. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710.

54. Jin, H., Gulhan, D.C., Geiger, B., Ben-Isvy, D., Geng, D., Ljungström, V., and Park, P.J. (2024). Accurate and sensitive mutational signature analysis with MuSiCal. Nat. Genet. *56*, 541–552. https://doi.org/10.1038/s41588-024-01659-0.

55. Musunuri, R.L., Zhu, B., Clarke, W.E., Hooper, W., Chu, T., Shelton, J., Corvelo, A., Chung, D., Sundar, S., Novak, A.M., et al. (2026). Lancet2: Improved and accelerated somatic variant calling with joint multi-sample local assembly graphs. NAR Genomics and Bioinformatics. https://doi.org/10.1093/nargab/lqag036.

56. Dou, Y., Kwon, M., Rodin, R.E., Cortés-Ciriano, I., Doan, R., Luquette, L.J., Galor, A., Bohrson, C., Walsh, C.A., and Park, P.J. (2020). Accurate detection of mosaic variants in sequencing data without matched controls. Nat. Biotechnol. *38*, 314–319. https://doi.org/10.1038/s41587-019-0368-8.

57. Chen, L., Zheng, Z., Su, J., Yu, X., Wong, A.O.K., Zhang, J., Lee, Y.-L., and Luo, R. (2025). ClairS-TO: a deep-learning method for long-read tumor-only somatic small variant calling. Nat. Commun. *16*, 9630. https://doi.org/10.1038/s41467-025-64547-z.

58. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods *18*, 170–175. https://doi.org/10.1038/s41592-020-01056-5.

59. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

60. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595. https://doi.org/10.1093/bioinformatics/btp698.

61. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

62. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv.

63. Vollger, M.R., Dishuck, P.C., Sorensen, M., Welch, A.E., Dang, V., Dougherty, M.L., Graves-Lindsay, T.A., Wilson, R.K., Chaisson, M.J.P., and Eichler, E.E. (2019). Long-read sequence and assembly of segmental duplications. Nat. Methods *16*, 88–94. https://doi.org/10.1038/s41592-018-0236-3.

64. Jha, A., Bohaczuk, S.C., Mao, Y., Ranchalis, J., Mallory, B.J., Min, A.T., Hamm, M.O., Swanson, E., Dubocanin, D., Finkbeiner, C., et al. (2024). DNA-m6A calling and integrated long-read epigenetic and genetic analysis with fibertools. Genome Res. *34*, 1976–1986. https://doi.org/10.1101/gr.279095.124.

65. Hickey, G., Monlong, J., Ebler, J., Novak, A.M., Eizenga, J.M., Gao, Y., Human Pangenome Reference Consortium, Marschall, T., Li, H., and Paten, B. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat. Biotechnol. *42*, 663–673. https://doi.org/10.1038/s41587-023-01793-w.

66. Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. *36*, 875–879. https://doi.org/10.1038/nbt.4227.

67. Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. *36*, 983–987. https://doi.org/10.1038/nbt.4235.

68. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339. https://doi.org/10.1093/bioinformatics/bts378.

69. Cameron, D.L., Baber, J., Shale, C., Valle-Inclan, J.E., Besselink, N., van Hoeck, A., Janssen, R., Cuppen, E., Priestley, P., and Papenfuss, A.T. (2021). GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. Genome Biol. *22*, 202. https://doi.org/10.1186/s13059-021-02423-x.

70. Saunders, C.T., Holt, J.M., Baker, D.N., Lake, J.A., Belyeu, J.R., Kronenberg, Z., Rowell, W.J., and Eberle, M.A. (2025). Sawfish: improving long-read structural variant discovery and genotyping with local haplotype modeling. Bioinformatics *41*, btaf136. https://doi.org/10.1093/bioinformatics/btaf136.

71. Keskus, A.G., Bryant, A., Ahmad, T., Yoo, B., Aganezov, S., Goretsky, A., Donmez, A., Lansdon, L.A., Rodriguez, I., Park, J., et al. (2026). Severus detects somatic structural variation and complex rearrangements in cancer genomes using long-read sequencing. Nat. Biotechnol. *44*, 247–257. https://doi.org/10.1038/s41587-025-02618-8.

72. Wala, J.A., Bandopadhayay, P., Greenwald, N.F., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S., Li, Y., Weischenfeldt, J., Yao, X., et al. (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. *28*, 581–591. https://doi.org/10.1101/gr.221028.117.

73. Denti, L., Khorsand, P., Bonizzoni, P., Hormozdiari, F., and Chikhi, R. (2023). SVDSS: structural variation discovery in hard-to-call genomic regions using sample-specific strings from accurate long reads. Nat. Methods *20*, 550–558. https://doi.org/10.1038/s41592-022-01674-1.

74. Zhou, W., Emery, S.B., Flasch, D.A., Wang, Y., Kwan, K.Y., Kidd, J.M., Moran, J.V., and Mills, R.E. (2020). Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. Nucleic Acids Res. *48*, 1146–1163. https://doi.org/10.1093/nar/gkz1173.

75. Chu, C., Borges-Monroy, R., Viswanadham, V.V., Lee, S., Li, H., Lee, E.A., and Park, P.J. (2021). Comprehensive identification of transposable element insertions using multiple sequencing technologies. Nat. Commun. *12*, 3836. https://doi.org/10.1038/s41467-021-24041-8.

76. Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., and 1000 Genomes Project Consortium, and Devine, S.E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. *27*, 1916–1929. https://doi.org/10.1101/gr.218032.116.

77. Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., and Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. *21*, 189. https://doi.org/10.1186/s13059-020-02107-y.

78. Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., et al. (2024). Detection of mosaic and population-level structural variants with Sniffles2. Nat. Biotechnol. *42*, 1571–1580. https://doi.org/10.1038/s41587-023-02024-y.

79. McDonald, T.L., Zhou, W., Castro, C.P., Mumm, C., Switzenberg, J.A., Mills, R.E., and Boyle, A.P. (2021). Cas9 targeted enrichment of mobile elements using nanopore sequencing. Nat. Commun. *12*, 3586. https://doi.org/10.1038/s41467-021-23918-y.

80. Smit, A.F.A., Hubley, R., and Green, P. (1996). RepeatMasker Open-4.0. https://www.repeatmasker.org.

81. Dong, G., Ma, C.C., Mao, S., Brown, K.S.-M., Naik, S.M., McDonough, G.A., Wijethunga, S.P., Kim, J., Kirkham, S.L., Shao, D.D., et al. (2025). Diverse somatic genomic alterations in single neurons in chronic traumatic encephalopathy. Science *390*, eadu1351. https://doi.org/10.1126/science.adu1351.

82. Zhao, B., Wu, Q., Ye, A.Y., Guo, J., Zheng, X., Yang, X., Yan, L., Liu, Q.-R., Hyde, T.M., Wei, L., et al. (2019). Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. PLOS Genet. *15*, e1008043. https://doi.org/10.1371/journal.pgen.1008043.

83. Liu, M.H., Costa, B.M., Bianchini, E.C., Choi, U., Bandler, R.C., Lassen, E., Grońska-Pęski, M., Schwing, A., Murphy, Z.R., Rosenkjær, D., et al. (2024). DNA mismatch and damage patterns revealed by single-molecule sequencing. Nature *630*, 752–761. https://doi.org/10.1038/s41586-024-07532-8.

84. Bae, J.H., Liu, R., Roberts, E., Nguyen, E., Tabrizi, S., Rhoades, J., Blewett, T., Xiong, K., Gydush, G., Shea, D., et al. (2023). Single duplex DNA sequencing with CODEC detects mutations with high sensitivity. Nat. Genet. *55*, 871–879. https://doi.org/10.1038/s41588-023-01376-0.

85. Cheng, A.P., Rusinek, I., Sossin, A., Widman, A.J., Meiri, E., Krieger, G., Hirschberg, O., Tov, D.S., Gilad, S., Jaimovich, A., et al. (2025). Paired plus-minus sequencing is an ultra-high throughput and accurate method for dual strand sequencing of DNA molecules. Preprint at bioRxiv, 2025.08.11.669689. https://doi.org/10.1101/2025.08.11.669689.

86. Coorens, T.H.H., Spencer Chapman, M., Williams, N., Martincorena, I., Stratton, M.R., Nangalia, J., and Campbell, P.J. (2024). Reconstructing phylogenetic trees from genome-wide somatic mutations in clonal samples. Nat. Protoc. *19*, 1866–1886. https://doi.org/10.1038/s41596-024-00962-8.

87. Suvakov, M., Panda, A., Diesh, C., Holmes, I., and Abyzov, A. (2021). CNVpytor: a tool for copy number variation detection and analysis from read depth and allele imbalance in whole-genome sequencing. GigaScience *10*, giab074. https://doi.org/10.1093/gigascience/giab074.

88. Zhao, Y., Luquette, L.J., Veit, A.D., Wang, X., Xi, R., Viswanadham, V.V., Zhang, Y., Shao, D.D., Walsh, C.A., Yang, H.W., et al. (2025). High-resolution detection of copy number alterations in single cells with HiScanner. Nat. Commun. *16*, 5477. https://doi.org/10.1038/s41467-025-60446-5.

89. Abyzov, A., Mariani, J., Palejev, D., Zhang, Y., Haney, M.S., Tomasini, L., Ferrandino, A.F., Rosenberg Belmaker, L.A., Szekely, A., Wilson, M., et al. (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature *492*, 438–442. https://doi.org/10.1038/nature11629.

90. Johnson, B.K., Majewski, M.F., Jang, H.J., Siwicki, R.A., Wegener, M., Semwal, A., Morrison, J., Gallik, K., Hagemann-Jensen, M., Foy, K., et al. (2025). Efficient profiling of total RNA in single cells with STORM-seq. Preprint at bioRxiv, 2022.03.14.484332. https://doi.org/10.1101/2022.03.14.484332.

91. Grochowski, C.M., Gibbs, R.A., and Doddapaneni, H. (2024). Validation of single-nucleotide mosaic variants through droplet digital PCR. Curr. Protoc. *4*, e1041. https://doi.org/10.1002/cpz1.1041.

92. Abascal, F., Harvey, L.M.R., Mitchell, E., Lawson, A.R.J., Lensing, S.V., Ellis, P., Russell, A.J.C., Alcantara, R.E., Baez-Ortega, A., Wang, Y., et al. (2021). Somatic mutation landscapes at single-molecule resolution. Nature *593*, 405–410. https://doi.org/10.1038/s41586-021-03477-4.

93. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498. https://doi.org/10.1038/ng.806.

94. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinform. *43*, 11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43.

95. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

96. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

97. Aldana, R., and Freed, D. (2022). Data processing and germline variant calling with the sentieon pipeline. Methods Mol. Biol. *2493*, 1–19. https://doi.org/10.1007/978-1-0716-2293-3_1.

98. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311. https://doi.org/10.1093/nar/29.1.308.

99. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

100. De Coster, W., and Rademakers, R. (2023). NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics *39*, btad311. https://doi.org/10.1093/bioinformatics/btad311.

101. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257. https://doi.org/10.1186/s13059-019-1891-0.

102. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

103. Pedersen, B.S., and Quinlan, A.R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics *34*, 867–868. https://doi.org/10.1093/bioinformatics/btx699.

104. Pedersen, B.S., Bhetariya, P.J., Brown, J., Kravitz, S.N., Marth, G., Jensen, R.L., Bronner, M.P., Underhill, H.R., and Quinlan, A.R. (2020). Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. Genome Med. *12*, 62. https://doi.org/10.1186/s13073-020-00761-2.

105. Zhang, F., Flickinger, M., Taliun, S.A.G., InPSYght Psychiatric Genetics Consortium, Abecasis, G.R., Scott, L.J., McCaroll, S.A., Pato, C.N., Boehnke, M., and Kang, H.M. (2020). Ancestry-agnostic estimation of DNA sample contamination from sequence reads. Genome Res. *30*, 185–194. https://doi.org/10.1101/gr.246934.118.

106. Kokot, M., Dlugosz, M., and Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. Bioinformatics *33*, 2759–2761. https://doi.org/10.1093/bioinformatics/btx304.

107. Sirén, J., Eskandar, P., Ungaro, M.T., Hickey, G., Eizenga, J.M., Novak, A.M., Chang, X., Chang, P.-C., Kolmogorov, M., Carroll, A., et al. (2024). Personalized pangenome references. Nat. Methods *21*, 2017–2023. https://doi.org/10.1038/s41592-024-02407-2.

108. Sirén, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C., Sibbesen, J.A., Hickey, G., Chang, P.-C., Carroll, A., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science *374*, abg8871. https://doi.org/10.1126/science.abg8871.

109. Talenti, A., and Prendergast, J. (2021). Nf-LO: A scalable, containerized workflow for genome-to-genome lift over. Genome Biol. Evol. *13*, evab183. https://doi.org/10.1093/gbe/evab183.

110. Cleary, J.G., Braithwaite, R., Gaastra, K., Hilbush, B.S., Inglis, S., Irvine, S.A., Jackson, A., Littin, R., Nohzadeh-Malakshah, S., Rathod, M., et al. (2014). Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. J. Comput. Biol. *21*, 405–419. https://doi.org/10.1089/cmb.2014.0029.

111. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53. https://doi.org/10.1126/science.abj6987.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Cell lines		
COLO829 lymphoblastoid line	ATCC	COLO829BL Cat# CRL-1980, Lot# 70022927; RRID:CVCL_1999
COLO829 melanoma line	ATCC	COLO829 Cat# CRL-1974, Lot# 70024393; RRID:CVCL_1137
COLO829 49:1 mixture	UW-SCRI	COLO829BLT50
HG005 lymphoblastoid cell line	Coriell Institute	HG005 Cat# GM24631; RRID:CVCL_1C97
HG02622 lymphoblastoid cell line	Coriell Institute	HG02622 Cat# HG02622; RRID:CVCL_GV49
HG002 lymphoblastoid cell line	Coriell Institute	HG002 Cat# GM24385; RRID:CVCL_1C78
HG02257 lymphoblastoid cell line	Coriell Institute	HG02257 Cat# HG02257; RRID:CVCL_Q703
HG02486 lymphoblastoid cell line	Coriell Institute	HG02486 Cat# HG02486
HG00438 lymphoblastoid cell line	Coriell Institute	HG00438 Cat# HG00438; RRID:CVCL_P889
HapMap lymphoblastoid cell line mixture	WashU-VAI	HapMap6
Biological samples		
LA2 primary fibroblasts from donor LB	Yale University	LB-LA2
iPSC line #1 from LA2 fibroblasts	Yale University	LA2#1
iPSC line #2 from LA2 fibroblasts	Yale University	LA2#2
iPSC line #4 from LA2 fibroblasts	Yale University	LA2#4
iPSC line #52 from LA2 fibroblasts	Yale University	LA2#52
iPSC line #60 from LA2 fibroblasts	Yale University	LA2#60
Liver from donor ST001	TPC/NDRI	ST001-1A; RRID:SCR_000550
Lung from donor ST001	TPC/NDRI	ST001-1D; RRID:SCR_000550
Skin from donor ST001	TPC/NDRI	ST001-1K; RRID:SCR_000550
Colon from donor ST002	TPC/NDRI	ST002-1G; RRID:SCR_000550
Lung from donor ST002	TPC/NDRI	ST002-1D; RRID:SCR_000550
Skin from donor ST002	TPC/NDRI	ST002-1K; RRID:SCR_000550
Brain from donor ST003	TPC/NDRI	ST003-1Q; RRID:SCR_000550
Brain from donor ST004	TPC/NDRI	ST004-1Q; RRID:SCR_000550
Deposited data		
Data for COLO829, COLO829BL, COLO829BLT50	dbGaP	phs004193
Data for COLO829, COLO829BL, COLO829BLT50	SMaHT DAC	https://data.smaht.org
Data for HapMap6	dbGaP	phs004193
Data for HapMap6	SMaHT DAC	https://data.smaht.org
Data for LB-LA2, LA2#1, LA2#2, LA2#4, LA2#52, LA2#60	dbGaP	phs004425
Data for ST001, ST002, ST003, ST004	dbGaP	phs004193
Software and algorithms		
BSMN pipeline	Wang et al.[11]; Bae et al.[44]	https://github.com/bsmn/bsmn-pipeline
All[2]	Sarangi et al.[45]	https://github.com/abyzovlab/All2
TNHaplotyper2 (Sentieon Genomics 202112.06)	Freed et al.[46]	https://support.sentieon.com/docs/usages/driver/TNhaplotyper2
Haplotyper (Sentieon Genomics 202112.06)	Freed et al.[46]	https://support.sentieon.com/docs/usages/driver/Haplotyper
TNFilter (Sentieon Genomics 202308.03)	Freed et al.[46]	https://support.sentieon.com/docs/usages/driver/TNfilter

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Strelka2 (v2.9.2)	Kim et al.[47]	https://github.com/Illumina/strelka
VarNet (v1.1.0)	Krishnamachari et al.[48]	https://github.com/skandlab/VarNet
GATK	Auwera & O'Connor[49]	https://gatk.broadinstitute.org
Mutect2 (v4.4.0.0)	Cibulskis et al.[50]	https://gatk.broadinstitute.org
RUFUS	Ostrander et al.[51]; Richter et al.[52]	https://github.com/jandrewrfarrell/RUFUS
Manta (v1.6.0)	Chen et al.[53]	https://github.com/Illumina/manta
MuSiCal	Jin et al.[54]	https://github.com/parklab/MuSiCal
Dynamic Read Analysis for GENomics (DRAGEN) (v4.3)	Illumina	https://www.illumina.com/products/by-type/informatics-products/dragen-secondary-analysis.html
Lancet2 (2.8.5-main-cfbc7d5143)	Musunuri et al.[55]	https://github.com/nygenome/Lancet2
MosaicForecast	Dou et al.[56]	https://hub.docker.com/r/yanmei/mosaicforecast
ClairS-TO	Chen et al.[57]	https://github.com/HKU-BAL/ClairS-TO
Verkko v2.2.1	Rautiainen et al.[34]	https://github.com/marbl/verkko
Hifiasm	Cheng et al.[58]	https://github.com/chhylp123/hifiasm
Minimap2 v2.28	Li et al.[59]	https://github.com/lh3/minimap2
pbmm2 v1.14.99	PacBio Inc.	https://github.com/PacificBiosciences/pbmm2
bwa v0.7.17 bwa v0.7.19	Li et al.[60]; Li et al.[61]; Li et al.[62]	https://github.com/lh3/BWA
Flagger v0.3.3	Liao et al.[35]	https://github.com/mobinasri/flagger
NucFlag v0.3.3	Vollger et al.[63]	https://github.com/logsdon-lab/NucFlag
fibertools v0.6.0	Jha et al.[64]	https://github.com/fiberseq/fibertools-rs
FIRE v0.1.2	Vollger et al.[43]	https://github.com/fiberseq/FIRE
Minigraph-cactus v2.9.9	Hickey et al.[65]	https://github.com/lh3/minigraph, https://github.com/ComparativeGenomicsToolkit/cactus
vg v1.65.0	Garrison et al.[66]	https://github.com/vgteam/vg
DeepVariant v1.6.1	Poplin et al.[67]	https://github.com/google/deepvariant
DeepSomatic v1.9.0	Jimin et al.[40]	https://github.com/google/deepsomatic
Picard v2.23.3	GitHub	http://broadinstitute.github.io/picard
Delly	Rausch et al.[68]	https://github.com/dellytools/delly
GRIDSS	Cameron et al.[69]	https://github.com/PapenfussLab/gridss
pbsv	GitHub	https://github.com/PacificBiosciences/pbsv
Sawfish	Saunders et al.[70]	https://github.com/PacificBiosciences/sawfish
Severus	Keskus et al.[71]	https://github.com/KolmogorovLab/Severus
SvABA	Wala et al.[72]	https://github.com/walaj/svaba
SVDSS	Denti et al.[73]	https://github.com/Parsoa/SVDSS
PALMER v2.0.1	Zhou et al.[74]	https://github.com/WeichenZhou/PALMER
xTea_mosaic v0.1.9	Chu et al.[75]	https://github.com/parklab/xTea
xTea_long v0.1.0	Chu et al.[75]	https://github.com/parklab/xTea
RetroSom v2	Zhu et al.[12]	https://github.com/Czhuofu/RetroSomV2
MELT v2.2.2	Gardner et al.[76]	https://melt.igs.umaryland.edu/index.php
cuteSV v2.1.2	Jiang et al.[77]	https://github.com/tjiangHIT/cuteSV
Sniffles2 v.2.6.3	Smolka et al.[78]	https://github.com/fritzsedlazeck/Sniffles
NanoPal	McDonald et al.[79]	https://github.com/Boyle-Lab/NanoPal-Snakemake
RepeatMasker	Smit et al.[80]	http://www.repeatmasker.org
Minimera	Losh et al.[81]	https://github.com/Boyle-Lab/minimera/
eHAT-seq (v0.1.0)	Zhao et al.[82]	https://github.com/ealee-lab/HATseq-pipeline/
HiDEF-seq v3.1	Liu et al.[83]	https://github.com/evronylab/HiDEF-seq
CODECsuite	Bae et al.[84]	https://github.com/broadinstitute/CODECsuite
CompDuplex pipeline	Niu et al.[27]	https://github.com/zonglab/CompDuplex

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PPMseq	Chen et al.[85]	https://github.com/Ultimagen/healthomics-workflows/tree/main/workflows/single_read_snv
VISTA-seq pipeline	Dong et al.[81]	https://github.com/gldong/duplex-snv
SCAN2	Luquette et al.[7]	https://github.com/parklab/SCAN2
Sequoia	Coorens et al.[86]	https://github.com/TimCoorens/Sequoia
CNVpytor	Suvakov et al.[87]	https://github.com/abyzovlab/CNVpytor
HiScanner	Zhao et al.[88]	https://github.com/parklab/HiScanner
HDP	GitHub	https://github.com/nicolaroberts/hdp

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

LB is a living 29-year-old male donor. ST001 is 22-year-old male. ST002 is a 74-year-old male donor. ST003 is a 22-year-old male. ST004 is a 73-year-old male.

Informed consent was obtained from the LB participant according to the regulations of the Institutional Review Board (IRB protocol # 1104008337) and Yale Center for Clinical Investigation at Yale University.

Authorization for tissue donation for donors ST001 and ST002 was provided by the donor next-of-kin at a TPC-partnering organ procurement organization (OPO). Ethical approval for the tissue collection was obtained by National Disease Research Interchange (NDRI) through the University of Pennsylvania (IRB#5 FWA00004028). Frozen dorsolateral prefrontal cortex tissue from deceased donors ST003 and ST004 was obtained from the NIH NeuroBioBank repository at UMBTB. Collection of brain tissue from donors was performed as approved by the IRB of the University of Maryland, Baltimore and the University of Miami School of Medicine where the respective cases were recovered. Due to the authorization structure, only the age and sex of the benchmarking post-mortem donors can be made public.

