OSC (Common Fund)

The Common Fund

Concept Clearance: New Common Fund Program

TITLE: Somatic Mosaicism across Human Tissues (SMaHT) Program

Objective: Systematically illuminate somatic variation and capture the role SM play in the formation of the expanded personal genome that underlies biological processes of human health

Phase 1 Initiatives:

1. Generate a census of somatic variants in select tissues from diverse human donors
2. Develop innovative tools that optimize identification of variants
3. Create data and analysis toolkits using an open, FAIR workbench

Funds Available: $150M over 5 years for Phase 1

Program Period: 10 years; Phase 1 (Years 1-5), Phase 2 (Years 6-10)

Council Action: Vote on support of the Program

Somatic Mosaicism Across Human Tissues (SMaHT) Program

May 20-21, 2021

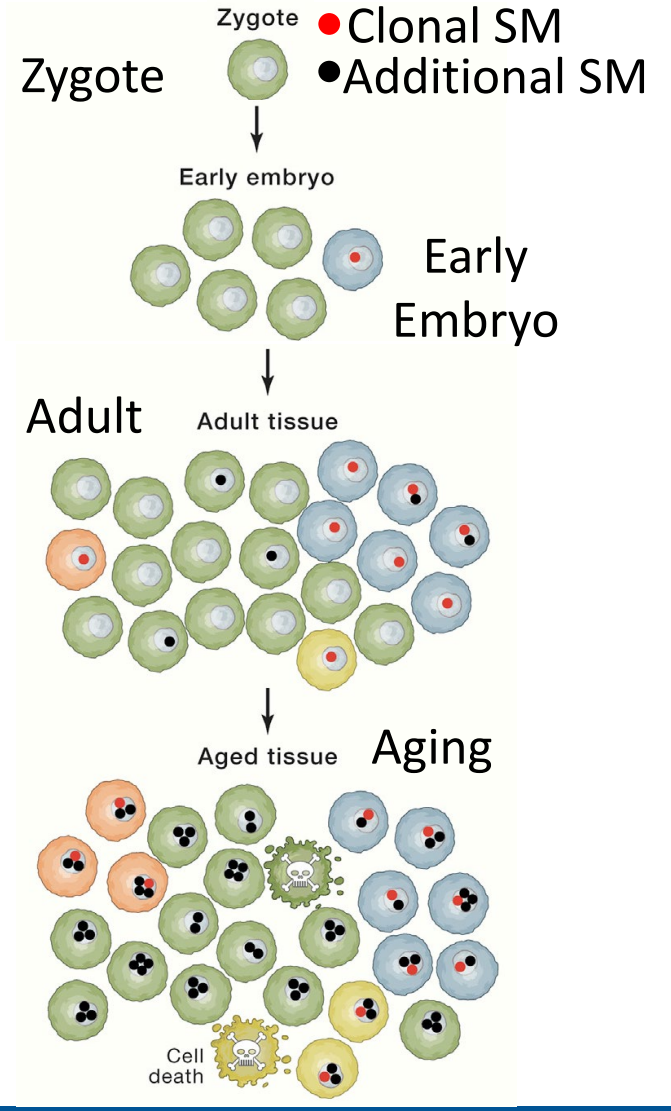
NIH Council of Councils

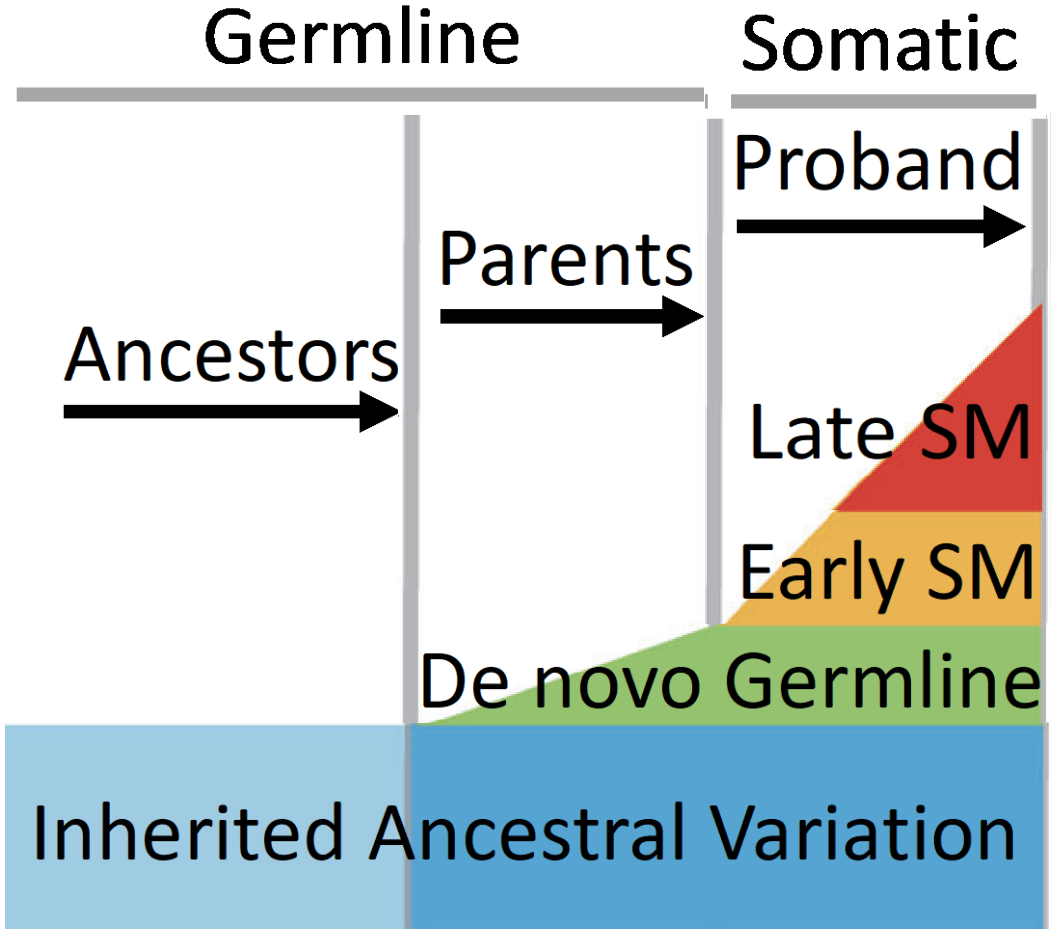
National Institutes of Health

Office of Strategic Coordination - The Common Fund

Mosaicism Expands the Personal Genome

The Common Fund





Adapted from Nishioka et al., 2019

Evolution of a Personal Genome

The Common Fund

Somatic Mosaicism across Human Tissues (SMaHT)

