Title: Somatic Mosaicism across Human Tissues (SMaHT): Stage 2

Background: All cells accumulate DNA mutations throughout the life of an individual. Germline mutations, which only occur in germ cells, can be transmitted to the next generation. The mutations that occur in other cells (e.g. somatic mutations, SM) are transmitted to daughter cells and are maintained as *scars* in the genome. DNA mutations encoded within the zygote diverge at the two-cell stage, accumulate throughout the lifespan, propagate throughout the lifetime of a cellular lineage, and are unique to each individual. Some tissues, such as sun-exposed skin and the esophagus, acquire larger numbers of somatic mutations compared to other tissues, such as the brain. The constellation of somatic mutations within an individual is called *somatic mosaicism*, as different cells within a tissue and across an individual can have both shared and private somatic mutations. Like fingerprints, the constellation of somatic mutations found in Identical Twin A would differ greatly from the constellation found in Identical Twin B. This constellation of somatic mutations reflects the trajectories that cells took throughout development and reflect the repertoire of exposures, DNA replication errors, oxidative damage, and stochastic events that cell encountered across its lifespan.

While the broader research field demonstrated a role for somatic mutations in diseases such as cancer, VEXAS syndrome, and focal cortical dysplasia, SMaHT will deliver a suite of coordinated tools and technologies to systematically study somatic mutations in healthy tissues and clinically-relevant samples. This infrastructure will be designed to catalog, validate, and tease out the function of somatic mutations across a common set of tissues encompassing all three germ layers.

Stage 1 of SMaHT produced:

- Improved sensitivity and specificity via integrating short- and long-read sequencing pipelines
- Improved detection and validation of rare variants via duplex sequencing and single-cell studies
- Evidence that molecular profiles cluster by tissue type rather than donor or sequencing platform
- Evidence of early (shared across tissues) and later (tissue-specific mutations) mutations
- Methods for mitochondrial analysis, structural variant detection, genome assembly, mobile element discovery, and ultra-low-frequency variant identification
- A flagship benchmarking package of 28 manuscripts describing tools and standards for the field
- 125 publications cited by over 1300 non-SMaHT investigators

Program Goal: SMaHT Stage 2 asks three questions that transition the program from somatic mutation characterization to a comprehensive functional analysis platform: Which mutations matter? Why do they matter? How can we rigorously study them? To accomplish this, SMaHT Stage 2 will:

- Expand the healthy tissue reference to 200 total donors (Stage 1 = 150; Stage 2 = 50)
- Recruit at least three cohorts (N≥50 per cohort) of diseased or clinically-relevant tissues with matched controls and at least two tissues linking results to the SMaHT Variant Catalog
- Create a standardized, reproducible platform to identify, validate, and interpret the potential function of somatic mutations across tissues from healthy (N=50) and disease-relevant samples
- Connect mutations and clonal lineages to cellular, tissue, and organismal phenotypes
- Release all resulting data, metadata, benchmarks, workflows, and protocols at regular intervals
- Ensure tissue collection protocols are interoperable with other programs (e.g. GTEx, dGTEx)
- Incorporate methods established by gnomAD, ClinVar, and GA4GH to ensure interoperability

At the end of Stage 2, SMaHT will deliver a flexible, customizable platform that can be broadly adopted by the community. The platform will consist of standard reference samples, highly curated datasets, validated protocols, benchmarked tools, and portable, customizable workflows designed to optimize detection, validation, and functional characterization of somatic mutations. The SMaHT platform will showcase an integrated platform where researchers can compare their findings to SMaHT datasets and human genome references, select methods that fit their studies, and test how specific somatic mutations affect cells, tissues, and disease traits.

Proposed Initiatives:

The four parallel initiatives support each other and have safeguards in place so that if one component is delayed, the others can use Stage 1 resources until the delays are mitigated.