METHOD DETAILS

Sample acquisition, preparation and distribution

Preparing HapMap samples and mixture, contributed by Sheng Chih Jin and Benpeng Miao

Each cell line was individually thawed, recovered, and expanded before pooling. Cells were cultured in T25 flasks containing 10 ml culture medium (84% RPMI-1640, 15% FBS, and 1% Glutamax), maintained within the optimal viable concentration range (2×10^5 – 5×10^5 cells/ml), and split by dilution every 3–5 days based on cell counts measured using a Vi-Cell Blu cell counter (Beckman Coulter) with trypan blue viability assessment. On the day of harvest, cells from each line were pooled into a 250 ml conical tube, thoroughly mixed, and approximately 10 ml was transferred to a T25 flask. Aliquots were taken for cell counting and sterility assessment; the latter was performed by streaking samples onto two sheep's blood agar plates and incubating them for two weeks at 30°C and 37°C. Based on cell counts, volumes of individual cell suspensions were calculated to achieve the target mix ratio: 83.5% HG005, 10% HG02622, 2% each of HG002, HG02257, and HG02486, and 0.5% HG00438. The calculated volumes were combined in a secondary tube, centrifuged at 228 RCF for 10 minutes, resuspended in 800 ml cryopreservation medium (65% RPMI-1640, 30% FBS, and 5% DMSO), and dispensed into cryovials at 5×10^6 cells/vial. Cryovials were then cryopreserved using a controlled-rate freezer and stored in liquid nitrogen vapor.

Preparing COLO829 samples, contributed by Nancy Parmalee

COLO829BLT50 Cell Culture and Mixing Methods. Pure populations of COLO829BL cells (cat # CRL-1980, lot # 70022927) and COLO829 cells (cat # CRL-1974, lot # 70024393) were obtained from ATCC (6/21/2023) and expanded. All cells were grown at 37°C with 5% CO_2.

COLO829BL lymphoblastic suspension cells were thawed in recovery media which consisted of 1 part RPMI-1640 (Fisher cat# 11875093) 15% fetal bovine serum (FBS, Fisher cat# 10082147) to 1 part 100% FBS. Cells were centrifuged in recovery media at 300 rcf for 5 minutes to remove DMSO. The pellet was resuspended in recovery media and plated in a T-25 flask with 5 mL total media for a volume of 0.2 mL media per cm^2 to allow for maximum gas exchange. After approximately 2 days in culture cells were split by dilution into RPMI-1640 15% FBS and split by dilution daily thereafter. Flask sized was increased with expansion of the culture maintaining a volume of 0.2 ml media per cm^2. When the culture reached a volume of 100mL it was transferred to a 500 mL Erlenmeyer cell culture flask and maintained with shaking at 90 rpm. The maximum volume of the culture was 200 mL per 500 mL flask. COLO829BL cells were harvested on a continuous basis. Cells were harvested by centrifugation in 50 mL conical tubes at 300 rcf for 5 minutes. The pellet was resuspended in 500 mL RPMI-1640 15% FBS. Days in culture ranged from 9-18 days. Cells were counted using the Countess cell counter (Invitrogen Thermo Fisher) with trypan blue (Fisher cat#15250061) and cryopreserved in freezing media consisting of RPMI-1640 15% FBS with 10% DMSO (Sigma Aldrich cat# D2650-100ML) at a cell density of 3 million viable cells per mL in 2 mL cryovials and stored in liquid nitrogen vapor phase.

COLO829 adherent melanoma cells were thawed and plated directly into a T-75 culture flask in 15 mL RPMI-1640 10% FBS. Media was changed approximately every 4 days. Cells were grown to near confluence and split by removing media, washing with PBS pH 7.2 (Fisher cat# 20012027), and trypsinizing for five minutes at 37° C using 0.05% trypsin EDTA (Fisher cat# 25300062). Volumes of both PBS and trypsin were 5 mL for T-75 flasks and 15 mL for T-175 flasks. COLO829 cells were harvested at confluence from 90 T-175 flasks on a single day at passage 4. Cells from all 90 flasks were combined and mixed to equilibrate flask-specific somatic mutations across the mixture. Cells were counted using the Countess cell counter with trypan blue and cryopreserved in freezing media consisting of RPMI-1640 10% FBS with 10% DMSO at a cell density of 3 million viable cells per mL in 2 mL cryovials and stored in liquid nitrogen vapor phase. 433 vials of 2 mL per cryovial were produced.

On a single day cryopreserved COLO829BL and COLO829 cells were thawed from the previously cryopreserved vials, kept on ice, and counted using the Countess cell counter with trypan blue. Based on the Countess readings the COLO829 and COLO829BL cells were mixed at a 1:49 ratio using viable cell counts. Viability of the COLO829 cells was 93% and viability of the COLO829BL cells was 91%. Cells were aliquoted into cryovials with constant swirling to maintain the homogeneity of the mix. A total of 380 vials were cryopreserved at a cell density of 2.55 million viable cells per mL for a total of 5.1 cells per 2 mL vial and stored in liquid nitrogen vapor phase. To assure homogeneity of the mixture DNA was isolated from the first vial aliquoted and the last vial aliquoted to test for differential settling of cells during the mixing process. Droplet digital PCR assays of the final mix showed no significant difference between the first and the last vial.

COLO829BLT50 Bio-Rad Droplet Digital PCR (ddPCR) Quality Control Methods. Bio-Rad–designed droplet digital PCR (ddPCR) assays were used to target 5 mutations present in COLO829 tumor cells detected: *MADD* p.S1620F, *MAP4K1* p.P298P, *RELN* p.E331K, *SCN11A* p.T340T, *ZNF217* p.P651S. Each assay was validated by a temperature gradient experiment to determine the annealing temperature for ideal cluster separation. All PCR reactions were set up in a UV-treated hood with positive airflow, and samples were run in quadruplicate. Each 22-μL PCR mixture contained 10 μL of ddPCR SuperMix, 1 μL of assay, 1 μL (10 ng) of DNA, and 10 μL of water.

Twenty microliters of the PCR mix were used to generate droplets using the Bio-Rad Droplet Generator. PCR was performed using the following parameters: 10 minutes at 95°C, followed by 40 cycles of 30 seconds at 94°C, 60 seconds at the annealing temperature of the assay (52.7°C for *MAP4K1* p.P298P, *SCN11A* p.T340T, *ZNF217* p.P651S and 55°C for *MADD* p.S1620F, *RELN* p.E331K), and 10 minutes at 98°C, and held at 12°C. We used the Bio-Rad QX200 Droplet Reader within 24 hours and analyzed the data with Bio-Rad Quantasoft software. Each run included a no-template control, WT control, and positive mutant control.

Samples were positive if the variant fluorescence was significantly different from the fluorescence of the WT control using 95% confidence intervals for total error. The total error is displayed by the Quantasoft software and defined as the greater of either the technical error (Poisson error) or the empirical error (standard error of the mean). Variant allele fractions (VAFs) were calculated as the concentration of variant droplets out of the total concentration of droplets containing at least one copy of variant or WT DNA.

Preparing iPSC and fibroblast samples, contributed by Livia Tomasini

Fibroblasts sample collection and iPSC generation. The skin biopsy was collected from the inner side of the left upper arm and fibroblast primary cultures were selectively expanded as previously described[32,89] using the explant method and fibroblast culture media (DMEM high-glucose-based medium supplemented with 10% FBS, 1:100 L-glutamine, 1:100 N.E. amino acids, 1:100 Pen/Strep, and 10 ng/mL FGF2, all reagents from Gibco Life Technologies). Cells were passed twice using TrypLE Express (Gibco Life Technologies). One million cells per vial were cryopreserved using Recovery Cell Culture Freezing Medium (Thermo Fisher Scientific). iPSCs were generated using the Epi5 Episomal iPSC Reprogramming Kit (Thermo Fisher Scientific) following manufacturer's recommendations. The iPSC lines were propagated using mTeSR1 media (Stem Cell Technologies) and passed using Dispase (Stem Cell Technologies). After expansion, iPSC lines were cryopreserved in Cryostor CS10 (Stem Cell Technologies).

LB-LA2 fibroblasts culture and expansion. One cryovial of the fibroblast cell line from individual LB, left arm location (line LB-LA2) was thawed in fibroblast culture media (DMEM high-glucose-based medium supplemented with 10% FBS, 1:100 L-glutamine, 1:100 N.E. amino acids, 1:100 Pen/Strep, and 10 ng/mL FGF2, all reagents from Gibco Life Technologies). After centrifugation at 1000 rpm for 4 minutes, cells were plated in a tissue culture treated 75 cm^2 flask (Falcon). Cells were cultured at 37°C with 5% CO2 in room air and the medium was changed every three days. Cells were passed and expanded when reaching about 70% confluency using TrypLE Express (Gibco Life Technologies), following manufacturer's recommendations. The initial 75 cm^2 flask was passed after two days in culture and plated in three 75 cm^2 flasks. Three days later, cells were passed to five 175 cm^2 flasks using the same procedure. Four days later, the cells were passed to twenty 175 cm^2 flasks. After five days, the cells were passed again to forty 175 cm^2 flasks. After five days, the cells were trypsinized as above, the pellets were resuspended in fresh fibroblast culture media and counted using a manual hemocytometer. Cells were resuspended in Recovery Freezing media (Gibco Life Technologies) at a density of 3 million cells per mL. A total of 52 cryovials with 2 mL each (6 million cells per vial) were produced at passage 6 to minimize culture-acquired mutations. This procedure ensured that each site received the identical fibroblast cell suspension for sequencing.

LB-LA2 iPSC clones culture and expansion. iPSC clones derived from the fibroblast line LB-LA2 (LB-LA2 clone #1; clone #2; clone #4; clone #52 and clone #60) were thawed in mTeSR1 media (Stem Cell Technologies), centrifuged at 800 RPM for 4 minutes and each plated in a well of a 6-well tissue culture treated plate (Corning), pre-coated with a 1% solution of Geltrex LDEV-Free hESC-qualified Reduced Growth Factor Basement Membrane Matrix (Gibco Life Technologies). iPSC lines were cultured at 37°C, 5% CO2 and 5% O2. The media was changed every two to three days. After one week, the cells were passed using Gentle Cell Dissociation Reagent (Stem Cell Technologies) following manufacturer 's instructions. Each clone was plated on one 10cm diameter tissue

culture dish (Falcon) pre-coated with Matrigel as above. Between 5 to 8 days later depending on the clone, cells were passed again following the same procedure, and each plated on four 10cm Matrigel pre-coated dishes. iPSCs were harvested between 7 to 9 days later depending on the clones. At the time of harvesting, iPSC colonies were dissociated using TrypLE Express (Gibco Life Technologies), resuspended in fresh mTeSR1 media and counted using a manual hemocytometer. The single cell suspensions were resuspended in Cryostor CS10 (Stem Cell Technologies) at a density of 3 million cells per mL and 2 mL were aliquoted per cryovial (6 million cells per vial). A total of four cryovials were produced for clone #4 at passage 6; a total of five cryovials for clone #1 and for clone #2, both at passage 6; a total of two cryovials for clone #60 at passage 5; and a total of three cryovials for clone #52 at passage 4.

Sample distribution

The shipping of the fibroblast sample to all GCCs was done at once. Cells for each of the 5 iPSC lines were shipped to Broad GCC for genomic characterization.

Human Tissue Collection, Preparation, and Distribution, contributed by M. Kathryn Leonard, Melissa VonDran, Eric Ho, Thomas Blanchard, and Thomas J. Bell

Non-brain tissues for the SMaHT benchmarking study were recovered by the TPC from two post-mortem donors. Donors over the age of 18 years were considered eligible for inclusion in the study if tissues could be collected within 24 hours of death. Samples of liver, lung, and colon were recovered from each donor using SMaHT-specific recovery, processing, and preservation protocols at the OPO. Briefly, for each tissue, five 10 cm x 1 cm x 1 cm samples from defined anatomical locations in each organ were rinsed with saline, cut into several smaller pieces, placed in individual 50 ml conical tubes and snap frozen. A companion 1 cm3 sample from each tissue was preserved in 10% formalin, praffin-embedded, and stained with Hematoxylin and Eosin (H&E) for subsequent histological analysis. Each tissue was reviewed by a pathologist at the University of Maryland Brain and Tissue Bank (UMBTB) to evaluate tissue composition and general pathological features such as inflammation, and organ-specific pathological conditions. Frozen dorsolateral prefrontal cortex tissue from deceased donors ST003 and ST004 was obtained from the NIH NeuroBioBank repository at UMBTB.

Homogenized Tissue Samples

To minimize regional variability and ensure consistency across sequencing centers, frozen tissues were pulverized into a powder using a pre-chilled mortar and pestle with liquid nitrogen. The homogenized material was pooled, mixed for uniformity, and aliquoted into pre-weighed cryovials using pre-chilled spatulas, then stored at -80°C at the TPC for distribution to GCCs or TTDs.

Intact, Non-homogenate Tissue Samples

For experimental procedures that required intact tissue, pre-chilled forceps were utilized to breakdown intact frozen tissue samples (~1 gram) into approximately 150mg aliquots. The intact tissues were aliquoted into pre-weighed cryovials using pre-chilled spatulas, then stored at -80°C at the TPC. Sample aliquots were distributed by TPC on dry ice to GCCs or TTDs.

Molecular Quality Control

RNA and DNA quality from snap-frozen tissue samples was examined by UMBTB prior to distribution of samples to sequencing centers. Tissues were lysed with ZrOBZO beads in a Bullet Blender (Next Advance) followed by isolation of RNA and DNA via QIACube using the RNeasy Lipid Tissue and DNeasy Blood and Tissue Kits, respectively, (Qiagen). RNA quality was assessed using the Agilent 2100 Bioanalyzer to generate RNA Integrity Numbers (RIN). Samples with RIN scores of 4.0 or higher were included for distribution and further analysis. DNA quality was assessed by measuring concentration and purity using a Nanodrop spectrophotometer. In addition, each GGC or TTD also performed quality control assays to confirm sample suitability for their experimental procedures.

Donor Metadata

The clinical data collected for each SMaHT donor belong to one of two categories: donor-level data or tissue sample-level data. Donor-level data encompass all clinical measures of the donor, including donor demographics (age, sex, height, weight, and body mass index (BMI)), medical history (cause of death and past medical history), social history (alcohol, tobacco, and illicit drug use), and serological testing results. Details on access to donor level data are contained within the official SMaHT data use policy available through the SMaHT web portal at (https://data.smaht.org).

Tissue Sample Metadata

Tissue sample-level data are attributes belonging to each sample collected and include the tissue type, ischemic time, comments from the prosector and pathology reviewer, and process metadata such as batch ID.

Data generation

Sequencing at WashU-VAI, contributed by Sheng Chih Jin, Robert Fulton, Ben Johnson, and Mary Majewski

Long-Read Sequencing (PacBio HiFi SPRQ Chemistry). PacBio HiFi SMRTbell libraries was prepared following PacBio protocol 'Procedure & Checklist – Preparing Whole Genome and Metagenome Libraries Using SMRTbell Prep Kit 3.0'. Genomic DNA was fragmented with a mode of ~20 kb using the Diagenode Megaruptor 3 instrument. Genomic DNA is initially processed with DNAFluid+ (P/N E07020001) using speed of 40 to dissociate aggregates and homogenize the DNA. The homogenized DNA was then sheared twice using Shearing kit (P/N E07010003) with speeds of 28 and 30. Sheared sample was assessed via fluorometry (Qubit High Senstivity DNA Kit) and Agilent Femto Pulse (Genomic DNA 165kb Kit). Libraries are made according to PacBio protocol utilizing barcoded adapters from the SMRTbell barcoded adapter plate 3.0 (PacBio P/N 102-009-200) to allow for multiplexing of samples during sequencing. Libraries were size selected using Sage PippinHT instrument and the 0.75% Agarose High-Pass 75E kit (P/N HPE7510) with a start size of 15000bp-17000bp. Size selected libraries were prepared for sequencing following instructions

generated in PacBio SMRT Link v13.0 Sample Setup and utilizing PacBio Revio polymerase kit (P/N 102-817-600). Sequencing was performed on PacBio Revio sequencer with an 'On Plate Concentration' of 170pM-200pM. 3 SMRTcells were generated for a total of ~100×/SMRTcell and 300Gb total coverage. Average Q-score values of Q33, Q33, and Q32 for the 3 SMRTcells was achieved.

Watchmaker Stranded Total RNA-seq. Total RNA integrity was determined using Agilent Bioanalyzer or 4200 Tapestation. Library preparation was performed with 100ng to 500ng of total RNA. Libraries were generated with Watchmaker Library Prep Kit with Polaris Depletion (Watchmaker). Briefly, ribosomal RNA was removed by an RNaseH method and purified with RNAClean beads (Beckman). mRNA was then fragmented in buffer by heating depending on RNA quality per protocol. mRNA was reverse transcribed to yield strand specific cDNA. A second strand and A Tailing reaction was performed to yield fragments with an A base added to the 3' ends. Illumina sequencing adapters were ligated to the ends. Ligated fragments were then amplified per protocol using primers incorporating unique dual index tags. Fragments were sequenced on an Illumina NovaSeq X Plus using paired end reads extending 150 bases.

KAPA Hyper PCR-Free Library Prep. Genomic DNA samples were quantified using the Qubit Fluorometer. Genomic DNA (~600-1000ng) was fragmented on the Covaris LE220 instrument targeting ~375bp inserts. Fragmented DNA was size selected using 0.8X ratio of Ampure XP beads (Beckman Coulter) to remove fragments less than 300bp. Dual indexed libraries were constructed utilizing the KAPA Hyper PCR-free library prep kit (Roche Diagnostics, Cat # 7962371001). Full length custom adaptors were used during ligation (IDT, UDI/UMI configuration with 10bp UDIs and a 9bp UMI in the i7 position). Libraries were run with KAPA Library Quantification kit (Roche Diagnostics) to measure molar concentration. Libraries are sequenced on NovaSeq X using paired end reads extending 150bp. For this application targeting 500× coverage, 4 libraries were constructed and utilized to generate the >500× coverage.

Single-cell Total RNA-seq (STORM-seq). Single cell suspensions from the benchmarking cell line mixtures were rapidly thawed in a 37C water bath and resuspended in increasing 1:1 dropwise volumes of warm (37C) flow buffer (HBSS with no divalent cations + 2% FBS + 25 mM HEPES) with gentle agitation. Resuspended cells were then pelleted at 300 x g at 4C for 5 minutes. Cells were washed in warm flow buffer and re-pelleted. Next, cells were counted and resuspended at 1 x 10 ^6 cells/mL in warm flow buffer containing 0.5 ug/mL DAPI for active viability surveillance during sorting. Single, live cells were index sorted into each well of a 384-well plate (Eppendorf) containing 2.17 uL Fragmentation Buffer (1.17 uL PBS pH 7.2 [Gibco 20012-027], 0.17 uL 10X Lysis Mix, 0.17 uL SMART scN6, 0.69 uL scRT buffer) using a BD FACSymphony S6 sorter running BD FACSDiva v9.1.3 and equipped with BD StepSort. The S6 was run with a 130 um nozzle at 14 psi. For cell deposition into 384-well plates, we used single cell sort mode. Single cell libraries were prepared using the STORM-seq kit (Takara Cat # 634751) and protocol (https://doi.org/10.5281/zenodo.15178455).[90] Briefly, after sorting, plates were immediately transferred to a pre-heated thermal cycler, heated to 85C for 3 minutes and snap cooled on ice for 2 minutes. Once cooled, 1.17 uL First Strand Master Mix (0.75 uL SMART scTSO mix, 0.08 uL RNase Inhibitor, 0.34 uL SMARTscribe RT) is added to each well. Immediately prior to dispensing, ERCC RNA Mix 1 (Thermo Fisher) is added to the First Strand Master Mix to create a 1:1000000 final dilution. First Strand Synthesis is performed at 42C for 180 minutes, followed by 70C for 10 minutes and 4C hold. After 1st Strand Synthesis, PCR1 (10 cycles of amplification: 94C 1 min, [98C 15s, 55C 15s, 68C 30s]x10, 68C 2 min, 4C hold) was performed with the addition of SMARTer RNA Unique Dual Index Sets A-D (SMARTer RNA Unique Dual Index Kits 96U sets A-D, Takara Cat # 634752, 634753, 634754, 634755). 4.67 uL of PCR1 Master Mix (0.33 uL nuclease free water, 4.16 uL SeqAmp CB PCR buffer, SeqAmp DNA Polymerase) is added to each well. Each well then receives 1 uL of a Unique Dual Index (UDI). UDIs were diluted 1:4 in 10mM Tris-HCl, pH 8.0 (Teknova) prior to addition. The plate was pooled and a bead-based cleanup (Beckman Coulter AMPure XP Beads) was performed. 162 uL of rRNA Depletion Master Mix (123.12ul nuclease free water, 16.2 uL 10X ZapR buffer, 11.02ul scZapR, 11.02ul sc-R Probes) is used to elute the depleted cDNA from the dried beads. The eluate is incubated at 37C for 60 min, 72C for 10 min, 4C hold. Finally, 12 cycles of amplification (94C 1 min, [98C 15s, 55C 15s, 68C 30s]x12, 4C hold) were performed for PCR2 (PCR2 Master Mix (208 uL nuclease free water, 400 uL SeqAmp CB PCR Buffer, 16ul PCR2 primers, 16 uL SeqAmp DNA polymerase) and the final library was eluted in 20 uL 10mM Tris-HCl pH 8.0. After QC, where necessary, an additional bead clean-up was performed to remove any remaining adapter-dimer.

Short-Read Sequencing. For both Kapa Hyper PCR-free libraries as well as the Bulk RNA-seq libraries, the molarity of each library was accurately determined through qPCR utilizing the KAPA library Quantification Kit according to the manufacturer's protocol (KAPA Biosystems/Roche) to produce cluster density appropriate for the Illumina NovaSeq X Plus instrument. Normalized libraries were sequenced on a NovaSeq X Plus Flow Cell using the151x10x10x151 sequencing recipe according to manufacturer protocol to generate a >500× WGS coverage, and >100M read pairs for the RNAseq library. For the STORM-seq libraries, sizing was performed using the Agilent Bioanalyzer HS kit and concentration was determined using a Qubit fluorometer and dsDNA HS kit before sequencing using an Illumina NovaSeq 6000 S2 2x150 bp flow cell. Libraries were sequenced to generate an average of 1M reads per cell.

Sequencing at UW, contributed by Jeffrey Ou, Christian D. Frazar, Caitlin N. Jacques, and Jeffrey M. Weiss

Short-Read DNA (Illumina WGS)

Short-Read whole genome sequencing was carried out using the Illumina NovaSeqX Plus sequencer. DNA was extracted from cell lines and tissues using the Qiagen DNAEasy Blood and Tissue kit (69506). Lysis time varied by tissue type and ranged from 4 hours to overnight. DNA from 1mL blood samples were extracted using the Qiagen MagAttract HMW DNA kit (67563) following the manufacturer's instructions. The QIAmp DNA Mini kit (51304) was used to purify DNA from buccal swabs, following the manufacturer's

instructions. The integrity of the extracted DNA was confirmed using the Agilent Femto Pulse instrument and the DNA yield was confirmed using the Invitrogen Qubit kit and Qubit 4 fluorometer. Starting with a minimum of 750ng of DNA, samples are sheared in a 96-well format using a Covaris R230 focused ultrasonicator targeting 380bp inserts. The resulting sheared DNA is cleaned with Takara NucleoMag beads to remove sample impurities prior to library construction. Shearing is followed by size selection and sample prep is performed using the KAPA Hyper Prep kit (KR0961 v1.14). End-repair, A-tailing, and ligation are performed as directed. Two final NucleoMag cleanups are performed after ligation to remove excess adapter dimers from the library. All library construction steps are automated on the Revvity Janus platform. Library yield is quantified using Invitrogen Quant-IT dsDNA High Sensitivity kit (Q33120). Libraries are validated in triplicate using the Biorad CFX384 Real-Time System and KAPA Library Quantification Kit (KK4824). Barcoded libraries are pooled using liquid handling robotics prior to loading. Massively parallel sequencing-by-synthesis with fluorescently labeled, reversibly terminating nucleotides is carried out on the NovaSeq X Plus sequencer. Base calls are generated in real-time on the instrument (RTA 4.29.2) and then demultiplexed, fastq files are produced by bcl-convert v4.2.7.