BSMN

COSMIC
Catalogue Of Somatic Mutations In Cancer

1000 genomes

Human Genome Project

Input from the Scientific Community

The Common Fund

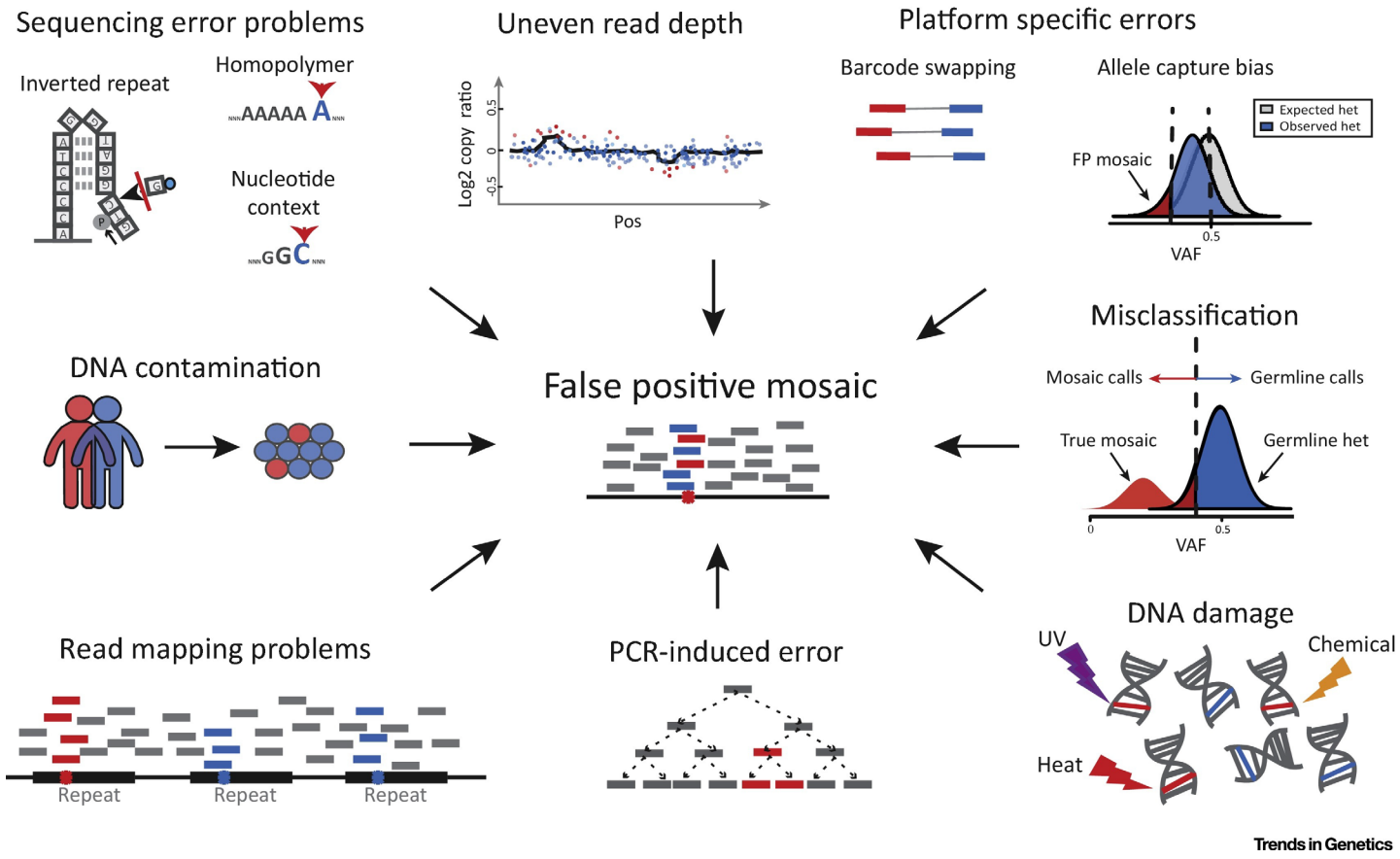
To identify gaps, challenges and potential programmatic scope, input was sought from the scientific community via an RFI (NOT RM 20-020), and two virtual Think Tanks held in July 2020. Five broad areas were identified:

- **Create a census of somatic variants** in different cell types, disease states, and life stages that can inform how they influence regulation and function
- **Build data analysis pipelines** to reliably **detect and annotate** structural variants and other somatic mutations
- Develop robust, **next-generation technologies** that **enhance sensitivity and spatial resolution** of somatic mutations across diverse tissue and cell types
- Develop carefully chosen model systems and new tools to **determine the biological consequences** of somatic variants
- Collaborate closely with similar programs to **build common benchmarks and analytical tools**

Challenges in Studying Somatic Variation

The Common Fund

- **Sensitivity:** low frequency (<5%) variants are hard to detect
- **Specificity:** many sources of technical variation provide significant background
- **Repetitive Regions:** CNVs, TEs etc. give rise to variants in ~45% of genome that is hard to sequence reliably