Initiative 1: Tissue Procurement (TPC): The TPC will expand the healthy tissue reference from 150 to 200 donors and provide standardized tissue collection and distribution through:

- Collecting and distributing 12-15 tissues from 50 additional healthy donors
- Maintaining and distributing tissues from the entire healthy donor cohort
- Providing standardized tissue collection, preservation, pathology, metadata, and distribution protocols that are interoperable with other programs (e.g. GTEx, dGTEx)

Initiative 2: Genome Characterization (GCC) Centers will generate broad coverage, deep sequencing data from 200 total donors (50 new donors) and conduct rigorous functional analysis in healthy and diseased/clinically-relevant samples through:

- Conducting deep, multi-omic analysis of 12-15 tissues from 50 healthy donors
- Providing standardized sequencing pipelines for many tissue types
- Developing methods to capture the functional consequences of somatic mutations
- Adding clinically-relevant cohorts (patients where somatic mutations are suspected to drive their phenotype) for at least 3 conditions, using short-read, long-read, RNA, duplex, and single-cell sequencing to link variants to cellular/tissue phenotypes

GCCs will be invited to bring well-characterized clinical cohorts into the SMaHT production pipeline. The GCCs will be expected to demonstrate strong evidence and/or rationale that somatic variants contribute to the phenotype and propose studies of at least 4 distinct sample types, which may include affected tissue, nearby unaffected tissue, blood, buccal swabs, fibroblasts, or other clinically-relevant samples.

Risk Reduction: Multiple GCC awards will reduce site-specific artefacts, ensure broad coverage of long-read sequencing platforms, and allow for investigator-initiated studies (e.g. use of donor-specific assemblies, mitochondrial analysis, single-cell studies, and duplex sequencing methods).

Initiative 3: National Somatic Mosaicism Commons Data Analysis & Organizational Center: This initiative will build off of the current Data Portal and continue the work of the Stage 1 OC to support rigorous, reproducible, and sustainable analysis, governance, and community use of SMaHT data by:

- Transforming the SMaHT data portal into the National Somatic Mosaicism Commons, which will use GA4GH standards to provide harmonized, analysis-ready data releases that are interoperable with gnomAD and ClinVar infrastructure
- Implementing a cost-effective operational model that will manage data ingestion, specimen metadata, and quality control, as well as publish analysis-ready releases, donor-by-tissue matrices, and disease catalogs
- Leading communication, outreach to broader communities, coordination, governance, and training efforts
- Supporting pilot projects for junior investigators inside and outside the Network and collaborative projects that bring together GCCs with outside teams that have small, well-characterized disease samples but lack the resources to run full somatic mutation sequencing pipelines

The Center will use GA4GH standards and be interoperable with gnomAD/ClinVar. Users will be able to see somatic mutations in a genome browser context, determine the reliability of specific mutations, learn the phenotypic effects of the mutation, and compare the mutation to background variation in healthy tissues. The collaborative projects, along with the larger projects proposed by GCCs in Initiative 2, will help SMaHT develop standardized protocols for small and large groups, broaden community participation, and bring high-value diseased samples into the program.

Risk Reduction: To ensure sustainability after Common Fund support, the National Somatic Mosaicism Commons will be built on interoperable NIH data-ecosystem standards with a staged transition plan to maintain catalog, workflows, and controlled-access data availability beyond the program period.

Initiative 4: Experimental and Computational Technology Development: This initiative will build from the tools generated in Phase 1 to rigorously detect somatic mutations while advancing integrated approaches focused on determining the functional consequences of somatic mutations and producing portable, benchmarked workflows. New experimental and computational tools will:

- Conduct spatial omics and single-cell multi-omic methods at scale
- Optimize detection of mutations at very low allele frequencies and in difficult-to-map regions
- Optimize pipelines to reduce tissue input, make use of lower-quality samples, increase sensitivity and recall, and reduce costs
- Develop methods for functional validation at scale (e.g. iPSC-derived products, epigenomics, etc.)
- Develop and disseminate portable workflows that improve rigor, reproducibility, comparability across SMaHT datasets, and usability by the broader research community