RNA Sequencing (PacBio bulk Kinnex)

Extracted total RNA was quality checked using UV-Vis spectroscopy (Denovix DS-11 FX) and Agilent Bioanalyzer 2100 using the Total RNA Nano 6000 kit (Agilent, G2939A & 5067-1511.) Kinnex full-length RNA libraries were generated per manufacturer's recommendations (PacBio, 103-072-000). Samples were sequenced on the Revio platform on SMRT Cells 25M with Revio Chemistry V1 (PacBio, 102-817-900) or SPRQ (PacBio, 103-520-200) with Adaptive Loading and 30-hour movies. Data were postprocessed using SMRT Link v13.1 or 13.3 with the ''Read Segmentation and Iso-Seq'' pipeline to segment and classify reads.

Long-Read DNA (PacBio HiFi/Fiber-Seq)

Extracted DNA samples were checked for quantity using Qubit dsDNA HS (Thermo Fisher, Q32854) measured on DS-11 FX (Denovix) and size distribution using FEMTO Pulse (Agilent, M5330AA & FP-1002-0275.) Depending on initial length distribution, samples were sheared to a target peak length of ~20 kbp. Some samples were left unsheared; others were sheared with Megaruptor 3 Hydropores (Diagenode, B06010003 and E07010003.) Moderately degraded or low-concentration samples were subjected to a light shear at setting 28 or 29, while intact DNA samples were sheared by processing twice, at settings 28 or 29 and 30 or 31. Highly intact and non-homogeneous DNA samples (e.g. derived from fibroblasts) were pre-sheared with Megaruptor 3 DNAFluid+ (Diagenode, E07020001) before final shearing. Sheared DNAs were subjected to PacBio HiFi library prep via the SMRTbell Prep Kit 3.0 (PacBio, 102-182-700) using barcoded adapters (PacBio, 102-009-200). When final library QC permitted (minimum of 500 ng of individual or pooled libraries,) size selection was performed with Pippin HT using a high-pass cutoff of 8-15 kbp (Sage Science, HTP0001 & HPE7510.) Low-mass libraries or those with short final size distributions (~5-10 kbp) were instead treated with a mild size-selection using diluted AMPure PB beads per the manufacturer's protocol. Libraries were sequenced on the Revio platform on SMRT Cells 25M with Revio Chemistry V1 (PacBio, 102-817-900) or SPRQ (PacBio, 103-520-200) with Adaptive Loading and 30-hour movies.

Long-Read DNA (ONT Standard WGS)

DNA was extracted from tissue samples using the NEB Monarch HMW DNA extraction kit for Tissues (#3060L). No more than 25mg of tissue was cut into small chunks, then placed into a 1.5mL tube and mashed with the pestle. The sample was lysed according to the protocol at a shaking speed of 1000rpm (liver at 2000rpm; colon at 2000rpm for 15min, no shaking 30min). After lysis and RNAse treatment, brain samples were put on ice for 3 min before protein precipitation and phase separation. 800uL of the upper phase is combined with isopropanol to precipitate the DNA. DNA washes and elution followed the manufacturer's protocol. DNA was quantified by Qubit and size distribution was measured on the Femto Pulse. In order to selectively remove smaller fragments of DNA, we performed a short read eliminator (PacBio 102-208-400) and cleaned up the DNA with a bead wash. Libraries were constructed using between 2-9ug of DNA and the Ligation Sequencing Kit from ONT (SQK-LSK114) with modifications to the manufacturer's protocol. End repair was incubated for 20 minutes, and the adapter ligation was incubated for 1 hour. The final library was eluted in 30ul of EB and quantified by Qubit. 200-400ng of library was loaded onto a FLO-PRO114M R10.4.1 flow cell for sequencing on the PromethION, with two nuclease washes and reloads after 24 and 48 hours of sequencing.

Long-Read DNA (ONT Ultra-long WGS)

Ultra-high molecular weight DNA was extracted from fibroblast cell lines using a phenol chloroform extraction protocol (Logsdon, protocols.io, 2020) or using the NEB Monarch HMW DNA extraction kit for Cells & Blooc (#T3050L) following the manufacturer's protocol with the following exceptions: 6 million cells was used for the starting input with a shaking speed of 600 rpm during the lysis step. DNA was precipitated with 300uL EEB (ONT) and solubilized at 4°C for two days. Libraries were constructed using the Ultra-Long DNA Sequencing Kit V14 (SQK-ULK114) following the manufacturer's protocol. For monarch DNA, two extractions were combined for library preparation and final elution volume was doubled from the protocol. For phenol chloroform extracted DNA, approximately 40ug of DNA was input into library prep and the final elution volume ranged from 2x-4x the protocol volume, depending on DNA visualized during the clean-up step. Final libraries were left at room temp over night to solubilize. 75 uL of library was loaded onto a primed FLO-PRO114M R10.4.1 flow cell for sequencing on the PromethION, with two nuclease washes and reloads after 24 and 48 hours of sequencing.

Hi-C DNA Sequencing

Cultured fibroblasts were harvested and counted. 1-2M cells were washed and frozen dry at -80°C. Omni-C library preparation was conducted using either the Dovetail Omni-C Kit (PN 21005G Dovetail® Omni-C® Kit) for human cells v2.0 or Dovetail Omni-C Kit for human cells v2.1 (updated July 2025) with minor modifications. Cell pellets were treated with 0.5-2uL of the Omni-C nuclease mix.

Sample digests were reviewed and compared against the recommended range in the Omni-C protocol prior to proceeding with proximity ligation. 150-200ng of proximity-ligated sample was used for library preparation. The number of PCR cycles used for library preparation was dependent upon the amount of starting material from proximity ligation and ranged between 9-12 cycles. The final library was sequenced on an Illumina NovaSeqX platform and generated at least 1.2 billion 2×150 read pairs.

Data used to create DSA

To generate DSA for COLO829BL, we used a combination of PacBio HiFi sequencing data, ONT ultra-long WGS data, and Hi-C. To generate DSA for the fibroblast (LB-LA2) and benchmarking donor samples (ST001–ST004), we leveraged PacBio HiFi-only assemblies. Where short-read Illumina sequencing data were available for a given sample, these data were used to calculate the quality value (QV) of that sample's genome assembly.

Sequencing at BCM, contributed by Chris Grochowski, Hsu Chao, Muchun Niu, Ziad Khan, Heer Mehta, Donna Muzny, Rui Chen, Chenghang Zong, and Harsha Doddapaneni

DNA Extraction

DNA from different sample sources (Cell lines, Tissue homogenates, Blood and Tissue Cores) was extracted using the Chemagic Prime 8 robot and Chemagen's proprietary Magnetic Bead technology, with the Chemagic Prime DNA Blood kit and the manufacturer's 2k protocol (Revvity).

RNA Extraction

RNA from cell lines material was extracted using the Chemagic Prime 8 robot and Chemagen's proprietary Magnetic Bead technology, with the Chemagic Prime Total RNA Blood 2.5k Kit (Revvity). RNA from other sample sources (Tissue homogenates, Blood and Tissue Cores) was extracted according to the manufacturer's Trizol protocol manually. The isopropanol/aqueous phase was then transferred to a RNeasy spin column (Qiagen), and the manufacturer's RNeasy protocol was then followed.

Short read WGS sequencing

Two independent methods (Picogreen assay and 1% E-gels)) were used to determine the quantity and quality of the DNA before library construction. For library preparation, DNA (1 ug) was sheared into fragments of approximately 450-600 bp in a Covaris E220 system (Covaris, Inc. Woburn, MA) in batches of 96 samples at a time, followed by double SPRI bead clean up to select a narrow band of sheared DNA for library preparation. DNA was end-repaired, 3'-adenylated and ligated using a set of 96 8-bp adapters (Illumina TruSeq UD Indexes v2, # 20040870) for sample barcoding. The final library size estimation and quantification are completed using the Fragment Analyzer (Agilent AAT, Inc) electrophoresis system and QuantStudio™ 6 Flex Real-Time PCR System (Applied Biosystems) respectively to achieve an average final library size of ∼530bp and must be greater than 470 bp. Libraries were sequenced on The NovaSeqX instrument to generate 150 bp, dual indexed and paired-end sequence reads in a format of multiplexed pools to generate 412× - 541× coverage.

Mapping and QC pipeline for internal quality assessment before data submission to DAC

WGS sequence data were aligned to the hg38 reference genome, followed by variant calling using Illumina's Dynamic Read Analysis for GENomics (DRAGEN) software, v4.3.6. Genome coverage was evaluated by calculating the mean coverage and the distribution of coverage across the genome, including the proportions of bases covered at ≥1×, ≥10×, and ≥20×. The proportions of bases meeting these thresholds and the total number of mapped bases at Q20 or higher were reported for internal QC tracking. Alignment data were assessed for contamination using VerifyBamID v1.1.3 and an orthogonal confirmation of sample identity was applied using the Error Rate In Sequencing (ERIS) software developed at the HGSC to rapidly compare sequence data to genotypes from SNP arrays via an ''exact match'' test.

Bulk RNA Sequencing

Whole transcriptome sequencing (total RNAseq) data was generated using the Illumina TruSeq Stranded Total RNA with Ribo-Zero Globin kit (20020612, Illumina Inc.) or Watchmaker RNA Library Prep Kit with Polaris® Depletion (7BK0002-096, Watchmaker Genomics). RNA quality and quantity was estimated using Agilent Bioanalyzer. To monitor sample and process consistency, 1 μl of the 1:50 diluted synthetic RNA designed by External RNA Controls Consortium (ERCC) (4456740, ThermoFisher) was added to 1 μg total RNA. In addition, as a process control, the Universal Human Reference RNA (UHR) (740000, Agilent Inc.), was processed in parallel with the RNA samples. Libraries were sequenced on the NovaSeq 6000 instrument using the S4 reagent kit (300 cycles) to generate 2x150bp paired-end reads. In order to generate a minimum of 157-300M read-pairs per sample.

Mapping and QC pipeline for internal quality assessment before data submission to DAC

The RNA-Seq analysis pipeline cleans and processes raw RNA sequencing data (FASTQs), providing robust QC metrics and has the flexibility to map the reads to either GRCh37 reference or GRCh38 (after excluding the alternate contigs). The pipeline aligns RNA-Seq reads, removes duplicates, and generates QC metrics. It also quantifies gene expression using RSEM, generates quality control metrics, and produces raw gene feature counts.

Long read WGS

WGS on ONT platform: Genomic DNA was quantified using Qubit dsDNA quantification broad range assay (Thermo Fisher Scientific). DNA size was determined using the Femto Pulse System (Agilent). A total of 3 libraries was prepared per sample to be sequenced on 3 flow cells with the aim of achieving 90× coverage. DNA was sheared using Covaris g-tubes (Covaris 520079) to achieve an average size of 20-25 kb for DNA from cell lines and 15-20 kb for DNA from tissues. Sheared DNA was size selected on the PippinHT instrument (Sage Science) using the 6-10 kb or the 15-20 kb High-Pass definition. ONT libraries were prepared using the SQK-LSK114 kit

following the manufacturer's instructions. Final libraries were eluted in 27 μL of the ONT elution buffer and quantified using the Qubit dsDNA quantification broad range assay. Libraries were loaded at 15 fmoles to be sequenced on R10.4.1 flowcells.

Mapping and QC pipeline

For each ONT sequencing event, the unaligned bams were consolidated into a single fastq, followed by alignment to hg38 reference genome using minimap2 v2.24. Additional sequencing is performed if the minimum mean genome coverage is not met. For sequencing data quality control, alignment data were assessed for sex concordance, contamination using VerifyBamID v1.1.3, and an orthogonal confirmation of sample identity using the HGSC's ERIS software. For additional downstream secondary analyses, we used Sniffles2 v2.0.5 for SV calling, Clair3 v3.0.1.11 for SNV calling, and WhatsHap v2.3 for phasing. These tools were conveniently executed via Princess2, a highly configurable open workframe. Post analysis, select files were archived. Our pipeline was implemented on our local high-performance compute cluster. All ONT data contained MM and ML methylation tags, which described the location of the base modifications and the likelihood of the modification (i.e. 5hmC, 5mC, or no modification) at each given location. Deliverables included ONT consolidated unaligned bams and fastqs.

WGS on PacBio platform

Genomic DNA was quantified using Qubit dsDNA quantification broad range assay (Thermo Fisher Scientific). DNA size was determined using the Femto Pulse System (Agilent). A total of 2 libraries were prepared for each sample to be sequenced across 3 SMRT Cells to achieve 90× coverage. DNA was sheared using Covaris g-tubes (Covaris 520079) to achieve an average size of 18-22 kb.Sheared DNA was size-selected on the PippinHT instrument (Sage Science) using the 6-10 kb or the 15-20 kb High-Pass definition. The size selected DNA was used as input for Pacbio SMRTBell Prep Kit 3.0 for library preparation. DNA damage repair, A-tailing and adapter ligation were performed as per manufacturer's instructions. SMRTBell Adapter Index Plate 96A was used for barcoding each library. Adapter ligated DNA was nuclease treated following the manufacturer's guidelines and purified using 1X Pacbio SMRTBell Cleanup beads. Final libraries were eluted in 30 μL Pacbio elution buffer and quantified using the Qubit dsDNA quantification high-sensitivity assay (Thermo Fisher Scientific). Final library size was determined using the Agilent Femto Pulse.Sequencing primer annealing and polymerase binding were performed using the Revio Binding Kit 3.0. Libraries were loaded onto the PacBio Revio machine utilizing SMRTlink v13 for workflow setup and loaded at 325pM loading concentration with 30 hours of movie time.

Mapping and QC pipeline for internal quality assessment before data submission to DAC

For PacBio, the PacBio unaligned HiFi bams were demultiplexed and aligned to hg38 reference genome using PacBio's SMRTLink v13 tools lima v2.9.0 and pbmm2 v1.13.1, respectively. Additional sequencing is performed if the minimum mean genome coverage is not met. For sequencing data quality control, alignment data were assessed for sex concordance, contamination using VerifyBamID v1.1.3, and an orthogonal confirmation of sample identity using the HGSC's ERIS software. For additional downstream secondary analyses, we used Sniffles2 v2.0.5 for SV calling, Clair3 v3.0.1.11 for SNV calling, and WhatsHap v2.3 for phasing. These tools were conveniently executed via Princess2, a highly configurable open workframe. Post analysis, select files were archived. Our pipeline was implemented on our local high-performance compute cluster. All PacBio data contained MM and ML methylation tags, which described the location of the base modifications and the likelihood of the modification (i.e. 5hmC, 5mC, or no modification) at each given location. PacBio unaligned HiFi bams were delivered.

Key QC metrics and thresholds have been established on both platforms (mean coverage, % genome coverage at 10×, HIFI yields, mean read length, contamination, and Q30 or Q10 Mapped bases) and sequence data is screened for this before submission to DAC.

Single cell RNA sequencing

Frozen samples were minced on dry ice, lysed in a nuclei extraction buffer, and dissociated using gentleMACS. Nuclei were filtered, centrifuged, resuspended, and triturated. After magnetic labeling with anti-nucleus beads, nuclei were purified using column separation, washed, and resuspended for counting, morphology assessment, and downstream library preparations. The library preparation and sequencing of single-nuclei cDNA were carried out following the manufacturer's protocols (https://www.10xgenomics.com). To obtain single cell GEMS (Gel Beads-In-Emulsions) for the reaction, single-nuclei suspension was loaded onto a Chromium X. The library for single nuclei RNA-seq was prepared with the Chromium GEM-X Single Cell 5' Reagent Kits v2 (10x Genomics), while the library of single nuclei ATAC-seq was prepared with the Chromium Next GEM Single Cell ATAC Reagent Kits v2 (10x Genomics). The constructed libraries were subsequently sequenced on an Illumina Novaseq 6000 (https://www.illumina.com). The sequencing data were processed using the 10X Genomics software (Cell Ranger v7.2.0 for gene expression (GEX) and Cell Ranger ATAC v2.1.0 for ATAC-seq analysis).

Droplet Digital PCR (ddPCR) Validation

A competitive TaqMan probe approach was used to validate the sample concentrations for the COLO829BLT50 as well as the HapMap sample mixture using the method we recently described (Table S13).[91] To assess the level of COLO829BL versus COLO829 we confirmed the presence and fraction of multiple well-known somatic variants from the COSMIC database within the mixture (Table S13).

Five BioRad designed ddPCR Mutation Assays targeting variants identified in the COLO829 cell line were ordered from Bio-Rad Laboratories (Hercules, CA). The assays targeted the following variants: MAP4K1 p.P298P (c.894C>T; Assay ID dHsaMDS546407091), RELN p.E331K (c.991G>A; Assay ID dHsaMDS550923980), MADD p.S1620F (c.4859C>T; Assay ID dHsaMDS851278352), SCN11A p.T340T (c.1020G>A; Assay ID dHsaMDS432990804), and ZNF217 p.P651S (c.1951C>T; Assay ID dHsaMDS581392264).

Each assay was run using standard ddPCR protocols and run on a Bio-Rad QX200 Droplet Digital PCR System.[91] Droplets were generated using the QX200 Droplet Generator, thermal cycled under recommended conditions, and read on the QX200 Droplet Reader. Data were subsequently analyzed using QuantaSoft software.

Sequencing at NYGC, contributed by Alexi Runnels

DNA/RNA Dual Extraction

DNA and RNA for all benchmarking samples were extracted using Qiagen's AllPrep kit (80204), using manufacturer instructions. For RNA samples with low RIN output from AllPrep extractions, additional RNA was extracted using Qiagen's RNeasy Mini kit (74106).

HMW DNA Extraction

HMW DNA for long read WGS was extracted using New England Biolabs' Monarch HMW Extraction kit (T3050L and T3060L).

Short read WGS Library Prep

Whole genome sequencing (WGS) libraries were prepared using the NEBNext Ultra II FS DNA PCR-free Library Preparation Kit (NEB E7430L) in accordance with the manufacturer's instructions. 500ng of DNA was sheared enzymatically and was subsequently end-repaired and adenylated. DNA fragments were ligated to Illumina sequencing adapters and the libraries underwent bead-based size selection. Final libraries were quantified using the QuantStudio5 Real-Time PCR System (Applied Biosystems) and Fragment Analyzer (Agilent).

*Bulk RNA Library Prep***.** mRNA libraries were prepped using Illumina TruSeq Stranded mRNA Library prep (Illumina 20020595), in accordance with manufacturer recommendations, and using IDT for Illumina custom 10nt Indices (IDT). Briefly, 500ng of total RNA was used for purification and fragmentation of mRNA. Purified mRNA underwent first and second strand cDNA synthesis. cDNA was then adenylated, ligated to Illumina sequencing adapters, and amplified by PCR (using 10 cycles). The cDNA libraries were quantified using Fragment Analyzer (Agilent) and Spectramax M2 (Molecular Devices). Libraries were sequenced on an Illumina NovaSeq sequencer, using 2 x 100 bp cycles. FASTQ files were provided to the DAC, for further analysis.

Long-Read WGS Library Prep. High molecular weight (HMW) DNA samples were first sheared using the Megaruptor 3 (Hologic Diagenode, catalog number B06010003) to a target fragment size of 45 kB, following the manufacturer's recommendations. Fragmented DNA quality was assessed using NanoDrop (ND-2000), Qubit Broad Range Assay (Q32850), and the Genomic TapeStation (G2964AA), all according to manufacturer instructions.

To improve 260/230 ratios, a 3X buffer exchange clean-up was performed using AMPure XP Beads for DNA Cleanup (A63882). The cleaned, fragmented HMW DNA was then prepared using the Oxford Nanopore Ligation Sequencing Kit V14 (SQK-LSK114), following the manufacturer's protocol. Final libraries were quantified again using the Qubit Broad Range Assay (Q32850) and Genomic TapeStation (G2964AA).

Illumina Sequencing. Short read WGS libraries were sequenced on the Illumina NovaSeq X Plus, using 2 x 150 bp cycles. Bulk mRNA libraries were sequenced on the Illumina NovaSeq X Plus, using 2 x 100 bp cycles. FASTQ files were provided to the DAC, for further analysis.

Oxford Nanopore Sequencing. HMW libraries were sequenced on Promethion P24 (PRO-SEQ024) using R10.4.1 flow cells (FLO-PRO114M) according to manufacture instructions. FASTQ and unaligned BAM files were provided to the DAC, for further analysis.

Sequencing at Broad, contributed by Shadi Zaheri

Long-Read DNA (PacBio HiFi)

High-fidelity long-read WGS was performed using the PacBio Revio platform. DNA was extracted from tissues and cell lines using Qiagen's MagAttract HMW DNA Kit. Libraries were prepared with the SMRTbell prep kit 3.0, size selected via Sage Science's Pippin HT to ~15kb and sequenced on the PacBio Revio. The resulting raw data files were processed through circular consensus error correction. Sequencing was performed to a target depth of 24× (single-cell derived iPSC lines) or 96× (COLO829BLT50, tissues, HapMap) per sample, with average HiFi read lengths of 14.8 kb. The data were aligned to GRCh38 (no ALT contigs) using pbmm2 (v1.13.0).

Short-Read DNA (Illumina WGS)

Short-read whole genome sequencing was carried out using the Illumina NovaSeq X plus 10B platform. DNA was extracted from tissues and cell lines using the Qiagen AllPrep DNA/RNA/miRNA Universal Kit. Libraries were prepared with Kapa Biosciences HyperPrep library construction kit. Each sample was sequenced to an average depth of 80× across two libraries, achieving a combined coverage of 160× using Illumina 2×150 bp paired-end read chemistry. Reads were aligned to GRCh38 using DRAGEN aligner through the Sentinel pipeline.Short-Read RNA-seq (Illumina).

Short-Read RNA

RNA-seq was performed on total RNA extracted from tissues and cell lines using the Qiagen AllPrep DNA/RNA/miRNA Universal Kit. RNA integrity was assessed using the Caliper LabChip GX. Libraries were prepared using either Illumina's TruSeq Stranded mRNA HT Kit or the Watchmaker RNA Library Prep Kit. TruSeq libraries were sequenced on the Illumina NovaSeq SP platform using 2×151 bp paired-end reads to a target depth of 75 million read pairs per sample. Watchmaker libraries were sequenced on the Illumina NovaSeq X Plus 25B platform using 2×146 bp paired-end reads to a target depth of 100 million read pairs per sample. Alignment was performed using the DRAGEN aligner (v07.031.677) against the GRCh38 reference genome.