Dou et al., 2018

Phase 1 SMaHT Outcomes

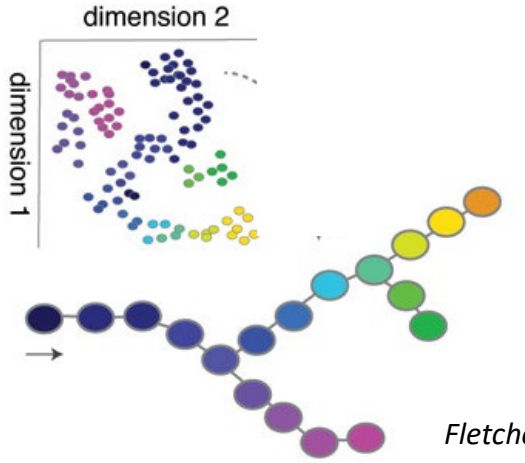
The Common Fund

Developmental Trajectory

Ravindran, 2019

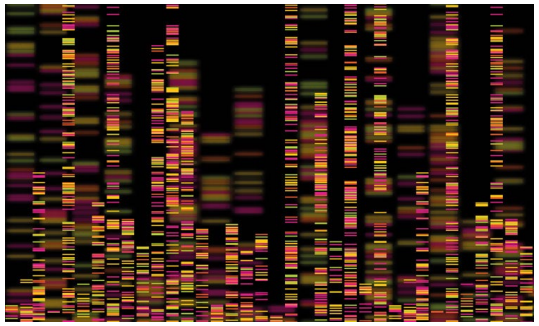
Image Credit Watchara

Cell Lineage Mapping



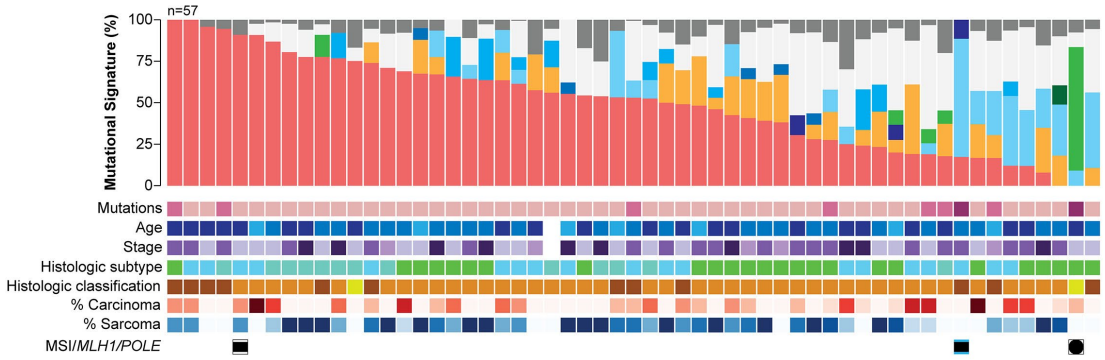
Fletcher et al., 2018

The “dark genome”



Muthuirulan, 2018

Mutational Signatures



Ashley et al., 2019

SMaHT Phase 1 Goals (Years 1-5)

The Common Fund

1. **Build personal genomes via systematic documentation of SNVs, Structural Variants, and Mobile DNA** in humans to understand biology of SM across the lifespan
2. **Develop next-generation tools and technologies** that improve sensitivity and resolution of somatic variants
3. **A FAIR, standards-driven data workbench** for visualization, analysis, and modeling of SMaHT data alongside data from other sources

Initiative 1: Tissue Mapping Centers

The Common Fund

Purpose: Comprehensively catalog somatic variants in core tissues from ~70-90 individuals

Deliverables:

- Biorepository of well-characterized tissues
- Reference catalog of tissue-specific variants
- Benchmarks, tools, and standardized data analysis pipelines

Deeper Understanding of:

- Variant location, frequency and tissue specificity
- Cell lineages and cell fate
- Variant accumulation in normal cells
- Types and extent of somatic variation in core tissues

Image Credit Watchara

Ectoderm
Hippocampus
Cortex

Mesoderm
Heart
Blood
Skin
Vagina
Ovary
Testis

Endoderm
Colon
Esophagus
Lung
Liver
PBL

Initiative 2: Tool & Technology Development

The Common Fund

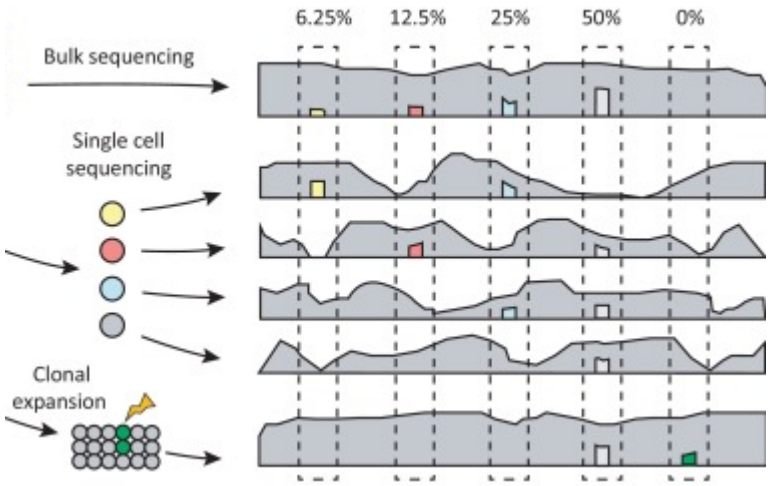
Purpose: Accelerate development, optimization and implementation of tools and data analysis pipelines for significantly improving sensitivity and specificity of variant detection and for integrated multi-omics analysis

Deliverables:

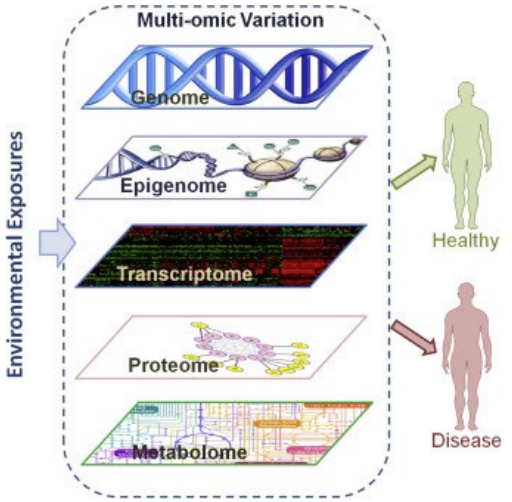
- Improved detection of low frequency variants
- Improved detection of somatic mosaicism in repetitive regions

Scientific Advances:

- Increased accuracy of detection of variants
- Analysis of structural variants, mobile DNA and repetitive DNA in small samples
- Integrate structural variants, mobile DNA and repetitive DNA into germline genetic studies



Dou et al., 2018



Sun et al., 2015

Initiative 3: Data Analysis and Program Coordinating Center

The Common Fund

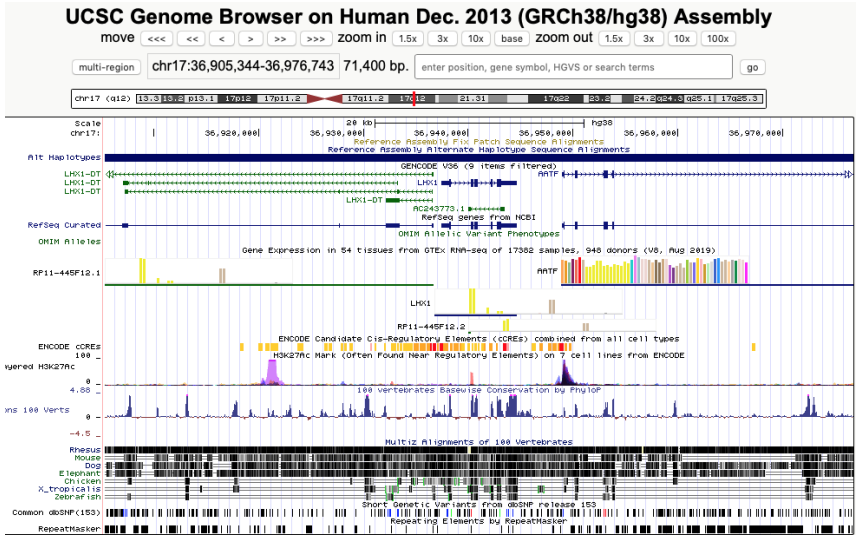
Purpose: Build a data coordination and organizational hub for the consortium that coordinates with other related programs across the NIH

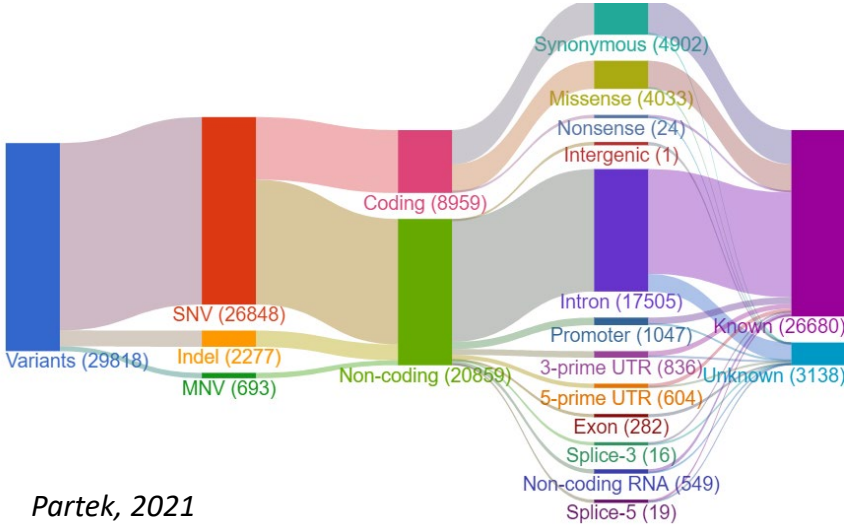
Deliverables:

- Rapid access and sharing of biospecimens, experimental protocols, datasets, and analytical pipelines
- Tools for analysis of changes across the lifespan and inter-individual variability
- Data workbench for studying somatic mosaicism that integrates with existing genomic data resources (e.g. AnVIL)
- Harmonization of SMaHT products with related programs