Stage 1 projects produced methods for detecting mutations in single-stranded DNA, repetitive regions, and mobile DNA; single-cell spatial DNA/RNA/epigenomic analyses; single-cell duplex sequencing; and broad single-cell genome coverage. In year 3 of Stage 1, SMaHT adopted duplex sequencing methods into the standard pipeline and is adding single-cell studies by the end of year 4. In coordination with Initiative 3, validated computational methods will be packaged with documentation, training, and compatibility reports and linked to National Somatic Mosaicism Commons data releases. These tools can be used in virtually all types of sequencing studies, not just those designed for detection of somatic mutations.

Risk Reduction: Multiple awards will foster development of experimental assays and robust analysis workflows. Projects will have clear Go/No Go milestones and metrics.

Deliverables:

By the end of Stage 2, deliverables from this program will include:

- A somatic mutation resource from 200 healthy donors
- Disease cohorts from at least three clinically-relevant conditions
- Standardized SOPs for tissue procurement, preservation, metadata, validation, and distribution
- Standardized functional analysis pipelines based on deep multi-omics data from 50 donors
- Standardized sequencing-to-function-to-analysis pipelines for healthy and diseased cohorts
- The National Somatic Mosaicism Commons Data Analysis & Organizational Center with versioned, analysis-ready, GA4GH/FAIR data releases
- Portable computational workflow bundles with documentation, training, and compatibility reports
- Benchmarked experimental and computational tools for low-abundance and hard-to-detect variant classes
- Standard approaches to test the effects of somatic mutations and epimutations on molecular, cellular, and tissue phenotypes
- A gnomAD/ClinVar research framework to document variant context and supporting evidence
- Interoperability with human germline variant detection, analysis, and classification infrastructure
- Interoperability with similar programs (GTEx, dGTEx, SenNet, HuBMap, etc.)
- Coordinated community infrastructure for workshops, cross-network collaboration, junior investigator pilots, and clinical seed projects

These products will move the field from isolated, assay-specific studies to a reproducible, cross-tissue framework for measuring, confirming, testing, and interpreting somatic mosaicism in health and disease.

Budget: $202.5M over 5 years.

TABLE 1 - BUDGET SUMMARY. In Table 1, for each initiative, indicate the lead IC (if known) and the requested budget from the Common Fund (not IC dollars) in millions of dollars for each year. Approximate RMS costs should be included in the budget now.

TABLE 1

Initiative	Lead IC	FY28	FY29	FY30	FY31	FY32	Total
Initiative 1 - Tissue Procurement	TBD	$2.5M	$2.5M	$2.5M	$2.5M	$2.5M	$12.5 M
Initiative 2 - Genome Characterization Centers	TBD	$14.5 M	$14.5 M	$14.5 M	$14.5 M	$14.5 M	$72.5 M
Initiative 3 - National Somatic Mosaicism Commons Data Analysis & Organizational Center	TBD	$9M	$9M	$9M	$9M	$9M	$45M
Initiative 4 - Experimental and Computational Technology Development	TBD	$14M	$14M	$14M	$14M	$14M	$70M
RMS - NIH staff salary/travel and NIH-organized workshops (TBD within total request)	TBD	$500 K	$500 K	$500 K	$500 K	$500 K	$2.5M
TOTAL		$40.5 M	$40.5 M	$40.5 M	$40.5 M	$40.5 M	$202.5 M

Budget component notes: Initiative 1 supports one Tissue Procurement Center ($2.5M/year). Initiative 2 supports 3-5 Genome Characterization Centers ($14.5M/year). Initiative 3 supports one National Somatic Mosaicism Commons Data Analysis & Organizational Center, including data ingest, quality control, releases, governance, clinical sample seed awards, junior investigator pilots, meetings, and communications ($9M/year). Initiative 4 supports 8-10 experimental technology development projects ($8M/year) and 3-4 computational methods and workflow awards ($6M/year). Activity codes, lead ICs, and RMS estimate remain to be finalized.