Sequencing at Yale, contributed by Livia Tomasini

iPSC lines from LB donor have been previously sequenced by BGI to a coverage of 30-60× and the data has been made accessible at NDA (study no. 1057). DNA for the fibroblast sample was extracted using the DNeasy Blood & Tissue kit (Qiagen) following manufacturer's instructions for cultured cells. Whole genome sequencing (WGS) libraries were prepared using the Elevate Mechanical Library Prep Kit (Element Biosciences) PCR-free. Genomic DNA was sheared using a Covaris LE220 sonicator. A double-sided SPRI (Solid Phase Reversible Immobilization) selection method was used to target a fragment size of 600 bp. DNA fragments were subsequently end-repaired, adenylated, and ligated using adaptors from the Elevate Long UDI Adapter Kit (Element Biosciences). Libraries were evaluated using qPCR. Four libraries were sequenced on an Element AVITI sequencing system using 2x150bp read length to an average depth of 220× genome coverage.

Generating MEI targeted sequencing data, contributed by Eunjung Alice Lee and Weichen Zhou

MEI benchmarking is accompanied by a dedicated manuscript.[21] MEI-targeted sequencing methods, eHAT-seq on short reads and TEnCATS on long reads, were also benchmarked in HapMap mixture experiments. Somatic calls identified from WGS data were further orthogonally validated using these targeted sequencing approaches.

For eHAT-seq, libraries were prepared from 500 ng genomic DNA using an optimized protocol based on the original HAT-seq method,[82] with major improvements to both experimental procedures and the analytical pipelines. Briefly, DNA was fragmented, end-repaired, A-tailed, and adapter-ligated, followed by L1 enrichment PCR to amplify the L1-genome junctions. To improve sequence diversity of the semi-amplicon library and sample labeling, L1HS-targeting primers with variable staggers and unique dual indexes were introduced during L1 enrichment and indexing PCR, respectively. The resulting stranded eHAT-seq libraries were size-selected, quantified, pooled, and sequenced on an Illumina NovaSeq platform (2 × 150 bp). Reads were trimmed, and read-pairs containing L1Hs sequences in Read 2 were retained. Corresponding Read 1 (representing 3' genomic flanks) were aligned to hg38/GRCh38, and candidate insertion sites were identified by peak calling and annotation against reference MEI databases and genomic features.

The TEnCATS framework was previously described in McDonald et al.[79] and further optimized here. Libraries were generated from HapMap mixture genomic DNA using an optimized protocol for the SQK-LSK114 kit and ONT R10.4.1 flow cells for targeting both L1Hs and Alu elements. Briefly, genomic DNA was dephosphorylated and subjected to Cas9-mediated target enrichment using ribonucleoprotein complexes assembled from in vitro transcribed guide RNAs, followed by cleavage, A-tailing, adapter ligation, bead purification, and long-fragment cleanup. Libraries were sequenced for 72h on MinION and PromethION2 Solo instruments and a wash and reload was completed at 24 hours for PromethION libraries. Reads were processed with NanoPal, aligned to the reference genome, and analyzed to identify non-reference MEI insertions. On-target reads for L1Hs and Alu elements were defined by BLASTn, clustered by locus, and filtered to exclude looping and chimeric artifacts by Minimera, with only calls supported by at least two reads retained. Somatic calls identified from WGS data were further orthogonally validated using these targeted sequencing approaches.

Generating duplex sequencing data, contributed by Shadi Zaheri, Muchun Niu, Chenghang Zong, Harsha Doddapaneni, Nisrine T. Jabara, Gilad Evrony, Sangita Choudhury, and Alexi Runnels

Duplex sequencing experimental methods. Four donor tissue samples (ST001, ST002, ST003, and ST004), a COLO829 cell line, and a HapMap DNA mixture underwent duplex sequencing using the following six methods: CODEC, NanoSeq (HpyCH4V and MNB), CompDuplex-seq, HiDEF-seq, ppmSeq, and VISTA-seq. Benchmarking of duplex methods is accompanied by a dedicated manuscript.[22]

CODEC. Libraries were sequenced on an Illumina NovaSeq S4 platform, targeting approximately 1× duplex coverage. CODEC library preparation and data processing followed previously described methods.[84] Two distinct end repair and dA-tailing strategies were implemented to suppress error propagation during library preparation. An enhanced version of this protocol (DRv2, unpublished data) was also used to further minimize damage-related artifacts. Following ER/dA tailing, CODEC specific quadruplex adapters were ligated. Libraries were PCR-amplified and sequenced on the Illumina NovaSeq 6000 (S4 flow cells).

Compduplex and Nanoseq (HpyCH4V and MNB). NanoSeq (HpyCH4V and Mung Bean) libraries for the benchmark materials were prepared using 50-100ng of DNA as described by Abascal et al.[92] with minor changes as described in Chao et al. https://www.protocols.io/view/optimized-mung-bean-nuclease-nanoseq-library-prepa-d4h38t8n.html. Libraries were pooled and sequenced on Illumina NovaSeq sequencing platforms for 30× coverage. CompDuplex libraries were prepared using 20-50ngof gDNA as described by Niu and Zong (https://www.protocols.io/view/compduplex-accurate-detection-ofsomatic-mutations-kxygx3x4og8j/v1). Libraries were pooled and sequenced on Illumina NovaSeq sequencing platforms for 30× coverage.

HiDEF-seq. Genomic DNA was isolated from tissue and cultured cell samples using either the PacBio Nanobind PanDNA kit with Proteinase K and RNase A incubations at 37 °C or the Qiagen Puregene system with Proteinase K/RNase A incubations at room temperature. Libraries were prepared with an updated HiDEF-seq v3 workflow designed to streamline the procedure and provide genome-wide coverage via random fragmentation.[83] A detailed step-by-step protocol is available at: https://www.protocols.io/view/hidef-seq-kxygxy9mwl8j/v3. To reduce single-strand mismatch artifacts, we used a non-A-tailing strategy with blunt-end ligation to hairpin adapters. Short fragments were removed prior to fragmentation using the Short Read Eliminator (SRE) kit (PacBio), DNA was sheared to ~4 kb and then end-repaired. Nicks were ligated with E. Coli DNA ligase, residual nicks were blocked with dideoxy-C/G/T nucleotides, hairpin adapters were ligated, and residual non-circularized molecules were depleted by exonuclease treatment. Final libraries were run on a PacBio Revio instrument with 30-hour movie times and with per-strand output settings.

VISTA-seq. VISTA-seq adapts the META-CS framework for duplex consensus sequencing to accommodate inputs ranging from single cells to bulk DNA containing many alleles. A detailed step-by-step protocol is available at: https://www.protocols.io/view/tn5-duplex-sequencing-tn5-duplex-seq-for-low-input-6qpvr3nbzvmk/v1. After gentle lysis and cleanup, genomic DNA was tagmented with engineered Tn5 preloaded with strand-specific adapters, enabling simultaneous fragmentation and symmetric tagging of both strands. Strand-specific barcodes were then introduced through two rounds of indexed synthesis with high-fidelity polymerase, each followed by exonuclease cleanup to remove excess primers and reduce background. The resulting duplex-tagged libraries were PCR-amplified, purified, and size-selected to enrich ~400–600 bp fragments. Sequencing was performed as 150 bp paired-end reads on the Illumina NovaSeq X Plus platform with a PhiX spike-in for run quality calibration.

ppmSeq (Ultima) sequencing. ppmSeq (Paired Plus Minus Sequencing) libraries were prepared using the NEBNext UltraShear module (NEB M7634L) and NEBNext Ultra II DNA PCR-free Library preparation Kit (NEB E7410L) in accordance with the manufacturer's instructions. 250ng of DNA was sheared enzymatically and was subsequently end-repaired and adenylated. DNA fragments were ligated to Ultima Genomics Indexed Native Duplex Adapters and the libraries underwent bead-based size selection. Final libraries were quantified using the QuantStudio5 Real-Time PCR System (Applied Biosystems) and Fragment Analyzer (Agilent). The ppmSeq libraries were sequenced on Ultima Genomics Solaris (UG100) instrument according to manufacturer instructions.

Primary template-directed amplification (PTA), contributed by Joe Luquette, Abhiram Natu, and Milovan Suvakov

Benchmarking single cell analyses using PTA is accompanied by a dedicated manuscript.[23] Single nuclei were isolated from frozen post-mortem lung and colon tissues and amplified by primary template-directed amplification (PTA) across four TTDs from Yale University, Boston Children's Hospital (BCH), Weill Cornell College (Conell), and Mayo/Yonsei University. Sample input was frozen tissue cores at all sites except BCH, which used tissue homogenate.

The groups employed different nuclei preparation strategies, but final analysis of data that passed all QC steps revealed no significant differences between preparation protocols. Tissue was dissociated mechanically using Dounce homogenizers or the Singulator 100 automated platform. Nuclei extraction was done in sucrose-based buffers or using Minute detergent-free nuclei isolation kit (Invent Biotechnologies) with NP40 buffer. FACS-based single-nucleus sorting was performed at all sites, but DNA staining varied and included DAPI, PI, AO/PI, SYTO9, or no staining. Whole-genome amplification was performed using the BioSkryb ResolveDNA PTA kit at all centers: V1 was used at Yale, BCH, and Cornell, while Yonsei used both V1 and V2. Amplification quality was assessed by a 4-locus multiplex PCR assay at Yale, BCH, and Yonsei, with only samples yielding all four amplicons (and DNA concentrations exceeding 100 ng by Qubit) retained for sequencing. Cells that were not subject to 4-loci PCR had substantially lower rate of passing data QC post sequencing.

A total of 102 single nuclei (27 by Yale University, 20 by BCH, 7 Cornell, and 48 Yonsei University) were sequenced at Baylor College of Medicine, the Broad Institute, NYGC, WashU-VAI, and Theragen Bio. Sequencing library preparation varied across GCCs. All libraries were sequenced on Illumina NovaSeq X or NovaSeq X Plus platforms with 2x150 bp paired-end reads, targeting approximately 30× coverage, although final coverage varied per cell. Reads were uniformly processed at DAC as described in the following section.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data Processing at the Data Analysis Center (DAC), contributed by Michele Berselli and Hye-Jung Elizabeth Chun

Illumina paired-end whole genome sequencing (WGS) data were processed using a standard pipeline based on the Genome Analysis Toolkit (GATK) Best Practices.[93–95] Reads containing artifactual poly-G repeats associated with Illumina two-color chemistry were initially detected and discarded using fastp[96] (v0.23.2). The remaining reads were aligned to the GRCh38 reference genome (GCA_000001405.15), which excludes ALT contigs and decoy sequences from hs38d1, using BWA-MEM[60] (v0.7.17) as implemented in the Sentieon software[97] (v202308.01). This specific version of GRCh38 was chosen, because inclusion of ALT contigs, large variations with very long flanking sequences nearly identical to the primary human assembly, was deemed to reduce the sensitivity of variant calling and many other analyses due to most aligners giving MQ of 0 to reads mapped in the flanking sequences. Aligned reads were sorted by genomic coordinates, and read group information, including sample, flowcell, lane, and library, was assigned to each read. Duplicate reads were marked using Sentieon *LocusCollector* and *Dedup*, equivalent to *MarkDuplicates* in Picard (v2.9.0; https://broadinstitute.github.io/picard). To improve mapping accuracy, local realignment around indels was performed using Sentieon *Realigner*, equivalent to *RealignerTargetCreator* and *IndelRealigner* in GATK (v3.8). Base quality score recalibration (BQSR) and base alignment quality (BAQ) recalibration were performed using Sentieon *QualCal*, corresponding to *BaseRecalibrator* and *ApplyBQSR* in GATK (v4.1). Known sites used for BQSR included variants from dbSNP[98] (v138) and gold-standard indels from Mills and 1000 Genomes Project,[99] as provided in the GATK resource bundle. Original base qualities were retained in the OQ tag, allowing recovery of the initial quality scores from the alignments to enable re-generation of the complete original FASTQs from the final BAM files. To reduce file size, we removed the indel-related base quality score tags BI and BD, as recommended by GATK Best Practices. While these tags may offer refined base quality estimates near indels, they are not currently used by GATK or other standard variant callers in common practice and take a lot of disk space to store.

Processing PacBio HiFi WGS data

Data were processed using a standardized pipeline as follows. Reads were aligned to the GRCh38 reference genome (GCA_000001405.15), which excludes ALT contigs and decoy sequences from hs38d1, using pbmm2 (v1.13.0; https://github.com/PacificBiosciences/pbmm2). By default, all tags from the original unaligned reads were annotated to the final alignments. To reduce

the file size while retaining relevant alignment information, the –strip flag was used to remove non-essential kinetic tags (dq, dt, ip, iq, mq, pa, pc, pd, pe, pg, pm, pq, pt, pv, pw, px, sf, sq, st). This preserves methylation-related tags MM and ML, and assay-specific tags such as nucleosomes position in Fiber-seq data. Aligned reads were then sorted by genomic coordinates, and read group information, including sample and library, was assigned to each read.

Processing Oxford Nanopore Technologies (ONT) WGS data

Data were processed using a standardized pipeline as follows. Reads were aligned to the GRCh38 reference genome (GCA_000001405.15), which excludes ALT contigs and decoy sequences from hs38d1, using minimap2[59] (v2.26). To ensure consistency with pbmm2, the following flags were applied: -Y to enable soft clipping for supplementary alignments, -L to produce long CIGAR strings (used in the CG tag), –eqx to represent matches and mismatches explicitly using = and X, and –secondary=no to suppress secondary alignments. Aligned reads were then sorted by genomic coordinates, and read group information, including sample and library, was assigned to each read. Methylation-related tags MM and ML were retained from the original unaligned reads and annotated to the final alignments using Methylink (v0.6.0; https://github.com/projectoriented/methylink).

Processing RNA-seq data

Data were processed using a pipeline based on the GTEx consortium analysis pipeline. Reads were aligned to the GRCh38 reference genome (GCA_000001405.15), which excludes ALT contigs and decoy sequences from hs38d1, using STAR (v2.7.10b) as implemented in the Sentieon software (v202308.01). The HLA region was also excluded from the reference genome to generate the STAR index files. Genome aligned reads were sorted by genomic coordinates, and read group information, including sample and library, was assigned to each read. Duplicate reads were marked using Sentieon LocusCollector and Dedup, equivalent to MarkDuplicates in Picard (v2.9.0). Transcriptome aligned reads were used for quantification.

To match the GTEx pipeline, quantification and quality control were performed using two independent software packages. RSEM (v1.3.3) was used to generate gene- and transcript-level quantification, while RNA-SeQC (v2.4.2) was used to generate gene- and exon-level quantification as well as quality control metrics. In GTEx, RNA-SeQC results are used for gene- and exon-level read counts and TPM values, and RSEM results are used for transcript quantification. Consistent with this approach, we release both sets of results, allowing users to select gene quantification from either RSEM or RNA-SeQC as needed. GENCODE v47 was used to generate the RSEM index files, and the collapsed gene model for RNA-SeQC.

Data quality assessment

Data quality was assessed using a multi-step pipeline at the Data Analysis Center (DAC) at the time of data processing and after alignment for both original raw (FASTQ) and aligned (BAM/CRAM) sequence data to ensure that the data generated by the Network is of acceptable quality for variant analyses The QC status is determined based on the results of select QC checks (Table S5), and the QC status of each BAM/CRAM file is indicated in the files of metadata (i.e., "manifest" files), which are visualized in and available to download from the SMaHT Data Portal (https://data.smaht.org). All datasets released from the Network is of either "PASSED" or "FLAGGED" QC status, and datasets with "FAILED" QC status are not released. For Illumina data generated on the NovaSeq platforms, raw reads were evaluated using FastQC (v0.12.0; http://www.bioinformatics.babraham.ac.uk/projects/fastqc) to assess base quality scores, GC content distribution, and the presence of overrepresented sequences including adapter or other known artifacts, such as poly-G runs. For PacBio and ONT data, unaligned reads were evaluated using NanoPlot[100] (v1.44.1) to calculate read length distributions and other sequencing summary statistics. To assess potential microbial contamination of both bacterial and viral origin, all datasets were also processed using Kraken2[101] (v2.1.3), applied to the raw reads. To optimize computational efficiency while achieving accurate bacterial detection, the Standard-16 database provided by the authors was used.

Standard quality metrics, including mapping rates and other alignment-based statistics, were calculated using Samtools[102] (v1.17), Picard (v3.0.0), and a custom in-house Python script (https://github.com/smaht-dac/qc-pipelines/blob/main/dockerfiles/bamstats/bamStats.py). For all sequencing data, alignment-level statistics were summarized using the Samtools *stats* and *flagstat* commands. Sequencing coverage metrics, including mean and per-base coverage, were calculated using mosdepth[103] (v0.3.9). Additional metrics were generated using Picard tools, including *CollectAlignmentSummaryMetrics*, *CollectBaseDistributionByCycle*, *CollectGcBiasMetrics*, and *MeanQualityByCycle*. For paired-end Illumina data, insert size distributions were calculated using *CollectInsertSizeMetrics*, and overall sequencing performance was evaluated using *CollectWgsMetrics*. These metrics include the mean and standard deviation of insert sizes based on read pair alignment, read pair orientation, the fraction of aligned bases excluded due to low mapping quality or low base quality, and the proportion of bases reaching predefined coverage thresholds. For RNA-seq data, additional metrics were generated using RNA-SeQC (v2.4.2), including the rate of rRNA sequences among mapped reads, 5'-3' coverage ratios, estimates of library complexity, the exonic–intronic ratio, the number of genes detected, the intergenic rate, and the percentage of chimeric reads. We also implemented an in-house software that predicts tissue types using the RandomForest classification approach, to check correct tissue labeling based on gene-expression profiles.

Sample identity and integrity were evaluated using Somalier[104] (v0.2.19), which calculates the degree of relatedness between samples based on alignments to detect potential sample swaps or mislabeling. Samples with relatedness scores of 90% or lower were flagged for further review at the Data Analysis Center and the sequencing center to determine whether a swap or mislabels had occurred. To assess contamination from other human samples, VerifyBamID2[105] (v2.0.1) was used with a panel of known polymorphic loci in GRCh38, based on the 1000 Genomes Project and provided by the authors (https://github.com/Griffan/VerifyBamID/tree/master/resource). To increase computational efficiency while achieving accurate detection for PacBio and ONT data, VerifyBamID2 was run from pileups instead of the alignments directly. Intermediate pileups were generated using Samtools *mpileup* with the

following options: -B to disable BAQ computation, -a to output all positions, -s to output base alignment quality, and -q for minimum alignment quality.

Datasets with low to moderate levels of human contamination (<3%) were considered still useful for detecting somatic variants and made available for download at the SMaHT Data Portal (https://data.smaht.org). Datasets identified as having sample swaps or mislabels were retracted or deleted. The Data Analysis Center keeps a record of retracted and replacement datasets in the metadata, and these are made available to view in the Data Portal.

For all output files released by the SMaHT Data Analysis Center, the file format and integrity were checked using MD5 checksum validation and file-format sanity checks developed in-house to confirm EOF and integrity of all generated alignment files. All bioinformatics pipelines at the DAC were executed in the Amazon Web Services (WGS) infrastructure, designed and implemented to support modular, distributed processing at scale.

SNV and indel calling, contributed by Yeongjun Jang, Stephanie Gardiner, Stephanie Georges, and Yoo-Jin Jiny Ha

Benchmarking of mutation calling using the COLO829 cell lines is accompanied by a dedicated manuscript.[15] To define a benchmarking set of somatic mutations unique to COLO829, we applied four variant callers—TNHaplotyper2 (Sentieon Genomics 202112.06), Strelka2 (v2.9.2), VarNet (v1.1.0), and RUFUS (v0.1)—using COLO829BL as the matched control. PacBio HiFi support was assessed using bcftools mpileup (v1.17). For SNVs, PacBio allele counts were parsed per candidate site, and calls were retained if supported by $\geq$2 tumor reads and absent in the matched blood sample. For indels, additional breakpoint-aware annotation was performed to quantify allele balance and distinguish the targeted indel from other alternative alleles with Illumina COLO829BL data. We also established a negative control set to avoid misclassifying true variants as false positives in the benchmark, ensuring a fair evaluation of variant caller performance and preventing penalization for detecting true mutations missed during reference set construction. The negative control set consists of homozygous reference sites in both COLO829T and COLO829BL samples. Germline variant sites were also provided as a separate negative control set.

Somatic mutation calling for the COLO829BLT50 mixture was performed using 12 calling pipelines (M1-M12) that were based on DRAGEN, Lancet2, GATK/Mutect2, TNHaplotyper2, RUFUS and MosaicForecast.

M1 (BCM)

DRAGEN v4.3 with mosaic detection enabled; BAM headers were edited to allow correct sample-level merging. Calls flagged PASS or MOSAIC with AF $\leq$ 0.05 were retained and multi-allelic sites were excluded.

M2 (Broad)

DRAGEN v4.3 run with a single read-group per BAM (–use-single-read-group-for-bam-list) and a mosaic AF threshold of 0.2 (–vc-mosaic-af-filter-threshold); PASS calls were retained using the same post-calling filter as M1.

M3 (Broad)

the same DRAGEN v4.3 pipeline ("DRAGEN-Direct") applied directly to the original CRAMs.

M4 (Broad)

GATK Mutect2 v4.1.6.0 candidate somatic calls were filtered against a panel of normals (PoN) — a catalogue of recurrent germline and technical-artifact sites compiled by running the caller across unrelated normal genomes from the 1000 Genomes Project — together with gnomAD population allele frequencies. Sites in regions of very low (<50×) or abnormally high (>1000×) coverage were discarded, and residual germline variants were removed with a one-sided binomial test against an expected germline VAF $\geq$ 0.5, followed by Benjamini–Hochberg correction at $q < 10^{-5}$.

M5 (NYGC)

Lancet2 v2.8.5 (single-sample mode). Germline variants were removed using HaplotypeCaller/GenotypeGVCFs, gnomAD v3.1.2 ($AF > 10^{-5}$) and the High-Coverage 1000 Genomes callset, and UCSC simpleRepeats and segmental duplications were excluded. Retained calls required read depth within 0.65–1.25× the sample-mean coverage and alt-allele depth $\geq$ 4, and passed Lancet2's internal quality filters — a strand-bias filter (removing variants whose supporting reads are disproportionately from one strand), a base/allele-quality filter (requiring sufficiently high base-call quality at the variant allele), and a mapping-quality filter (requiring supporting reads to be confidently/uniquely aligned) — with final calls restricted to VAF $\leq$ 10%.

M6 (WashU-VAI)

GATK Mutect2 v4.4.0.0 in tumor-only mode, filtered against a public PoN and gnomAD (FilterMutectCalls), retaining calls with VAF < 2%.

M7 (WashU-VAI)

as M6, but additionally removing dbSNP v151 sites, calls with depth < 50× or > 1000×, and variants outside the canonical chromosomes.

M8 (UW-SCRI)

GATK Mutect2 in tumor-only mode using the GATK PoN and gnomAD for germline filtering (FilterMutectCalls); PASS calls were retained with VAF < 40% and depth > 20×.

M9–M11 (Utah-TTD)

RUFUS is a reference-free, k-mer–based caller (minimum k-mer count 5, k-mer length 25 bp) run against a matched normal (280× COLO829BL). The three pipelines differ only in the stringency of a variant-quality-score (VQS) classifier trained on the SEQC2 HCC1395 benchmark: M9 (lenient) applies no VQS filtering; M10 (moderate) applies the VQS cutoff giving ~90% sensitivity to true SNVs and indels; M11 (strict) applies cutoffs giving ~80% sensitivity for SNVs and ~60% for indels.