Scientific Advances:

- A better understanding of how timing, developmental trajectories, and mutational signatures expand the personal genome





Partek, 2021

Program Budget – Phase 1

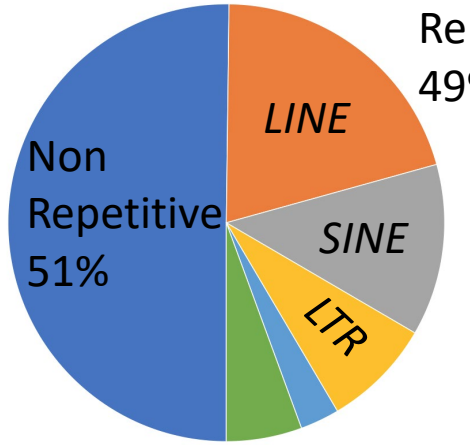
The Common Fund

	FY23	FY24	FY25	FY26	FY27	Total ($M)
Initiative 1: Tissue Mapping Centers (3-5 Centers)	14	16	18	20	20	88
Initiative 2: Technology Development Projects (5-10 Projects)	6	6	8	8	8	36
Initiative 3: Program Coordination and Data Analysis Center	2.5	3.5	4.5	5.5	7	23
RMS: NIH staff salary, travel and organized workshops	0.5	0.5	0.6	0.7	0.7	3
TOTAL ($M)	**23**	**26**	**31.1**	**34.2**	**35.7**	**150**