M12 (DAC)

MosaicForecast. Before genotyping, calls in segmental duplications and in "clustered" regions — genomic intervals enriched for SNVs supported by ≥ 3 haplotypes, which flag copy-number/repeat artifacts rather than true mosaics (parklab SegDup_and_clustered reference) — were removed; MosaicForecast's phasing- and random-forest–based model was then used to retain high-confidence mosaic calls.

We evaluated performance across three genomic context strata. Easy regions were defined by the 1000 Genomes strict mask (high-confidence, short-read–callable loci). Difficult regions were sites within the PanMask strict (pm151b) accessibility set but outside the 1KG strict mask, and extreme regions were loci outside PanMask, which are typically highly repetitive or structurally complex; PanMask provides a pan-genome–derived, less reference-biased accessibility framework (details and resources can be found on GitHub https://github.com/parklab/SMaHT_Regional_Categorization).

To identify cell culture-induced mutations that may have arisen during continuous expansion and harvesting of COLO829BL lymphoblastic cells, we examined variants present in COLO829BLT50 but absent from the COLO829 reference set. RUFUS and Mutect2 were each run in paired control mode on the five COLO829BLT50 Illumina WGS datasets—generated across different sequencing centers—using COLO829BL as the matched normal. The RUFUS and Mutect2 call sets were merged using bcftools isec (v1.23), to generate five sequencing center-specific COLO829BLT50 call sets. The COLO829 reference set was subtracted from these sequencing center-specific sets, to retain the set of candidate cell line artifacts. We hypothesized that mutations identified across all five sequencing centers (set C(5,5)) were more likely to be true variants than those found in four (C(5,4)), three (C(5,3)), or fewer center-specific data sets, which are more likely to be sequencing or mutation calling tool artifacts.

We have used available, deep-coverage PacBio HiFi long-read data as correlative evidence to determine if any given candidate is truly a culture-induced mutation or an artifact. We called a candidate a true culture-induced mutation if the following criteria were fulfilled: (1) the presence of $\geq$2 supporting PacBio reads in COLO829BLT50 (confirming the mutation in the cell-line mix); (2) no PacBio or Illumina supporting reads in COLO829BL (confirming the absence of the mutation in the original blood cell line); and (3) no PacBio or Illumina supporting reads in COLO892 (confirming the absence of the mutation in the tumor cell line). Next, we checked the presence of the mutation in a deeply sequenced sample from an unrelated individual, donor ST001, and accepted the site if the mutant allele was present < 2 PacBio reads (eliminating recurrent sequencing error at the candidate mutation site). Finally, we removed any variants defined in the negative control set (described above). This procedure yielded a total of 5,372 mutations across C(5,5),..., C(5,1) sets; as expected C(5,5) contributing the largest fraction of sites.

We carried out a similar procedure for INDEL mutations. Here, the fraction of sites confirmed by the filtering procedure described above was so low in C(5,3), C(5,2), and C(5,1) that we only kept confirmed sites from C(5,5) and C(5,4). The procedure yielded a total of 196 INDEL sites.

Benchmarking of mutation calling using fibroblasts and iPSC lines is accompanied by a dedicated manuscript.[16] To identify benchmarking somatic SNVs and indels in fibroblasts we performed an all-to-all exhaustive comparison of genomes of 25 clonal iPSC lines. These initial calls were then processed using the All2 filtering strategy[45] and calls with VAF > 25% in each iPSC line were retained. We designated unique mutations discovered in five LA2 iPSC lines as the benchmarking set. Mutation from select iPSC lines that originated from different biopsies formed the negative control mutation set.

Somatic mosaic mutation calling in the fibroblast sample was conducted using RUFUS tool and Mutect2 (GATK version 4.2.6) in three different modes to ensure comprehensive discovery: (1) Tumor with Matched Normal, (2) Tumor-Only with default parameters, and (3) Tumor-Only with high sensitivity options "–initial-tumor-lod 0 –tumor-lod-to-emit 0 –af-of-alleles-not-in-resource 5e-8 –adaptive-pruning-initial-error-rate 0.0001 –pruning-lod-threshold -4". In the Tumor with Matched Normal mode, the five iPSC lines comprising the negative control set were used collectively as the matched "normal" samples against which the fibroblast was compared. For the Tumor-Only mode, we used a Panel of Normals (PON), which we downloaded from the GATK resource bundle. Calls were filtered using a slightly modified version of the BSMN filtering workflow[11,44] Called mutations were confirmed using data from different sequencing platforms, specifically PacBio and AVITI. A minimum quality cutoff of 20 was applied for both mapping quality and base quality. A mutation call was considered confirmed in fibroblast sample if at least one read supporting the alternative allele met these quality criteria. When analyzing post-mortem samples from the ST002 sample the same workflow as for the fibroblasts was utilized but for additional stringency, we applied additional filters to generate the final set of mutations: MFRL (median fragment length) >300, MMQ (median mapping quality) >56, MPOS (mean position in read) $\geq$ 20, MBQ (mean base quality) > 28, and ECNT (the number of somatic event) < 3.

Comparing SNV calling on CHM13 and GRCh38 references, contributed by Mrunal Dehankar, Yeongjun Jang, and Yifan Wang

To reprocess Illumina short-read data using CHM13, we applied the same alignment pipeline as used for aligning to GRCh38, with the indel resource lifted over from GRCh38 to CHM13. To identify benchmark somatic SNVs and indels in fibroblasts, we performed an exhaustive all-to-all comparison of genomes from 25 clonal iPSC lines. These initial calls were then processed using the All2 filtering strategy,[45] and calls with VAF >25% in each iPSC line were retained. We designated unique mutations identified in five LA2 iPSC lines as the benchmark set.

Somatic mosaic mutation calling in the fibroblast sample was performed using Mutect2 (GATK version 4.2.6) in both default tumor-only mode and high-sensitivity tumor-only mode with the following options: "–initial-tumor-lod 0 –tumor-lod-to-emit 0 –af-of-alleles-not-in-resource 5e-8 –adaptive-pruning-initial-error-rate 0.0001 –pruning-lod-threshold -4". For variant calling, we used a Mutect2 Panel of Normals (PON), originally generated from 1000 Genomes Project data and lifted over from GRCh38 to CHM13. Calls were filtered using a slightly modified version of the BSMN filtering workflow,[11,44] with a gnomAD population germline variant resource downloaded from the official gnomAD website that had also been generated by lifting variants over from GRCh38. Filtering of low VAF artifacts by the BSMN PON mask[11] was applied by lifting over mutation calls to GRCh38. For comparison purposes, we also lifted over the genomic stratification masks (i.e., easy, difficult, and extreme regions) from GRCh38 to CHM13. Comparisons of mutation calling using CHM13 and GRCh38 were restricted to equivalent easy regions.

Using CHM13 allowed making additional 1,538 bulk-only calls majority of which were with VAF < 1%. However, only approximately two-third of the additional calls were confirmed. It is unlikely that this lower confirmation rate on CHM13 as compared to GRCh38 reflects insufficient PacBio and AVITI coverage (~600× combined), as the majority of the additional calls had VAF >0.5% in the Illumina data and would therefore be expected to have, on average, three supporting reads in the confirmation data. We hypothesize that the lower confirmation rate results from lifting over GRCh38-based resources to CHM13. In particular, BSMN PON background frequencies mask[11], which capture recurrent sequencing and alignment errors, are likely reference-specific; therefore, they may not adequately remove recurrent artifacts when aligning and calling against CHM13. Similarly, lifting over other GRCh38-based resources to CHM13 may contribute to false calls and the lower confirmation rate.

Methods for MEI analyses, contributed by Seunghyun Wang and Weichen Zhou

Benchmarking of MEI calling is accompanied by a dedicated manuscript.[21] We first established a benchmarking set in the HapMap mixture experiment. Computationally, germline MEIs were defined per HapMap donor by combining assembly-based callsets (i.e., DipCall v0.20 and Minigraph v0.34 on haplotype-resolved assemblies) with read-based calls (i.e., Sniffles2 v2.2.2). Candidate insertions were then systematically annotated and filtered to focus on young, active subfamilies (L1HS/L1PA2, AluY/AluYb8/AluYa5, SVA_E/F). A custom TPRT-signature pipeline assigned mechanistic evidence by detecting TSDs (via synthetic pseudo-reads and bwa-mem CIGAR inspection) and polyA/T tails, enabling confidence tiering (Tier 1: TSD+polyA/T; Tier 2: polyA-only; Tier 3: remaining young-subfamily insertions). Zygosity was inferred from VAF distributions using a Gaussian mixture model. Somatic truth in the mixture was simulated by labeling HG005-present insertions as "germline" and insertions private to minor donors as "somatic," with expected VAF computed as mixture ratio × zygosity.

We then conducted HapMap mixture benchmarks for computational sMEI detection and evaluation, emphasizing read-signal-driven callers and harmonized scoring. Single-sample-capable WGS methods (short-read: xTea_mosaic v0.1.9, RetroSom v2, MELT v2.2.2; long-read: PALMER v2.0.1; plus tuned germline SV callers xTea_long v0.1.0, cuteSV v2.1.2, Sniffles2 v2.6.3 with reduced support thresholds) were run with repeat-based annotation (RepeatMasker) and young-subfamily filters to prioritize biologically plausible events. Performance was quantified against the tiered truth sets using breakpoint matching (≤100 bp for WGS; ≤200 bp for targeted) to define TP/FP, with recall/precision/F1 computed globally and stratified by mappability (151-mer unique vs non-unique), repeat context (ME vs ME-free), expected/caller VAF bins, and insertion length bins. Two computational refinements were layered on top: (i) cross-platform integration combining callset overlap and raw clipped-read evidence between short- and long-read data, with caller-specific FP filters (e.g., exclude non-unique regions for SR-only; require young-subfamily BLAST support and remove supplementary-only evidence for LR-only), and (ii) local haplotype phasing around each insertion (high-quality hetSNPs within ±40 kb; long reads spanning SNP+insertion) to separate likely germline from low-read, haplotype-linked mosaic events. Finally, the impact of DSA alignment on false positives in segmental duplications was tested using a binomial Generalized linear model comparing FP composition in SD vs outside SD while adjusting for caller and mixture effects.

For benchmarking donors (tissue homogenates), we called sMEIs per sample using xTea_mosaic on Illumina WGS and PALMER on long-read WGS, then applied additional computational filters (BLAST-based subfamily annotation with removal of older subfamilies and supplementary-alignment-only calls). Germline candidates were reduced by haplotype phasing, cross-sample consistency checks (sites classified as germline in any sample were treated as germline across samples), and exclusion of known polymorphic MEI loci curated from prior studies. We then applied the same short-/long-read integration framework across all platform pairs per donor, followed by structured manual review using TPRT evidence (TSD/polyA), breakpoint concordance, and repeat identity to finalize high-confidence sMEIs and estimate VAF from supporting reads normalized by local long-read depth.

Methods for SV analyses, contributed by Yuwei Zhang and Fritz Sedlazeck

Benchmarking of SV calling use HapMap mixture is accompanied by a dedicated manuscript.[20] Human Pangenome Reference Consortium assemblies for the six HapMap samples were procured from https://github.com/human-pangenomics/HPP_Year1_Assemblies. Each of the twelve haplotypes were independently aligned to GRCh38 and variants called with minimap and paftools, and post processed with BCFtools and Truvari to consolidate SV across haplotypes and collapse highly similar, distinct SVs. This produces the benchmark of 21,384 pseudo-somatic insertions and deletions. The benchmark was restricted to high-confidence genomic regions supported by all 12 haplotypes, excluding poorly assembled, reference-gap-proximal, and highly complex regions, ultimately covering 90.2% of GRCh38 autosomes.

Eight SV callers capable of operating without a matched normal and detecting variants below 10% VAF were selected: GRIDSS2 (v2.13.2), SvABA (v1.2.0), Manta (v1.6.0) and Delly (v1.2.6) for short-read data; Severus (v0.1.1), Sniffles (v2.4), and Delly (v1.2.6) for long-read data; Sawfish (v0.12.4), and SVDSS (v2.0.0-alpha.1) for HiFi data only. All callers were executed with default parameters on single-sample inputs without matched-normal controls. For long-read callers that support read-backed phasing or phased alignment inputs, such as Severus and Sniffles2, calling was performed directly on unphased single-sample BAM files without pre-haplotagging. This setup ensured a uniform baseline evaluation across callers, as phasing information serves for genotype annotation rather than primary variant discovery in Sniffles2, and internal benchmarks showed that read phasing had a negligible impact on Severus's detection performance. Uniform VCF normalization was conducted for each call set, and we only evaluated duplication, insertions and deletions ($\geq$50 bp in length) with PASS flag.

Truvari was used to benchmark each call set against the set of benchmark SVs within the defined high-confidence regions, matching SVs of the same type using positional, sequence-similarity, and size-similarity criteria. Benchmark SVs present in the highest-abundance HG005 sample were treated as germline, whereas the remaining variants were considered somatic; caller calls with reported VAF <25% were treated as somatic to focus the evaluation on the pseudo-somatic range. Performance was assessed using recall and precision, both genome-wide and across different VAF bins. To identify the sources of caller failure, SVs were additionally stratified by the following features: proximity to other SV; breakpoint overlapping a tandem repeat; low VAF; small event; and insertions. Calls with none of the features were considered as "not-challenging". To evaluate platform-specific detection limits, replicate BAMs were merged per platform and then randomly down-sampled to 60×, 90×, 120×, 150×, and 180×.

Methods for assembly, contributed by Min-Hwan Sohn, Youngjun Kwon, and Andrew B. Stergachis

UW-SCRI Genome Characterization Center constructed the COLO829BL diploid DSA[19] using COLO829BL-derived PacBio HiFi/Fiber-seq (60×) and the combination of ONT Ultra-long WGS and ONT Standard WGS (60×) coupled with Hi-C sequencing data (30×) with Verkko v2.1.[34] To maximize N50 of the assembly, ONT reads were downsampled by selecting the longest reads until a targeted 60× coverage was achieved. Specifically, based on the index file (.fai) of the ONT FASTQ, read IDs were extracted in descending order of read length until reaching 60× coverage based on 3.1 Gbp as the genome size. Reads corresponding to these IDs were then stored separately to generate the downsampled ONT FASTQ file. Phased assembly was generated using Verkko v2.1.[34] For the benchmarking donors (ST001, ST002, ST003 and ST004) and LB-LA2 Fibroblast DSA, HiFi-only diploid assemblies were generated using hifiasm[58] with default parameters, including the –dual-scaf and –telo-m CCCTAA options. PacBio HiFi reads from available tissue data for ST001 and ST002 were downsampled to 60× coverage by selecting the longest reads, consistent with the approach described above, while ST003 (49.76×) and ST004 (37.70×) were assembled without downsampling. SMHTLBLA2 was assembled using 110× coverage.

After the initial assembly, we implemented additional filtering steps to ensure that contaminated sequences are removed using NCBI FCS41 v0.4.0 beta and BLAST42 v2.15. In short, contigs shorter than 10 bp, which can cause errors in FCS, were excluded, followed by the detection of foreign DNA and adapter sequences. Subsequently, mitochondrial, Epstein–Barr virus (EBV) and ribosomal DNA (rDNA) sequences were identified using a BLAST-based approach. The detected contaminant and adapter regions were trimmed, and rDNA contigs were separated from the assembly. The command used to detect foreign DNA is as follows: python3 fcs.py –image FCS_IMG screen genome –fasta HAPLOID_ASSEMBLY.fasta –out-dir OUTDIR –gx-db GXDB_LOC –tax-id 9606. The command used to detect adapter sequence is as follows: av_screen_x -o OUTDIR –euk HAPLOID_ASSEMBLY.fasta. The command used to detect mitochondrial, EBV, and rDNA sequences is as follows: blastn -query HAPLOID_ASSEMBLY.fasta -db TARGET_DB -outfmt 6 –evalue 1e-30 > OUTPUT. Finally, for each haplotype, a decontaminated assembly FASTA file was produced with foreign DNA (including EBV), adapter, mitochondrial sequences, and rDNA contigs removed.

Methods for duplex, contributed by Yang Zhang, Vinayak V. Viswanadham, and Michail Andreopoulos

Benchmarking of duplex methods is accompanied by a dedicated manuscript.[22] Duplex sequencing data were processed using platform-specific analysis pipelines optimized for each library architecture and benchmarked against low-mutation reference samples. Across all methods, reads were aligned to the hg38/GRCh38 reference genome without decoy sequences when applicable, duplex molecules were reconstructed according to platform-specific designs, and somatic variants were retained only after stringent filtering to suppress sequencing, amplification, and single-strand artefacts. Technology-specific thresholds and post-processing steps were calibrated separately for each platform rather than imposed uniformly across assays.

For ppmSeq, aligned CRAMs were processed using the manufacturer's preprocessing workflow together with the downstream SRSNV framework and an additional custom SNV-calling procedure that estimates sensitivity-corrected mutational spectra and mutation burden under a predefined false discovery rate. Analyses were restricted to the Ultima high-confidence genome regions.

For CODEC, data were processed from raw sequencing output to variant calls using the standard CODECsuite workflow implemented in a Terra/WDL-based cloud pipeline.

For CompDuplex-seq and NanoSeq, variant calling and filtering were performed using a unified CompDuplex-NanoSeq workflow that includes read preprocessing, alignment, consensus molecule generation, duplex validation, and stringent filtering based on sequence quality, mapping quality, and duplex support.

For HiDEF-seq, data were processed using the HiDEF-seq v3.1 workflow, which is compatible with PacBio Revio data and preserves the core read-processing and filtering logic of the previously published HiDEF-seq pipeline (https://github.com/evronylab/HiDEF-seq) while using a refactored Nextflow-based implementation.

For VISTA-seq, somatic SNVs were called using a customized workflow derived from pre-pe and lianti, with modifications for barcode extraction, pooled read handling, and additional filtering to reduce recurrent false positives in multi-allelic and pooled-input settings.

Methods for PTA analyses, contributed by Joe Luquette and Milovan Suvakov

Analyses of single cell via the PTA approach is accompanied by in a dedicated manuscript.[23] Small somatic mutation detection and analysis with SCAN2 was performed to call single nucleotide variants (SNVs), insertions and deletions <50 bp (indels), and dinucleotide variants (DNVs). SCAN2 is a somatic mutation caller specifically developed for analysis of PTA-amplified single cells. The pipeline utilized the GRCh38 reference genome, dbSNP v151 for common variant annotation, and Sentieon's implementation of GATK HaplotypeCaller. To enable sensitive somatic indel calling, a cross-sample panel was constructed from 100 kb bins spanning chromosomes 1-22, X and Y (excluding pseudo-autosomal regions and regions with anomalously high sequencing depth), using single cells and matched bulks from this study combined with previously published PTA single-cell and matched bulk data downloaded from dbGaP. Mutations were called separately for colon and lung tissues to allow single cells to be matched with bulks from the opposite tissue, thereby reducing the likelihood that SCAN2 would incorrectly discard clonal mutations due to read support in the matched bulk. To retain maximum read support during variant calling, all samples were analyzed simultaneously rather than employing a standard GVCF approach, which loses information by omitting low read support. Eagle v2.4.1 was used for phasing with the 1000 Genomes reference panel. Mutation rates were automatically estimated by SCAN2, and false discovery rates were estimated using previously established false positive rates for PTA-amplified genomes. Sequencing data quality control evaluated cells across five metrics generated by SCAN2: MAPD, allele balance (distribution of VAFs at heterozygous germline SNPs), read depth distribution, GC bias (assessed by fitting a loess curve to read counts in non-overlapping 1 kb bins as a function of GC content), and genomic read depth profiles. Cells exhibiting outlier status in two or more of these metrics were excluded from all small mutation analyses (SNVs, indels, and DNVs).

Single-cell lineage reconstruction was performed using Sequoia. To avoid discarding early somatic mutations that define cell phylogenies near the zygote, SCAN2's filters requiring no read evidence of a somatic mutation in the matched bulk sample were relaxed by modifying the maximum bulk alternate read count, allele frequency, and binomial probability parameters. Relaxed calls were extracted and a list of unique mutations, identified by chromosome, position, reference and mutated sequence, was compiled across cells. The numbers of reference- and mutation-supporting reads were retrieved for each cell and site to create site-by-cell matrices of reference and mutation read counts. These matrices formed the input for Sequoia, which further filtered the relaxed mutation calls and inferred the phylogenetic tree. To account for artificial overdispersion introduced by research site-specific variation, beta-binomial overdispersion-based filtering was expanded to calculate the maximum likelihood overdispersion parameter (rho) across all cells and within cells from each center, requiring rho > 0.3 for shared mutation calls. For comparison, somatic mutations independently identified in high-depth, multi-tissue bulk whole-genome sequencing were interrogated in the single cells by generating per-cell pileups, producing variant-by-cell read count matrices that were similarly processed by Sequoia. Topologies of the bulk- and single-cell-derived phylogenetic trees were compared using a normalized generalized Robinson-Foulds distance computed via the JaccardRobinsonFoulds metric from the R package TreeDist.

Mutational signature extraction was conducted using the HDP (hierarchical Dirichlet process) R package. Counts spanning 257 distinct categories were analyzed: 96 SNVs with trinucleotide contexts, 83 indels classified by length and repeat/microhomology characteristics, and 78 distinct types of DNVs. To prevent double counting of mutations, variants were assigned to individual phylogenetic tree branches as dependent nodes, with sequencing centers constituting the parent nodes in the hierarchical structure.

Genome-wide copy number alteration (CNA) detection was conducted using CNVpytor and HiScanner. CNVpytor utilized a two-dimensional approach integrating both read depth coverage and B-allele frequency (BAF) signals during segmentation, with germline SNPs phased using shapeit4. BAF was calculated at heterozygous SNPs within the 1000 Genomes Project strict mask in 100 kb bins. High-confidence events were identified by filtering for integer copy number changes expected in a single cell: deletions (CN < 1.5, |delta_BAF| > 0.4), duplications (CN > 2.5, |delta_BAF| 0.08 to 0.24), and copy-neutral loss of heterozygosity (CN-LOH; 1.6 < CN < 2.4, |delta_BAF| > 0.4), with all events required to span at least 10 bins of heterozygous SNPs and a minimum size of 10 Mb. HiScanner integrated read depth with BAF and phasing information from Eagle-phased germline SNPs created by SCAN2. Autosomal diploid segments were examined for CN-LOH by segmenting into approximately 5 Mb intervals and identifying those with mean BAF < 0.25. To smooth out small CNAs potentially representing noise, segments spanning 10 Mb or less had their copy number set to that of the nearest larger flanking segment, and adjacent segments with equal copy number were merged. HiScanner was also executed in read depth-only mode to quantify sex chromosome aneuploidies. Aneuploidies on the Y chromosome were further analyzed using majority-voting phasing of pseudoautosomal region (PAR1) SNPs from cells with chromosome Y loss, which allowed high-confidence definition of the Y-chromosome haplotype.

Somatic structural variants (SVs) were detected by running Manta with default parameters, comparing each single-cell sample against the matched bulk sample. Manta's deletion and duplication calls larger than 100 kb were treated as candidates and

subsequently validated using CNVpytor by requiring a corresponding integer-level copy number change and significant BAF shift. Of the confident calls, the majority were located in T-cell receptor loci, consistent with V(D)J recombination. For complex rearrangements, Manta's break-end (BND) calls were used to provide orthogonal, breakpoint-level support for large sub-chromosomal CNAs identified by CNVpytor.

Comparing benchmarking mutations set for SVs and MEIs, contributed by Nahyun Kong and Zitian Tang

To evaluate the agreement of two complementary somatic benchmarking strategies in HapMap cell line mixtures, we compared SV and MEI benchmark sets generated using pangenome-based and hybrid assembly-based approaches. This comparison was designed to determine the extent to which the two methods produce concordant results, identify variant sets unique to each approach, and clarify the factors underlying these differences. We focused specifically on SVs and MEIs because these variant classes are often underrepresented in published somatic benchmarking resources.

Prior to comparison, shared benchmark regions were defined by intersecting the pangenome-based[17] and hybrid assembly-based[20,21] benchmark regions using bedtools (v2.27.1) intersect, and all subsequent variant comparisons were restricted to these shared regions. Since the pangenome-based approach targets pseudo-somatic variants exclusively, with HG005 germline variants filtered out, non-HG005 SVs and non-HG005 MEIs were selected from the assembly-based benchmark set or MEI working group set before comparison. For SVs, the SV working group easy set and the pangenome-based SV benchmark set were evaluated using Truvari v4.0.0 with the -pick multi flag. For MEIs, concordance between the MEI working group's tier1 benchmark set and the pangenome-based MEI benchmark set was assessed by comparing insertion start positions within the shared benchmark regions using in-house scripts.

When comparing SV and MEI benchmark sets generated using a pangenome-based approach with those from hybrid, assembly-based approaches, we found substantial overlap. The primary differences laid in variant representation: the pangenome-based method preserved exact genotypes and haplotype structures, separating overlapping variants into distinct records, whereas the assembly-based approaches merged comparable variants to facilitate benchmarking. Each approach therefore offers distinct advantages, depending on the goals of benchmarking and downstream analyses. For MEI, counts were also different depending on whether they were classified based on repeat annotation (higher count) or mechanistic signature of target primed reverse transcription (lower counts).

Evaluating pangenome-based somatic variant detection approach, contributed by Qichen Fu, Benpeng Miao, and Juan F. Macias-Velasco

The pangenome-based somatic variant detection analyses are accompanied by a dedicated manuscript focused on this topic.[18] The assemblies from HPRC release 1, together with GRCh38 and CHM13,[35] were used to construct pangenome graphs using Minigraph-Cactus.[65] This generated both allele-frequency (AF)-filtered and unfiltered pangenome graphs. In this study, only the clipped graphs were used.

The pangenome-inferred diploid assembly was generated from the AF-unfiltered pangenome graph. First, HG005 whole-genome sequencing (WGS) data were used to generate k-mer profiles with KMC using the options -k29 -m128 -okff.[106] Haplotype sampling was then performed by matching k-mer profiles derived from the WGS data to haplotypes in the pangenome graph using vg (v1.64.1) haplotypes.[107] This procedure imputes a subgraph from the clipped, unfiltered HPRC Release 1 pangenome graph,[107] using GRCh38 as the reference backbone and incorporating two recombinant assemblies (hap1 and hap2). Because the HapMap mixture's "somatic" alleles are in fact germline variants contributed by the other constituent cell lines, k-mer profiles were generated from HG005 WGS data alone, rather than from the mixture, using KMC (options -k29 -m128 -okff103), to avoid any possibility of these pseudo-somatic alleles influencing the inferred haplotype structure. These recombinant assemblies closely approximate the HG005 diploid assembly[18] and together constitute the pangenome-inferred diploid assembly used in this study.

To generate HapMap mixture call sets across GRCh38, the pangenome, and the pangenome-inferred diploid assemblies, we used 500× Illumina short-read WGS data from the WashU-VAI GCC. Reads were aligned to GRCh38 using BWA-MEM (v0.7.17).[61] The same dataset was also aligned either to the clipped, AF-filtered pangenome graph or to personalized pangenome graphs containing GRCh38 and the pangenome-inferred diploid assembly using vg giraffe (v1.64.1).[108] Alignments were subsequently surjected to GRCh38 for the AF-filtered pangenome graph, or to the pangenome-inferred diploid assemblies for the personalized pangenome graph, using vg surject.[35] All alignments were processed with GATK (v4.1.0.0),[94] including duplicate marking, indel realignment, and base quality score recalibration (BQSR) before variant calling.

SNVs and indels were called from 500× short-read data using MuTect2 (v4.5.0.0)[50] in tumor-only mode. Variants were subsequently filtered with FilterMutectCalls, and only calls annotated as "PASS", "haplotype" or "clustered_events" were retained. Variants were then decomposed and normalized with bcftools norm (v1.20-24-g5d48dbc3)[102] using the options -a –atom-overlaps '.' –check-ref w -N, and finally separated into SNVs and indels using Picard SplitVcfs (v2.9.0).

For benchmarking, we used HapMap mixture benchmarking variant sets based on GRCh38 and the pangenome-inferred diploid assembly generated using a genome-graph approach[17] (https://zenodo.org/records/17544617). The benchmarking was performed within the confident regions on chromosomes 1-22, X, and Y. GRCh38 confident regions were further stratified into easy, difficult, and

extreme regions as defined by the SMaHT network (https://data.smaht.org). These stratifications were lifted over to pangenome-inferred diploid assembly using UCSC liftover and the corresponding chain files generated by nf-LO.[109]

The somatic SNVs were benchmarked using RTG vcfeval (v3.12.1)[110] using options –squash-ploidy and –all-records. SNVs were partitioned into nine variant allele fraction (VAF) bins: 0-0.5%, 0.5-1%, 1-1.5%, 1.5-2%, 2-3%, 3-4%, 4-5%, 5-10%, and 10-16.5%. For recall assessment, the benchmarking variant set was stratified by VAF, and each VAF bin was compared against the full call set. For precision assessment, the call set was stratified by VAF, and each bin was compared against the complete benchmarking variant set. Precision and recall were calculated as follows: Precision = TP / (TP + FP); Recall = TP / (TP + FN).

Direct alignment to an unfiltered population pangenome graph is not recommended, because alleles the sample does not carry add spurious alignment paths that reduce accuracy. Prior work mitigated this with a static AF filter that retains only variants above a fixed frequency (here, >10% in the population used to construct the graph). This filter, however, is not informed by the sample being analyzed: it discards genuine germline variation the sample carries below the threshold, while retaining many common variants irrelevant to that sample. The recommended alternative is adaptive, sample-informed filtering through pangenome personalization, in which vg haplotypes selects the graph traversals that best match a sample's own k-mer profile, in effect a tailored allele filter that removes spurious alleles. The pangenome-inferred diploid assembly is derived from this personalization and is the approach we recommend. We nonetheless include the static AF-filtered alignment as a historical comparator, since it reflects how this has previously been done and readers may wish to know how it performs, with those results presented in the supplement (Figure S6).

Evaluating DSA-based somatic variant detection approach, contributed by Min-Hwan Sohn and Andrew B. Stergachis

Benchmarking of mutation calling using DSA is accompanied by a dedicated manuscript.[19] We established a comprehensive benchmarking framework using *in silico* tumor-normal mixtures at three dilution ratios (1:4, 1:9, and 1:19 tumor:normal) corresponding to approximately 20%, 10%, and 5% variant allele fractions. These dilution ratios were designed to simulate the detection of low-fraction mosaic somatic variants in donor tissues within a variable range that depends on the timing of mutation acquisition during development and the proportion of affected cells in the sampled tissues. *In silico* mixtures were generated by independently subsampling pure normal (COLO829BL) and tumor (COLO829) data for Fiber-seq or Illumina short-read sequencing data aligned onto COLO829BL DSA (samtools *view* –subsample) to achieve target coverage ratios. The subsampled files were then merged into composite alignment datasets at total haplotype coverage of 20× for Fiber-seq and 50× for Illumina short-reads. We performed the alignment onto the DSA, leveraging pbmm2 v.1.14.99 for Fiber-seq and bwa v.0.7.19 with BWA-MEM algorithm (*-k 20 -w 105 -d 105 -r 1.3 -c 12000 -A 1 -B 4 -O 6 -E 1 -L 6 -U 18 -Y -K 100000000*) for Illumina short-read sequencing data. For comparative analysis against GRCh38, the DSA-mapped Fiber-seq mixtures were first converted to unaligned BAM format and then realigned to GRCh38, preserving identical read composition under different assembly coordinates. Short-read mixtures were converted to FASTQ using samtools fastq and remapped to the reference of interest with the same parameters.

Next, we performed SNV calling using DeepVariant v1.6.1[67] on COLO829BL Fiber-seq and two COLO829 Fiber-seq biological replicates independently. Somatic variants were identified by retaining only those variants present in both COLO829 replicates but absent from the COLO829BL with additional quality control and masking steps to ensure high quality variant set. This involves masking genomic positions containing variants that are either unique to single tumor replicates or present only in the normal sample from evaluation to prevent mis-categorization of variants into confusion matrix; the former representing potentially true but somatic events with lack of strong evidence, and the latter representing potential culture artifacts or assembly artifacts since these variants should not appear as somatic relative to an assembly derived from the same source. When evaluating precision in DSA space using short-read sequencing data in particular, we also made use of ''haplotype-agnostic'' DSA-based somatic SNV reference set created by projecting SNVs from one haplotype onto the other corresponding haplotype using the variation graph (vg) toolkit.[66] This enabled us to account for discovering false-positive SNVs arising in the diploid DSA, merely owing to the random assignment of short-reads to one of the two haplotypes, as such reads are often too short to be confidently mapped between near-identical haplotypes, which is an issue that long-read sequencing data generally does not suffer from. For GRCh38-based evaluation, DSA-based somatic SNV coordinates were moved to GRCh38 coordinates using a donor-specific graph (DSG)[19] constructed by minigraph-cactus[65] from Telomere-to-Telomere assembly of the CHM13 cell line[111] (T2T-CHM13v2.0), GRCh38 and COLO829BL DSA haplotypes 1 and 2, with SNV positions from COLO829BL DSA space injected into the graph and surjected onto GRCh38 using vg toolkit[66] to generate the corresponding reference set. The same approach was applied to generate a T2T-CHM13-based reference set by surjecting the injected SNV positions onto T2T-CHM13v2.0 coordinates. Genomic loci unique to the COLO829BL DSA with respect to each reference were defined in a similar fashion using DSG, retaining only regions that could not be surjected to GRCh38 or T2T-CHM13v2.0, respectively, and are ≥1kb in length.

Finally, we ran DeepSomatic v1.9.0[40] in tumor-only mode at different dilution ratios on DSA, GRCh38 and T2T-CHM13, respectively. Somatic SNVs were further filtered to include only PASS variants and exclude variants in masked genomic position. Furthermore, we restricted the evaluation only to pre-computed callable genomic regions in the DSA defined by accounting for the misassembled proportion of the genome and also large-scale deletions. GRCh38-based variants and T2T-CHM13-based variants were also filtered against a panel of normals (PoN) catalogues packaged within DeepSomatic to reduce germline contamination.[40] The

PoN records in GRCh38 coordinates were lifted over to T2T-CHM13v2.0 using the Picard LiftoverVcf (v.2.23.3) with default parameter following the exact method described in the link: https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/liftover/dbSNP.html. To define satellite regions of the DSA, we used repeatmasker annotation of the DSA and used entries with repClass "Satellite" as the indicator. Satellite regions for GRCh38 and T2T-CHM13v2.0 were obtained from the corresponding repeatmasker tracks in the UCSC Genome Browser. Non-satellite regions for each assembly were defined by subtracting satellite regions from the genome-wide coordinates.

Cell, Volume *189*

Supplemental information

Comprehensive benchmarking of somatic mutation detection by the SMaHT Network

The Somatic Mosaicism across Human Tissues Network

Supplemental information

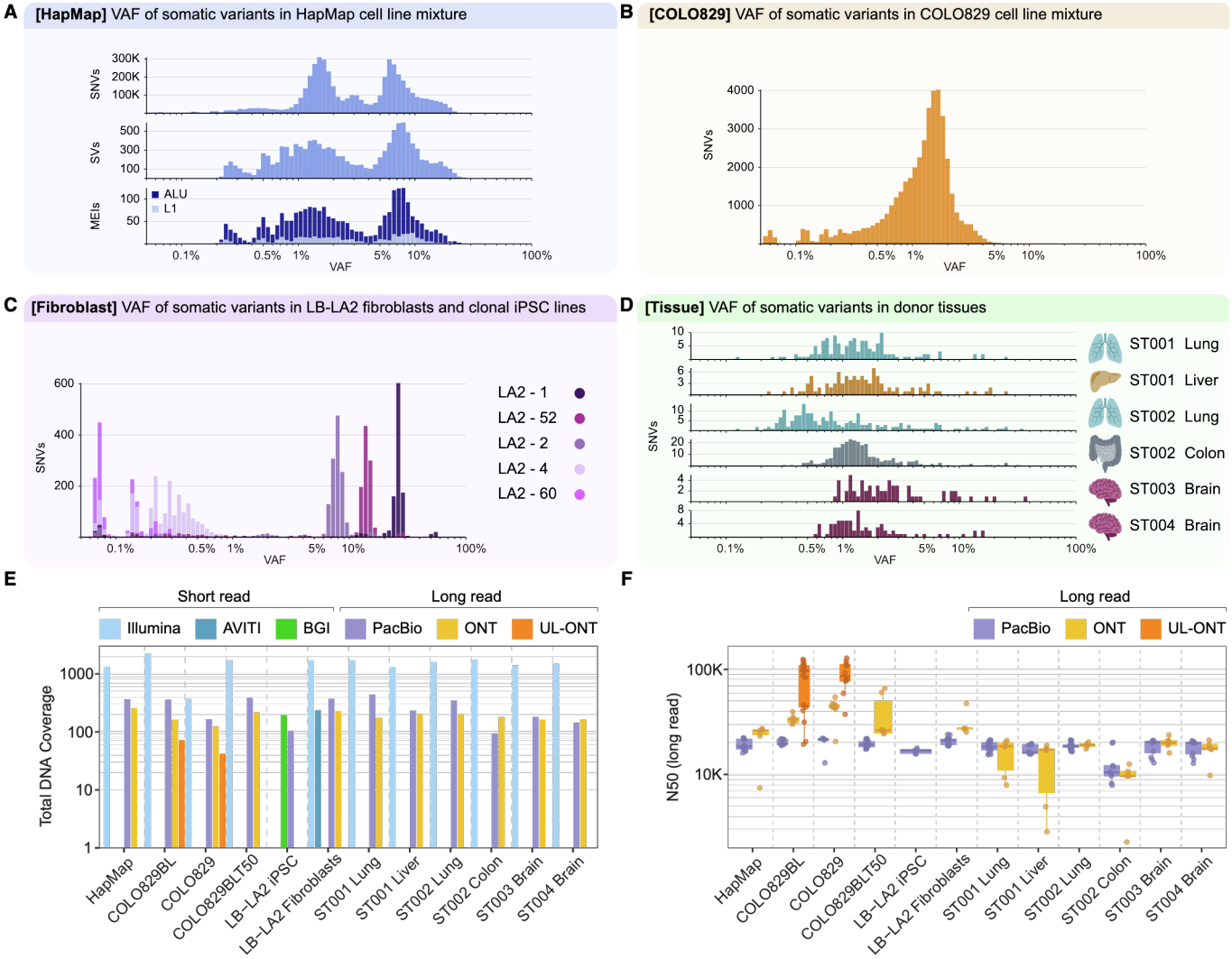


Figure S1. Summary of the SMaHT benchmarking experiment and data. Four benchmarking experiments were carried out by the SMaHT Network. **A)** In the HapMap experiment, germline variants present in some (but not all) lines simulated somatic mutations (SNVs, indels, SVs, and MEIs) at VAFs ranging from 0.25% to 16.5%. The resulting sample has a mix of 12 haplotypes. **B)** In the "COLO829" experiment, mutations in the melanoma line simulated SNVs and indels at VAF up to 5% in the mix. **C)** In the "Fibroblast" experiment, somatic SNVs and indels were present at VAF from 0.1% to 25% revealing the clonal structure in the fibroblasts. **D)** In the "Tissue" experiment, six frozen tissues obtained *post-mortem* from four phenotypically normal donors were analyzed to benchmark the detection of somatic mutations in production-like samples. The distribution of somatic mutations by VAF varied across samples. **E)** Total sequencing coverages of WGS datasets using all short-read and long-read sequencing technologies applied in the benchmarking experiments. **F)** N50 as a representative data quality metric for PacBio and ONT long-read sequencing. Details of data generation are provided in Tables S3 and S4.

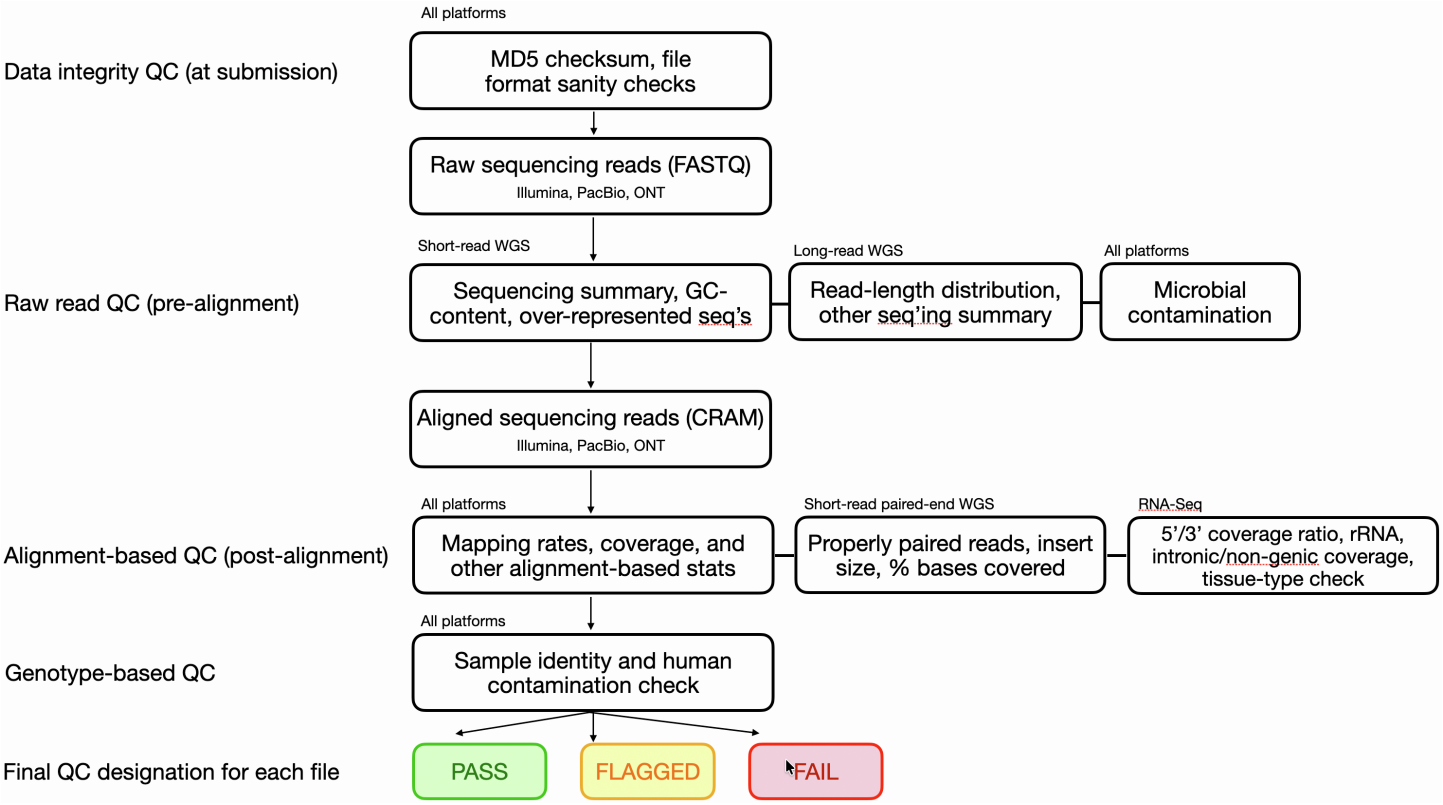
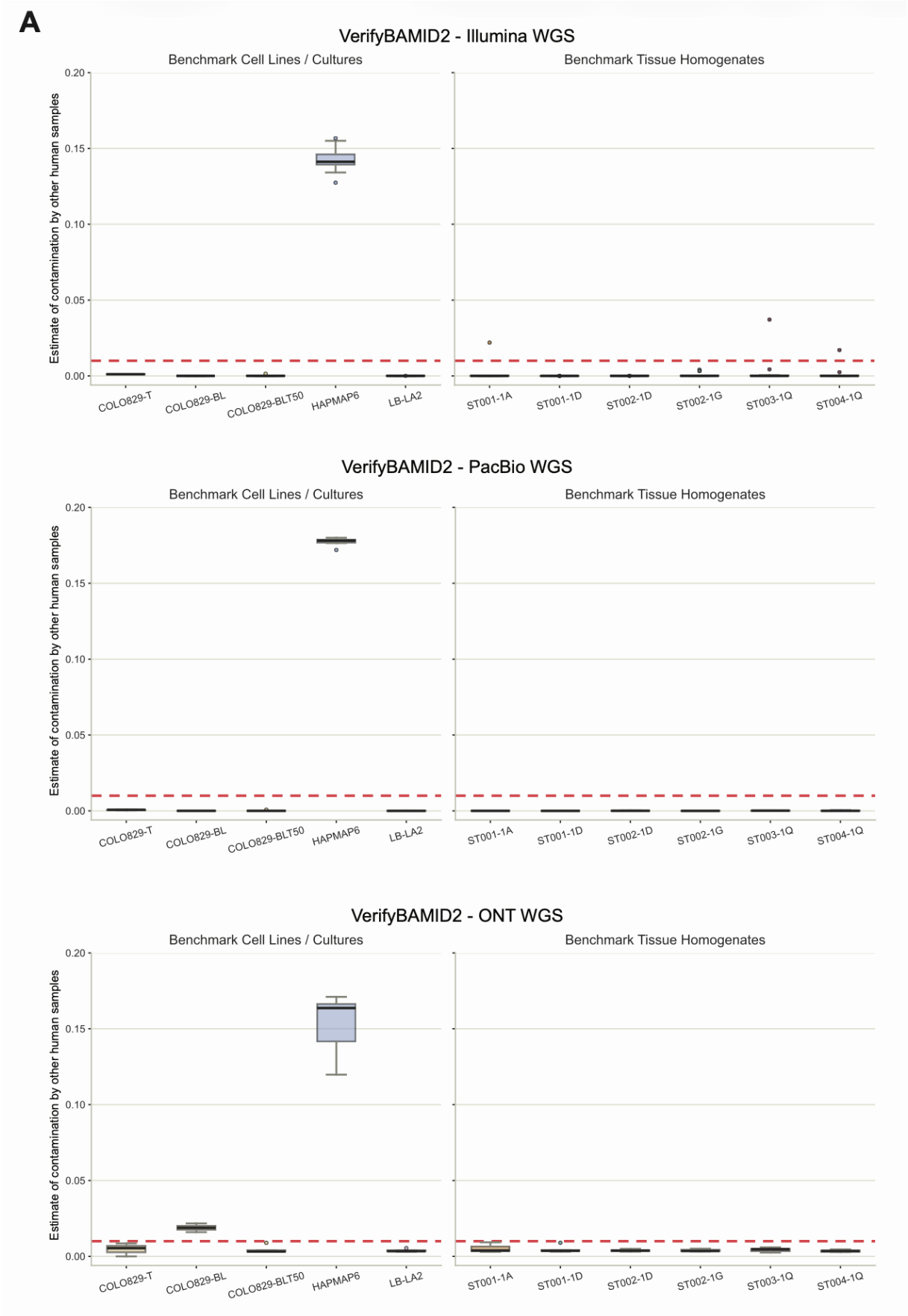


Figure S2. Overview of the SMaHT benchmark data quality control (QC) processes.