SMaHT Phase 2 Outcomes

The Common Fund

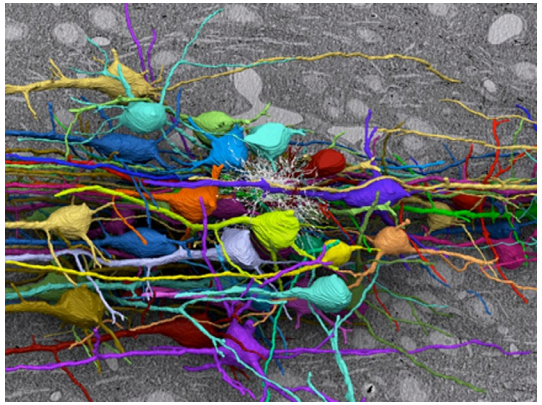
Repetitive Genome



Adapted from Rodic and Burns, 2013

Image Credit Watchara

Cellular Connectivity

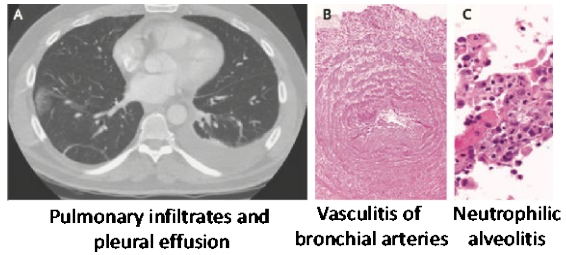


Nair, 2013

Population Level

Image Credit: Darryl Leja, NHGRI

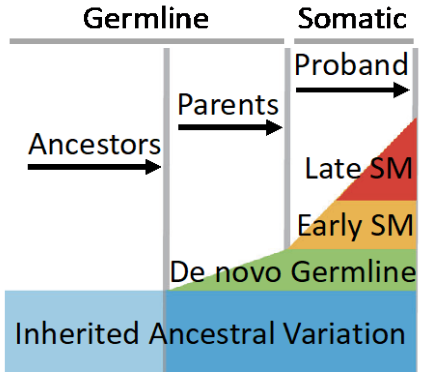
Undiagnosed Diseases

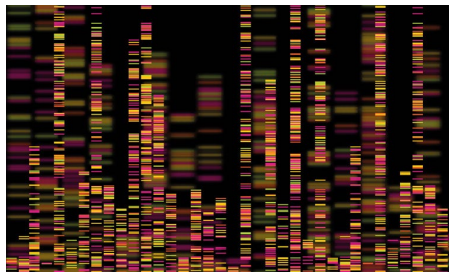


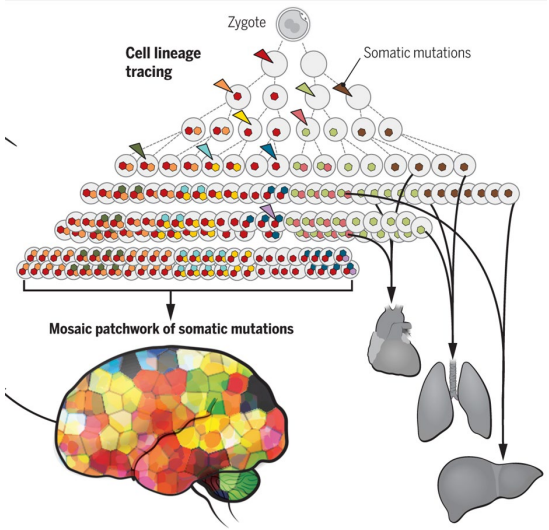
Beck et al., 2020

VEXAS Syndrome
UBA1 SM
Neutrophils
Monocytes

SMaHT as a Common Fund Program







- **Uniquely poised** to uncover the personal genome
- **Synergistic** and builds on other programs
- Community needs **benchmarks, standards, metrics, and analysis pipelines**
- **Cross-cutting** SMaHT tools, reference maps, and data analysis pipelines will **catalyze** future studies
- T**ransform** understanding of the genetics of disease and other biological processes

NIH Working Group

The Common Fund

Co-Chairs:

Walter Koroshetz (NINDS)
Rick Woychik (NIEHS)
Eric Green (NHGRI)
Roger Little (NIDA)
Joshua Gordon (NIMH)