A


VerifyBAMID2 - Illumina WGS
Benchmark Cell Lines / Cultures
Benchmark Tissue Homogenates
Estimate of contamination by other human samples
0.20
0.15
0.10
0.05
0.00
COLO829-T
COLO829-BL
COLO829-BLT50
HAPMAP6
LB-LA2
ST001-1A
ST001-1D
ST002-1D
ST002-1G
ST003-1Q
ST004-1Q
VerifyBAMID2 - PacBio WGS
Benchmark Cell Lines / Cultures
Benchmark Tissue Homogenates
Estimate of contamination by other human samples
0.20
0.15
0.10
0.05
0.00
COLO829-T
COLO829-BL
COLO829-BLT50
HAPMAP6
LB-LA2
ST001-1A
ST001-1D
ST002-1D
ST002-1G
ST003-1Q
ST004-1Q
VerifyBAMID2 - ONT WGS
Benchmark Cell Lines / Cultures
Benchmark Tissue Homogenates
Estimate of contamination by other human samples
0.20
0.15
0.10
0.05
0.00
COLO829-T
COLO829-BL
COLO829-BLT50
HAPMAP6
LB-LA2
ST001-1A
ST001-1D
ST002-1D
ST002-1G
ST003-1Q
ST004-1Q

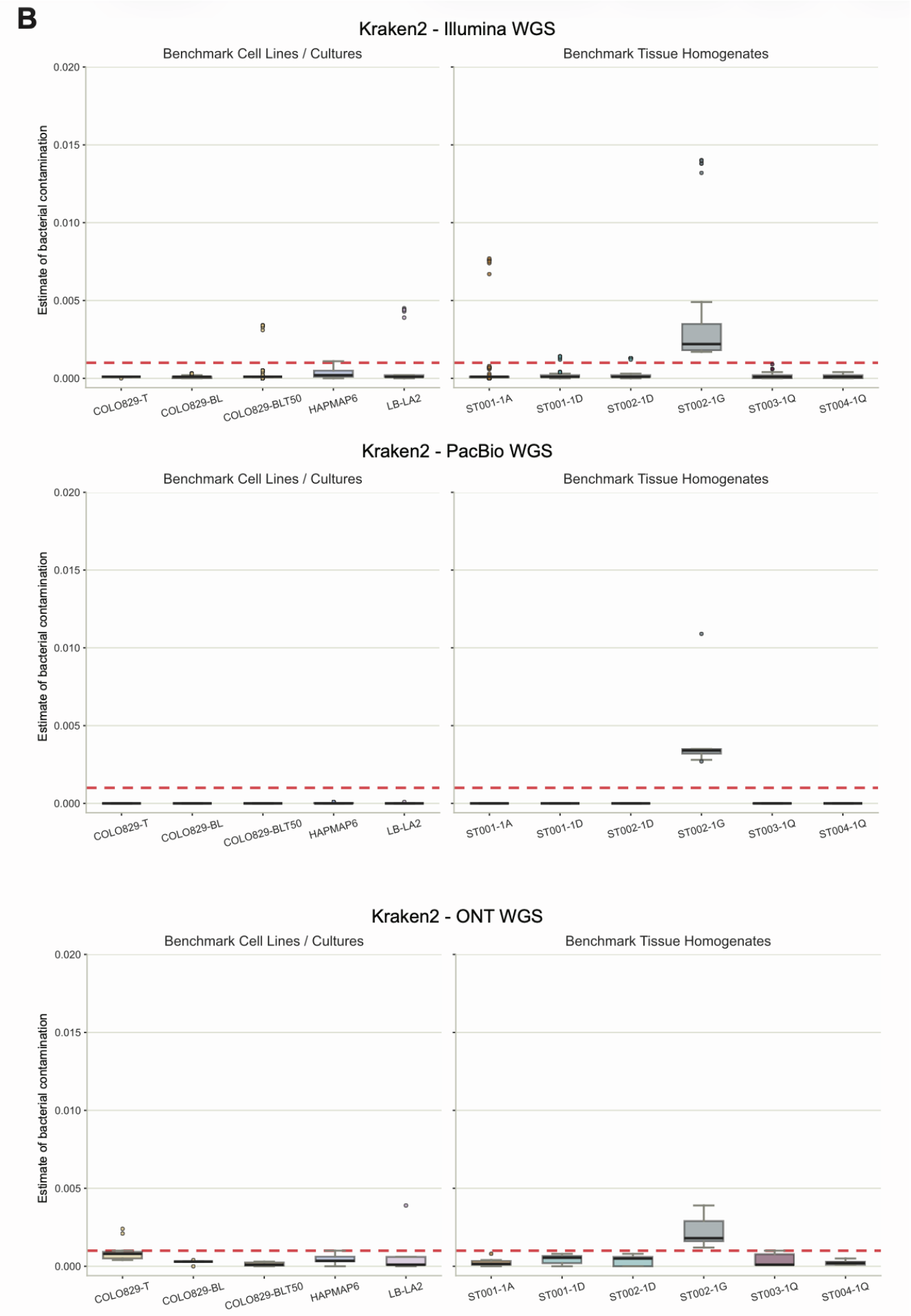

B
Kraken2 - Illumina WGS
Benchmark Cell Lines / Cultures
Benchmark Tissue Homogenates
Estimate of bacterial contamination
0.020
0.015
0.010
0.005
0.000
COLO829-T
COLO829-BL
COLO829-BLT50
HAPMAP6
LB-LA2
ST001-1A
ST001-1D
ST002-1D
ST002-1G
ST003-1Q
ST004-1Q
Kraken2 - PacBio WGS
Benchmark Cell Lines / Cultures
Benchmark Tissue Homogenates
Estimate of bacterial contamination
0.020
0.015
0.010
0.005
0.000
COLO829-T
COLO829-BL
COLO829-BLT50
HAPMAP6
LB-LA2
ST001-1A
ST001-1D
ST002-1D
ST002-1G
ST003-1Q
ST004-1Q
Kraken2 - ONT WGS
Benchmark Cell Lines / Cultures
Benchmark Tissue Homogenates
Estimate of bacterial contamination
0.020
0.015
0.010
0.005
0.000
COLO829-T
COLO829-BL
COLO829-BLT50
HAPMAP6
LB-LA2
ST001-1A
ST001-1D
ST002-1D
ST002-1G
ST003-1Q
ST004-1Q

Figure S3. Estimated contamination level across different benchmarking datasets. A) Levels of contamination from other human samples estimated in bulk Illumina (top), PacBio (middle), and ONT (bottom) WGS, using VerifyBAMID2 for the SMaHT benchmark cell line samples (left) and tissue homogenate samples (right). The red dotted horizontal line indicates the WGS QC threshold at 0.01 (1%). High estimated level of contamination for HapMap mix samples reflect thats this sample is a mixture of multiple cells lines from unrelated donors. **B)** Levels of bacterial contamination estimated in bulk Illumina (top), PacBio (middle), and ONT (bottom) WGS, using Kraken2 for the SMaHT benchmark cell line samples (left) and tissue homogenate samples (right). The red dotted horizontal line indicates the WGS QC threshold at 0.001 (0.1%). In both panels each dot represents a value for a bam file. Suffices in sample annotation mean the following: 1A -- liver, 1D – lung, 1G – colon, 1Q – brain.

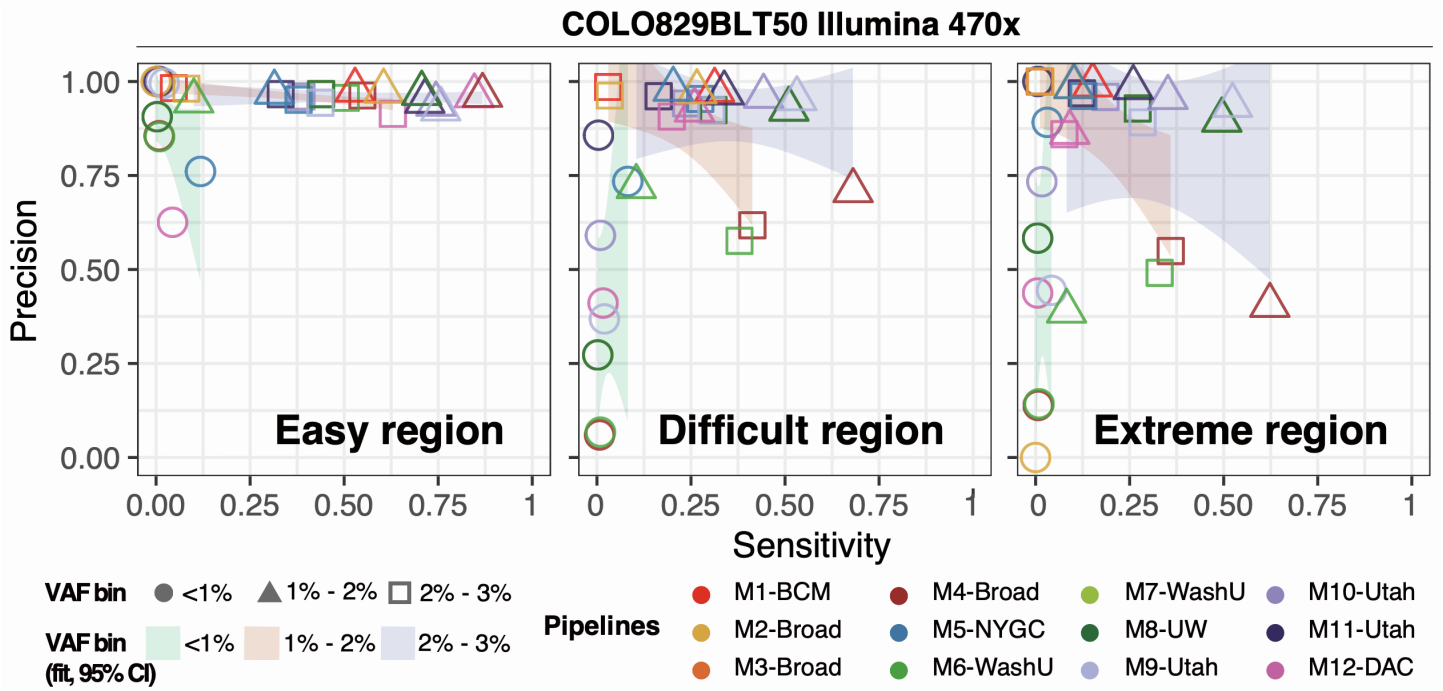


Figure S4. Precision–sensitivity performance for discovering SNVs and indels of each calling pipeline in COLO829 benchmarking in Easy, Difficult, and Extreme genomic regions, aggregated into three VAF bins (<1%, 1–2%, and 2–3%). Dots correspond to pipelines.

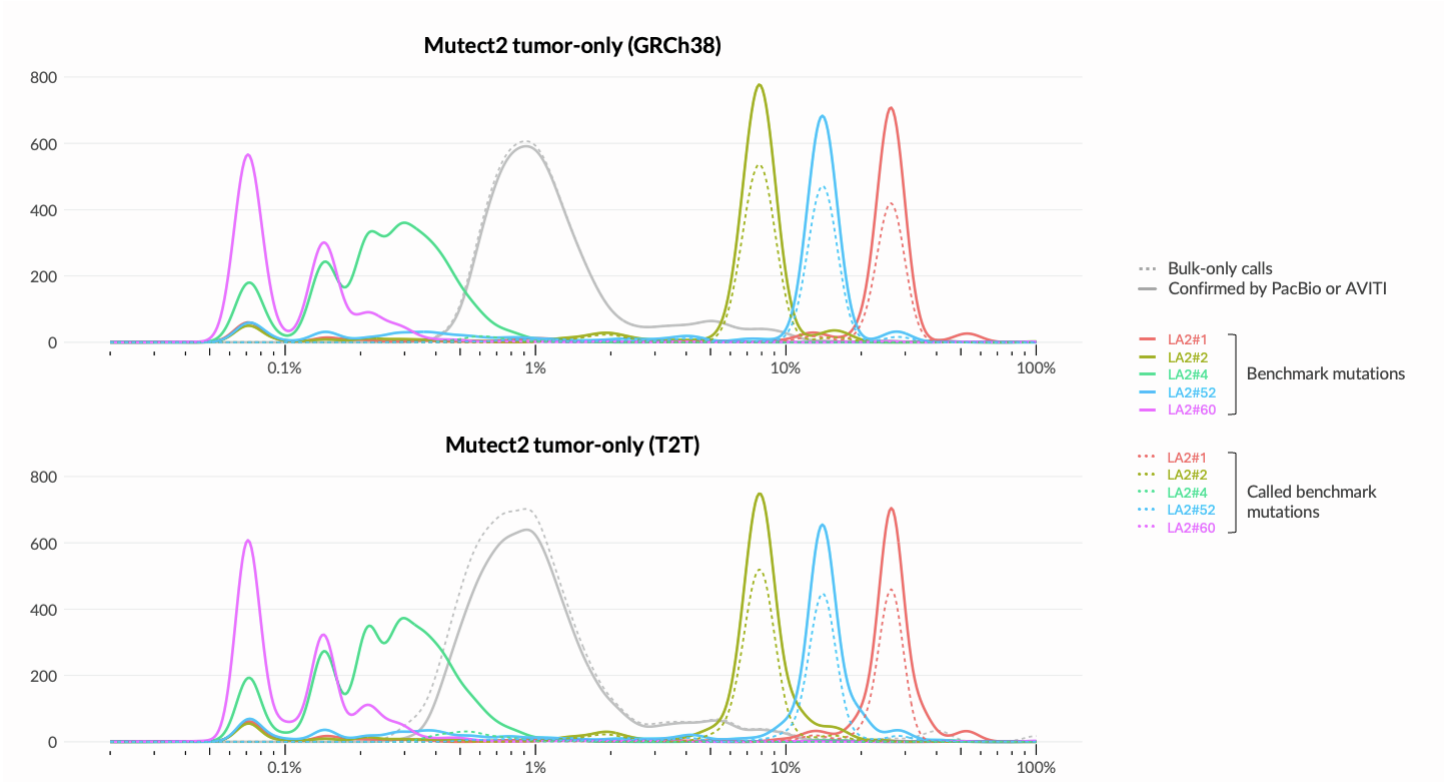


Figure S5. Performance of mutation calling in the fibroblast experiment using (A) GRCh38 and (B) CHM13 reference genomes. Approximately the same number of benchmark mutations (dashed colored lines) are detected using either reference at VAF >5%. In contrast, approximately twice as many benchmark mutations with VAF <1% are detected using CHM13. CHM13 also yields more fibroblast-only calls (i.e., calls outside the benchmark set; dashed gray line). However, a substantially smaller fraction of these calls is confirmed by either PacBio or AVITI when using CHM13 (solid gray lines).

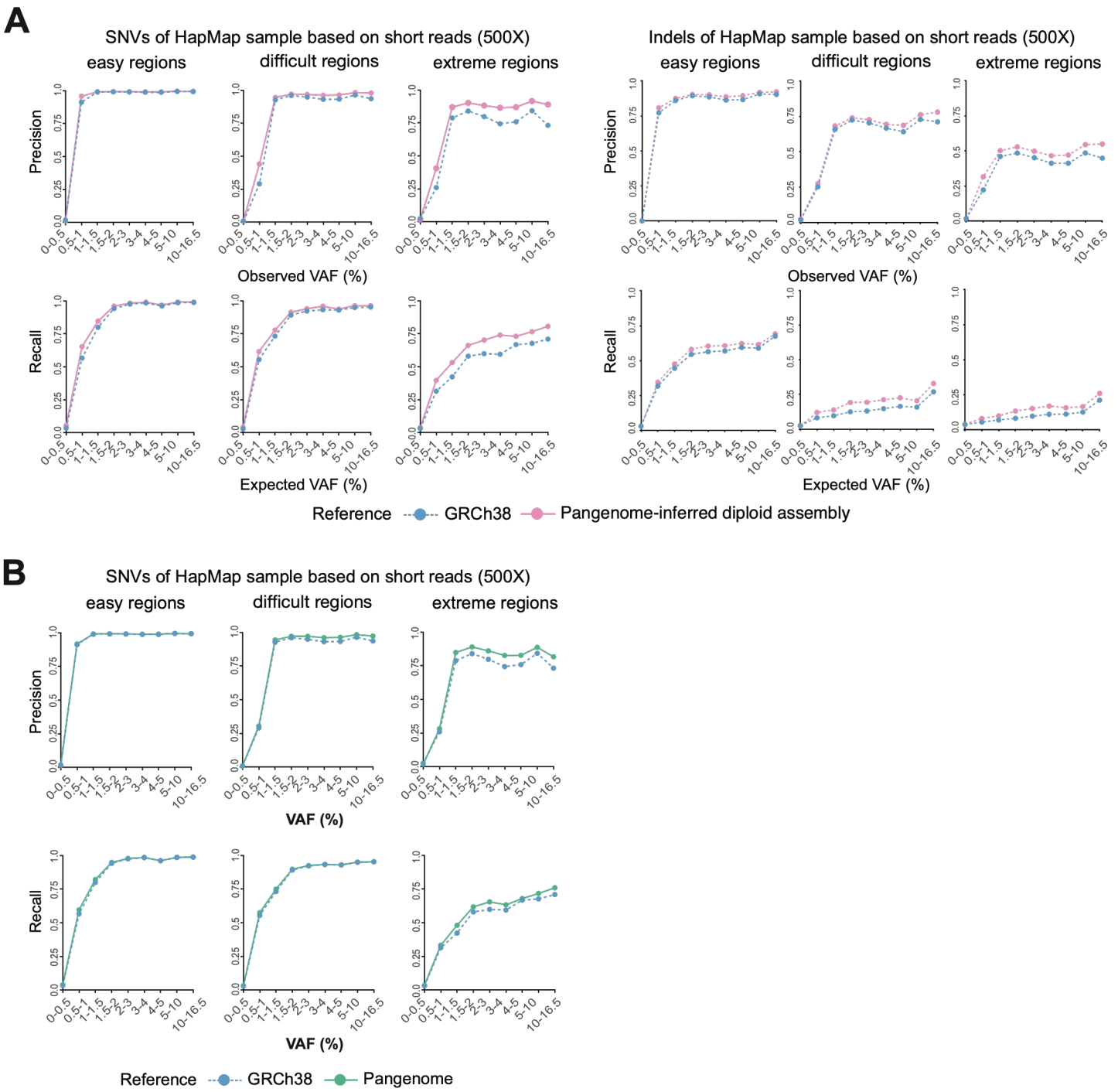


Figure S6. Pangenome improves somatic SNV and indel detection in the HapMap mixture. A) Construction of a pangenome-inferred diploid assembly for HG005, using k-mer content from HG005 short-read WGS data combined with haplotype sampling from the HPRC v1.1 pangenome graph (excluding all six HapMap samples). The resulting pseudo-diploid assembly partially reconstructs the HG005 genome. Somatic SNV (left) and indel (right) calling performance was benchmarked between GRCh38 and the pangenome-inferred diploid assembly using 500× short-read WGS data from the HapMap mixture; results shown are for haplotype 1 of the pangenome-inferred diploid assembly. **B)** Somatic SNV calling performance was benchmarked between GRCh38 and the pangenome-guided alignment (reads aligned to allele-filtered HPRC v1.1 pangenome followed by surjection to GRCh38) using 500× short-read WGS data from the HapMap mixture.

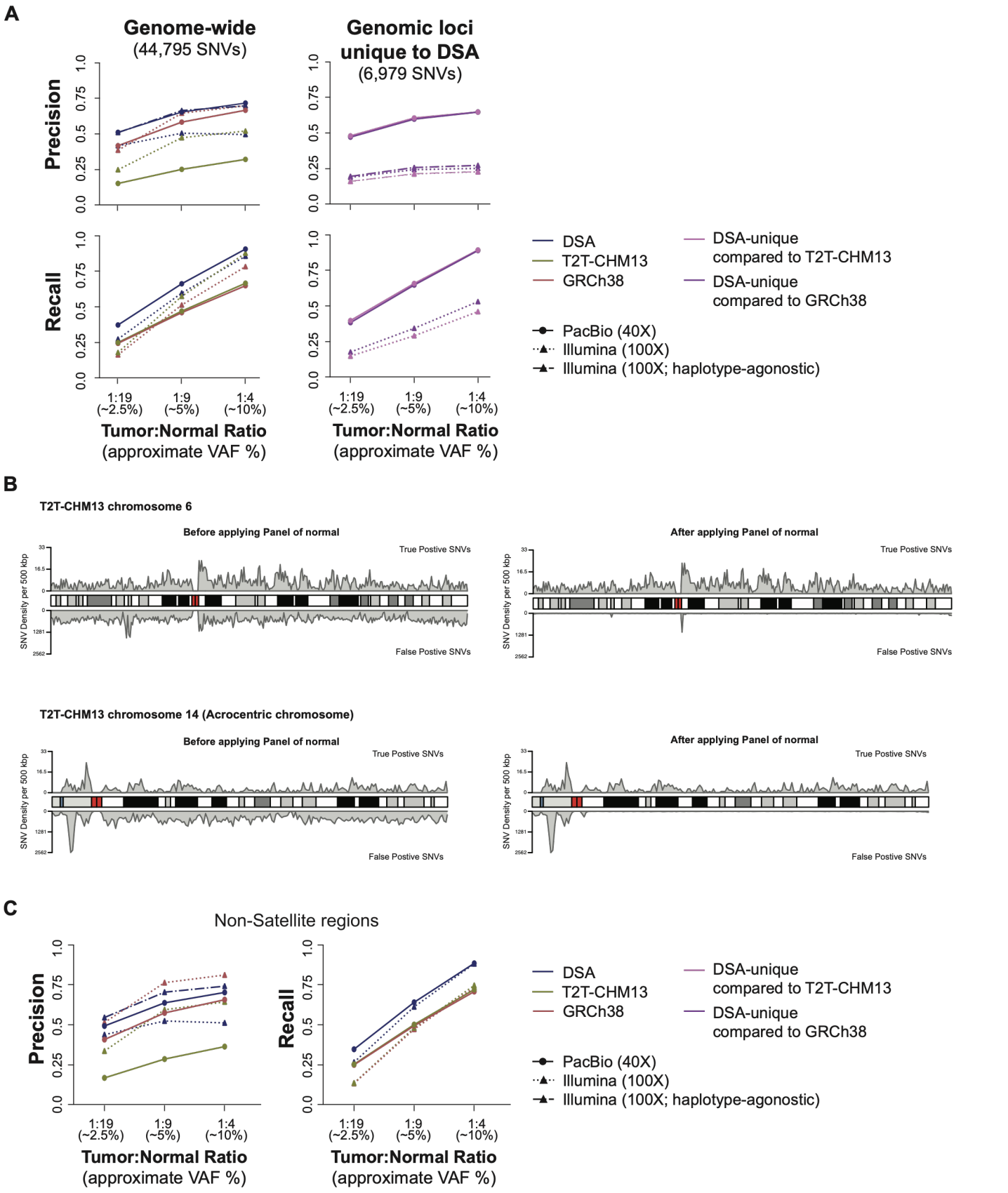


Figure S7. Benchmarking different assemblies (DSA, GRCh38 and T2T-CHM13) and sequencing modalities (short-read sequencing data and long-read sequencing data) for somatic SNV discovery. A) Precision and recall for sSNV discovery using 40x PacBio HiFi long-read sequencing data and 100x Illumina short-read sequencing data, each from a separate in silico mixture of the COLO829BL and COLO829 lines for GRCh38-, T2T-CHM13 and DSA-based approaches. **B)** Ideogram view illustrating the density of somatic SNVs discovered using T2T-CHM13 reference genome and the effect of applying panel-of-normal data. **C)** Precision and recall for sSNV discovery within non-satellite regions of the genome, separated by different assembly usage and sequencing modalities.

Table S1. Benchmark Data Generators. The following institutions generated the SMaHT benchmark sequencing data and metadata that were submitted to and analyzed at the Data Analysis Center.

Institution (PI)	**Abbreviation used in this study**	**Role in the SMaHT Network**
Baylor College of Medicine (Gibbs)	BCM	GCC (Genome Characterization Center)
Broad Institute (Ardlie)	Broad	GCC
New York Genome Center (Germer)	NYGC	GCC
University of Washington and Seatle Children Research Institute (Bennett)	UW-SCRI	GCC
Washington University at St. Louis and Van Andell Institute (Wang)	WashU-VAI	GCC
Boston Children's Hospital (Walsh)	BCH1	TTD (Technology and Tool Development)
Boston Children's Hospital (Choudhury)	BCH2	TTD
Baylor College of Medicine (Zong)	BCM1	TTD
Dana Farber Cancer Institute and Boston Children’s Hospital (Burns)	DFCI-BCH	TTD
Mayo Clinic and Yonsei University (Abyzov)	Mayo	TTD
Stanford University and Yale University (Urban)	Yale	TTD
New York University (Evrony)	NYU	TTD
University of Massachusetts Medical School (Fazzio)	UMass	TTD
University of Michigan (Mills)	UMich	TTD
Weill Cornell Medical College (Landau)	Cornell	TTD
Harvard Medical School (Park)	DAC	DAC (Data Analysis Center)
National Disease Research Interchange (Bell)	TPC	TPC (Tissue Procurement Center)

Table S2. Comparison of biological and technical advantages (+) and disadvantages (-) of reference cell systems used for mutation benchmarking.

	Feature	HapMap lines	COLO829	Fibroblasts / iPSC
Scientific	Mutation types	Broad spectrum across all variant classes (+)	High burden of SNVs, indels, SV/CNVs (+); low burden of MEI (-)	Medium burden (+); rare SVs and MEIs (-)
	Mutation origin	Germline variants (-)	Tumor-derived somatic mutations (-)	Physiological somatic mutations (+)
	Haplotype structure	Multiple unrelated haplotypes (-)	Matched diploid haplotypes (+)	Matched diploid haplotypes (+)
	Control of mutation frequencies	Defined by line proportions (+)	Defined by line proportions (+)	Not experimentally controlled (-)
Technical	Cell growth	Immortal, fast-growing (+)	Immortal, fast-growing (+)	Finite lifespan (-)
	Reference assembly	Complete T2T diploid assembly available (+)	Needed to be done (-)	Needed to be done (-)
	Existing sequencing data	Extensively characterized with short- and long-read data (+)	Extensively characterized with short- and long-read data (+)	Requires long-read characterization (-)
	Sample composition	Requires mixing of multiple lines (-)	Requires mixing for three times (-)	Natural mix of clones (+)
Considerations for a more streamlined and rational design	Consider utilizing controlled mixtures of multiple iPSC lines. It is feasible to derive several iPSC lines from a single donor, using different sample sources (e.g., fibroblasts, blood, and urine-derived cells), with each line carrying distinct somatic mutations. Mixing these lines would allow precise control of mutation frequencies while preserving the original diploid haplotype structure, closely mirroring real biological samples. "Negative control" mutations—those absent from the mixture—could be defined using iPSC lines not in the mix, as demonstrated in our fibroblast-based benchmark. iPSC lines are more genomically stable than cancer cell lines, making them less prone to culture-induced artifacts and providing an inexhaustible source of material for validation and follow-up studies. These lines, along with primary fibroblasts, could also serve as the foundation for generating telomere-to-telomere DSAs, enabling the development and benchmarking of methods that leverage personalized reference genomes. One limitation of this approach is the relatively low abundance of SVs, CNVs, and MEIs in iPSC lines, compared to SNVs, reflecting the relatively low frequencies of these variant types in normal human tissues. However, this could potentially be mitigated by inducing structural variation and MEI activation in primary cells prior to reprogramming.			

Table S5. Various QC checks performed on whole genome sequencing data at the SMaHT Data Analysis Center. The QC statuses are included in the File Manifest file, that are downloaded upon the files are selected (usually for downloads) in the SMaHT Data Portal (https://data.smaht.org). For more detailed descriptions of the QC checks and tools used, see Methods.

WGS file checked	QC check	Tool used	QC status when check failed
FASTQ	Estimated % of microbial and viral contaminations. PASS when <0.1% bacterial, <0.3% viral, and >95% human.	Kraken2	FLAGGED
BAM/CRAM	Sample relatedness to check for duplicate sample or sample swap	Somalier	FAILED
	Estimated average coverage checked against target coverage	Mosdepth	FLAGGED
	Estimated % contamination from another human sample. PASS when is <1%.	VerifyBamID2	FLAGGED
	Mean insert size. PASS when >250 bp.	Picard	FLAGGED
	% of properly paired reads	Samtools	FLAGGED
	% of mapped reads	Samtools	FLAGGED
	% of duplicate reads	Samtools	FLAGGED

Table S6. Data access levels of the SMaHT benchmarking sequence data and derived data, including germline and somatic mutations, gene expression, and epigenetic profiles. Authorizations for *post-mortem* donors within the main production phase of SMaHT allow for open access of most data, with the clinical metadata and germline sequence to remain under controlled access.

	COLO829, COLO829BL, COLO829BLT50	**HapMap mixture**	**LB-LA2 fibroblasts and iPSCs**	**Tissues from *post-mortem* donors**
Sequence data (FASTQ, BAM/CRAM)	Open	Open	Protected	Protected
Germline mutations	Open	Open	Protected	Protected
Somatic mutations: All mutation types from a cell line or individual donors as well as aggregated results from multiple donors	Open	Open	Protected	Protected
Gene expression profile	Open	Open	Protected	Protected
Epigenetic profile	Open	Open	Protected	Protected

Table S7. Summary of recall and precision for SNV discovery in different experiments. Results for the best performing methods by F1-score are provided. ^the number true SNVs was estimated from discovery in the bulk and validation by long reads. For such intervals the maximum estimate of recall is provided. For MEI the # of benchmarking mutations is given for easy + difficult and extreme regions. '-' indicates when precision couldn't be evaluated due to calls in that VAF range.

Mutation type	Experiment (coverage)	# of benchm. mutations	VAF range	Recommended method	Recall, %; Precision, %		
					Easy	Difficult	Extreme
sSNVs	COLO829BLT50 (470X)	7,277	2-3%	Mutect2 + PoN + gnomAD filt. (M4)	86.8%; 96.3%	68.1%; 71.0%	62.3%; 40.5%
		19,461	1-2%	MosaicForecast + repeat filt. (M12)	62.9%; 91.4%	19.9%; 90.7%	7.7%; 86.0%
		12,942	<1%	Lancet2 + germline/repeat filt. (M5)	11.8%; 76.1%	8.2%; 73.4%	3.1%; 89.1%
	Fibroblasts (1,360X)	2,787	>5%	MuTect2 + BSMN filt.	80.7%; 100%	N/A	N/A
		1,284^	1-5%	MuTect2 + BSMN filt.	95.7%; 99.3%	N/A	N/A
		1,862^	0.5-1%	RUFUS + BSMN filt.	79.6%; 93.6%	N/A	N/A
		2,004	0.1-0.5%	RUFUS + BSMN filt.	8.0%; 89.0%	N/A	N/A
	HapMap mix (500X)	3,957,025	>1.5%	MuTect2 on pangenome-inferred diploid assembly	98.1%; 99.4%	94.9%; 97.6%	74.8%; 89.7%
		197,636	1-1.5%		84.7%; 99.1%	77.7%; 94.6%	53.1%; 87.0%
		2,081,540	0.5-1%		65.3%; 95.7%	61.6%; 44.0%	39.6%; 40.6%
sIndels	COLO829BLT50 (470X)	305	2-3%	RUFUS + moderate filt. (M10)	38.8%; 40.7%	2.8%; 18.8%	3.6%; 44.4%
		902	1-2%	RUFUS + moderate filt. (M10)	31.9%; 50.9%	1.8%; 37.5%	2.6%; 33.3%
		695	<1%	Lancet2 + germline/repeat filt. (M5)	14.0%; 38.8%	2.0%; 25.0%	0.0%; 0.0%
	Fibroblasts (1,360X)	110	>3% Non-homo	MuTect2 + BSMN	86.4%; 100%	N/A	N/A
		27	1-3% non-homo	MuTect2 or RUFUS + BSMN	85.2%; 95.8%	N/A	N/A
		72	>3% homo	MuTect2 + BSMN	41.7%; 100%	N/A	N/A
	HapMap mix (500X)	846,316	>1.5%	MuTect2 on pangenome-inferred diploid assembly	62.3%; 91.1%	22.6%; 74.7%	17.1%; 52.4%
		43,767	1-1.5%		47.7%; 87.5%	13.8%; 68.2%	9.9%; 50.2%
		408,702	0.5-1%		34.6%; 80.8%	12.0%; 27.4%	8.1%; 31.6%
sMEIs	HapMap mix MEI-targeted method	172 / 45	0-1%	TEnCATS	79.1%; 69.5.%		48.9%; 33.3%
		65 / 9	1-3%		93.8%; 96.0%		77.8%; 52.3%
		79 / 12	3-5%		91.1%; 94.9%		75.0%; 50%

		76 / 8	5-%		92.1%; 82.4%	87.5%; 29.4%
	HapMap mix long-read WGS (60×)	172 / 45	0-1%	PALMER	32.6%; -	17.8%; -
		65 / 9	1-3%		52.3%; 61.7%	33.3%; 22.6%
		79 / 12	3-5%		84.8%; 78.2%	50.0%; 50.0%
		76 / 8	5-%		94.7%; 80.1%	100%; 42.9%
	HapMap mix short-read WGS (200×)	172 / 45	0-1%	xTea_mosaic	10.5%; -	4.4%; -
		65 / 9	1-3%		36.9%; 93.3%	0.0%; 33.3%
		79 / 12	3-5%		78.5%; 100%	33.3%; 28.8%
		76 / 8	5-%		88.2%; 89.3%	62.5%; 38.1%

Table S8. Recommended methods, precision, and recall for different WGS coverage by long or short read sequencing. Challenging categories are mutually exclusive.

Platform (coverage)	Method	# of benchmarking mutations or Recall, %; Precision, %					
		Not challenging	Challenging				
			Insertion	Small SVs (50-250 bp)	Low VAF (<5%)	Tandem Repeat	Adjacent SV
HPRC assemblies	N/A	995	1119	477	1248	133	24
Illumina (500X)	Manta	84%; 92%	1%; 48%	95%; 67%	46%; 30%	60%; 26%	67%; 17%
PacBio HiFi (110X)	SVDSS	97%; 99%	96%; 76%	97%; 68%	35%; 63%	94%; 66%	92%; 51%
ONT (115X)	Severus	98%; 97%	95%; 86%	97%; 80%	21%; 80%	94%; 76%	92%; 57%

Table S9. Advantages and disadvantages of DSA and pangenome-based somatic mutation detection.

Advantages	Disadvantages
Align reads accurately and improve somatic variant detection accuracy, especially in difficult and extreme regions	More complex analysis requires specialized knowledge and different data processing starting from alignment
Align reads and detect somatic variants in non-reference sequences (sequences missed in GRCh38)	Existing graph pangenomes clip highly repetitive sequences, limiting the overall genomic space represented.
Reduce germline contamination in alignments and somatic variant call sets	More computationally intensive and resource demanding (memory, storage)
	Pangenome is only an approximation of an individual's actual germline genome

Table S10. A summary of technical capabilities for different duplex methods assessed in the benchmarking project. The "Hpy" subtype of NanoSeq use the restriction enzyme HpyCH4V to produce blunt-ended 200-300 fragments spanning TGCA cut sites of the source DNA. The "MBN" subtype of NanoSeq uses Mung Bean Nuclease to randomly shear the source DNA, permitting whole-genome coverage in contrast to NanoSeq-Hpy. The ddBTP subtype of CODEC-seq was previously described in Bae et al 2023 (https://doi.org/10.1038/s41588-023-01376-0) and uses restriction enzymes HpyCH4V and AluI to shear the genome. The DRv1 (Duplex-Repair, version 1) and DRv2 (Duplex-Repair, version 2) subtypes of CODEC-seq refer to two successive versions of a whole-genome fragmentation and end-repair strategy (Xiong et al 2022, https://doi.org/10.1093/nar/gkab855; Xiong et al 2026, https://doi.org/10.1093/clinchem/hvag046). ddBTP: dideoxybase triphosphate. ddNTP: dideoxyribonucleoside triphosphate.

Assay	Subtype	Scale of genomic coverage	Uses ddBTP in ER/AT	Amenable to target capture	Indel Calling	Single-strand mismatch/damage detection	Quantify VAFs without targeted sequencing	Minimum input required	Amenable to single-cell adoption	Instrument/Platform for sequencing costs
NanoSeq	*Hpy*	HpyCH4V-restricted fragments	Yes	Yes	Yes	No	No	200 ng	Yes	NovaSeqX Plus 25B
NanoSeq	*MBN*	Genome-wide	Yes	Yes	Yes	No	No	100 ng	Yes	NovaSeqX Plus 25B
CODEC-seq	*ddBTP*	HpyCH4V and AluI-restricted fragments	Yes	Yes	Yes	No	Yes	100 ng	No	NovaSeqX Plus 25B
CODEC-seq	*DRv1*	Genome-wide	No	Yes	Yes	No	Yes	100 ng	No	NovaSeqX Plus 25B
CODEC-seq	*DRv2*	Genome-wide	No	Yes	Yes	No	Yes	100 ng	No	NovaSeqX Plus 25B
VISTA-seq	—	Genome-wide	No	No	Yes	No	No	200 pg	Yes	NovaSeqX Plus 25B
CompDuplex-seq	—	Genome-wide	Thiol-ddNTP	Yes	Yes	No	No	1 ng	Yes	NovaSeqX Plus 25B
HiDEF-seq	—	Genome-wide	Yes	Yes	No	Yes	No	2–4 µg	No	PacBio Revio
ppmSeq	—	Genome-wide	No, but optional	No	No	No	Yes	250 ng	No	Ultima UG100

Table S11. Recommendation of experimental approaches and analyses to address specific research and clinical challenges.

Challenge & mutation type		Analysis approaches and tools		
		Bulk	**Duplex**	**scWGA**
Detecting mutations of interest at >1% VAF (e.g., pathogenic or cancer driver)	**SNVs and indels**	Recommended WGS or targeted sequencing Use MosaicForecast at <300× coverage	Recommended targeted sequencing. Not efficient for shallow whole-genome duplex, and success can depend upon choice of bulk control	Not cost-efficient
	SVs and complex rearrangements	Recommended with long reads Use SVDSS for HiFi and Severus / Delly / Sniffles2 for either HiFi or ONT	Unresolved	Not cost-efficient
	MEIs	Recommended with long reads or multi-platform integration Use PALMER or xTEA mosaic	Unresolved	Recommended targeted sequencing Use TEnCATS or eHATseq
Detecting mutations of interest at <1% VAF (e.g., pathogenic or cancer driver)	**SNVs and indels**	Recommended WGS or targeted sequencing Use RUFUS or Lancet2 with high coverage data	Recommended if targeted sequencing. Feasible with shallow duplex, but VAF estimation not possible.	Not cost-efficient
	SVs and complex rearrangements	Limited sensitivity and precision	Unresolved	More reliable option for large CNVs and SVs; not cost-efficient Use Manta
	MEIs	Recommended if targeted Use TEnCATS or eHATseq	Unresolved	Recommended targeted sequencing Use TEnCATS or eHATseq
Detecting and monitoring mutational exposures (SNV and indel): burden and spectra		Limited to developmental early and clonally expanded mutations	Recommended	Recommended but more costly than duplex

Detecting and monitoring clonal expansion by SNVs and indels	Recommended but not definitive	Limited to deep targeted sequencing	Definitive but not cost-efficient
Analyzing mutations in repetitive genomic loci (satellites, VNTR, SegDups)	Recommended with long reads and DSA/pangenome Use Flagger and NucFlag to ensure region is appropriately assembled	Unresolved	Unresolved
Lineage tracing	Not conclusive	Not conclusive	Recommended Use SCAN2 for mutation calling Filter mutations with Sequoia to trace developmental lineages
Functional consequence of mutation(s) of interest	Applicable for homogeneous cell types with expression or epigenome readout	Unresolved	Applicable when performing expression or epigenome readout from the same cell

Table S12. Remaining challenges in discovering somatic mutations.

Discovery mode	Challenge	Current state-of-the-art	Solution outlook
SNVs using bulk WGS based on the reference genome	Discovering in STR	Existing machine learning approaches (like MosaicForecast) perform well for coverages up to 300-400X	Novel machine learning and deep learning approaches trained for high coverage data
Indels bulks WGS based on the reference genome	Discovering in homopolymers	Discovering indels with >3% VAF using assembly-based tools (like Lancet and Mutect2)	Using high fidelity sequencing (like provided by Elements) and incorporate platform specific error model into calling algorithms
MEIs using bulk WGS based on the reference genome	Accurate estimation of VAF at low levels (<3%)	MEI-targeted methods combined with ddPCR	Building models from higher-depth sequencing data
	MEI discovery in near-centromeric regions	Assembly-based alignment and tools	Leveraging T2T-level donor-specific assemblies
SVs from bulk WGS based on the reference genome	Discovering SV at VAF < 10%	Using long read sequencing	Increased coverage and improved analytics
	Adjacent SVs	Using long read sequencing and tools like Delly or Severus	Improved analytics
All mutation types using DSAs/graph genomes	Nascent computational tooling accounting for diploid nature of human genome	Somatic mutation discovery largely limited to linear reference genomes	Graph-genome and diploid-genome-based somatic mutation discovery methods
	Limited panel of germline variants	Pangenome graphs made from hundreds of diverse diploid assemblies	Larger diverse pangenomes with germline variant calls across large diverse human populations
	Limited ways to confirm mutation calls (e.g. indels) in DSA-unique regions	Unresolved	Application of linked-read-like approaches with short read sequencing
SVs, CNVs, and MEIs using duplex sequencing approaches	False-positive chimeric reads and read mapping to repetitive genomic regions	Unresolved	New duplex sequencing approaches, including long read duplex approaches together with leveraging DSA
SNVs and Indels using duplex approaches	Distinguishing artifacts reaching single-strand consensus from true single-strand lesions	HiDEF-seq (SNVs); unresolved for indels	New approaches to model error rates based on sequence and coverage; model per-sample and per-technology errors and biases in sequence capture
	Accurately calling mutations beyond the traditional a4s2 requirement	DupCaller (Cheng et al., DOI: 10.1101/2025.07.13.664565.)	
	Variable input mass requirements due	Picogram-scale inputs; single-cell duplex sequencing	Approaches to experimentally pre-empt artifacts while preserving input DNA mass;

	to different inefficiencies in library construction		filter out mismatches in data generated from low-input but error-prone solutions
All mutation types using PTA-amplified single cells (WGS)	Throughput and QC	Amplification via 96-well plate; manual 4-loci PCR for the initial QC	Automated approaches for single cell/nuclei isolation and amplification; SNP-based library-free amplicon sequencing QC
	Cost of sequencing	Sequence to at least 15×	Progressive reduction in sequencing cost
	Application to any post-mortem tissue	Application has been demonstrated for fresh cells and nuclei from post-mortem brain, heart, lung, and colon	Development of experimental protocols tailored for tissues and conditions (like cell density, disease, post-mortem interval)

Table S13. Details of the variants confirmed in the COLO829BLT50 and HapMap mixture. Listed below are the sense (forward), anti-sense (reverse) as well as the wild-type (FAM) and variant (HEX) fluorescent TaqMan probes. For the probe designs the capital letters denote locked nucleic acid (LNA) bases.

Variant	COSMIC ID	Bio-Rad Assay	VAF Tumor (UW-SCRI)	VAF Tumor (BCM)	Expected VAF COLO829BLT50	Observed ddPCR VAF (UW-SCRI)	Observed ddPCR VAF (BCM)
SCN11A p.T340T	COSM36715	dHsaMDS432990804	49.38%	50.1%	1.00%	1.44%	1.41%
ZNF217 p.P651S	COSM25218	dHsaMDS581392264	50.41%	50.2%	1.00%	1.33%	1.33%
MADD p.S1620F	COSM26934	dHsaMDS851278352	99.91%	100%	2.00%	1.79%	1.64%
MAP4K1 p.P298P	COSM21033	dHsaMDS546407091	32.70%	33.2%	0.83%	0.76%	0.71%
RELN p.E331K	COSM36731	dHsaMDS550923980	49%	48.9%	1.64%	1.53%	1.40%

Sample ID	dbSNP ID	Expected AF	Chr.	Position (hg38)	Change	Observed AF (BCM)
HG02622	rs73825128	10%	chr3	28,336,081	G>C	15.18%
HG002	rs747651768	2%	chr6	157,128,737	A>G	1.8%
HG02257	rs199853723	2%	chr20	10,640,763	T>G	2.6%
HG02486	rs16961993	2%	chr17	17,768,088	G>A	4.5%
HG00438	rs149078332	0.5%	chr1	63,703,537	G>A	0.6%
HG005	rs2269756	83.5%	chr19	35,308,468	C>A	79%

Primers/Probe	HG02622	HG002	HG02257	HG02486	HG00438	HG005
Forward	CACTGTCAT GATGTCTGA	AGAGCGTGA CCAATAGAA	GATGGCTTTATTG AATAGTATAATG	AGTGCTCAGT AAAGAGGA	ACAGCATTCTCA AACTTG	CCTTGTAGATGAAA GACTTTG
Reverse	CCTGAAGTC AACCACTAA	GCTGTCTTTA CTTCCTTTTC	GCAGAAGTAAGA GTTCAGA	CCTCATATTAC ACCTGTGAC	GGCTTTTGTTCC TTAGTTA	CAGGGTCTTGCTAT GTTC
Wild Sense Dual-Labeled Probe	ccaagCttGttAg aAcct	aagaGgcAaaC gaTata	tcacAcaAacTagTcc ca	ctacCctGagGatT gctg	tacaTtcTggGgtTttgt	tggcTtgCgcCttAtaa
Mutant Sense Dual-Labeled Probe	ccaagCttCttAg aAcct	aagaGgcGaaC gaTata	tcacAcaAacGagTcc ca	atggcaActAgcCa aTctc	tacaTtcTggAgtTttgt	tggcTtgAgcCttAtaa