OD
- **Richard Conroy**
- **Dena Procaccini**
- Tony Casco
- **Brionna Hair**

NCI
- Ian Fingerman
- Kevin Howcroft
- Philipp Oberdoerffer
- Wendy Wang

NHGRI
- **Adam Felsenfeld**

NIA
- Max Guo
- Alison Yao

NIAMS
- Yan Wang

NICHD
- Tuba Fehr
- Lesly-Anne Samedy-Bates

NIDA
- **Amy C. Lossie**

NIDCR
- Chiayeng Wang
- Lu Wang

NIEHS
- **Daniel Shaughnessy**
- Fred Tyson

NIMH
- Miri Gitik
- **Geetha Senthil**

NINDS
- Lyn Jakeman
- Jill Morris

Thank you

National Institutes of Health

Office of Strategic Coordination - The Common Fund

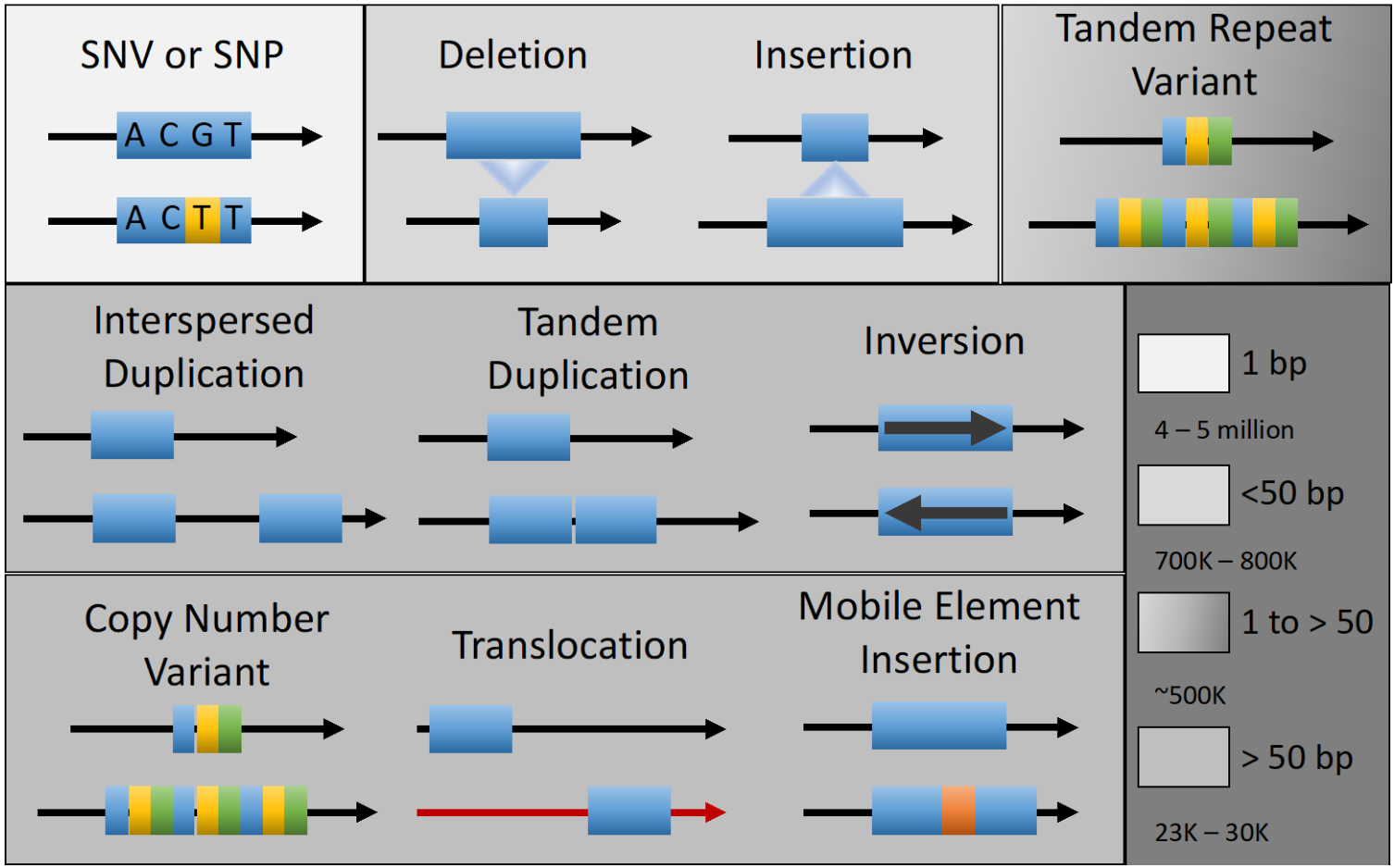
Supplemental Slides

National Institutes of Health
Office of Strategic Coordination - The Common Fund

Germline and Somatic Variants (SVs)

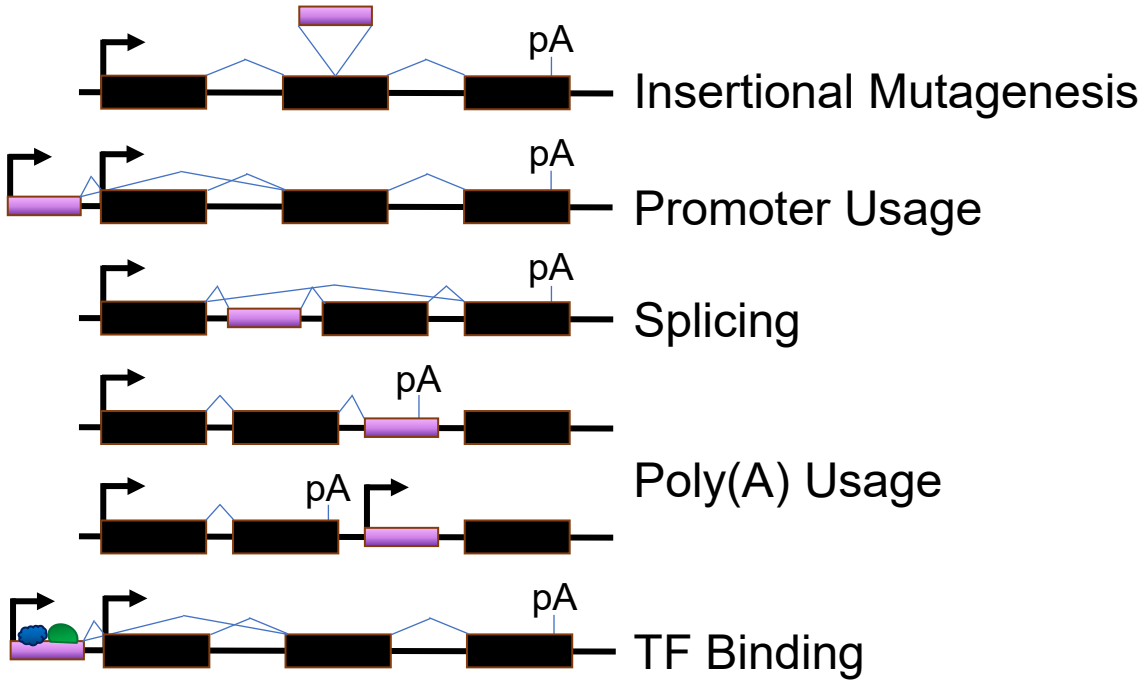
The Common Fund

- Arise through multiple mechanisms
- Affect diverse cellular pathways
- Frequency varies widely across tissues
- Occur in different genomic regions in a cell-specific manner

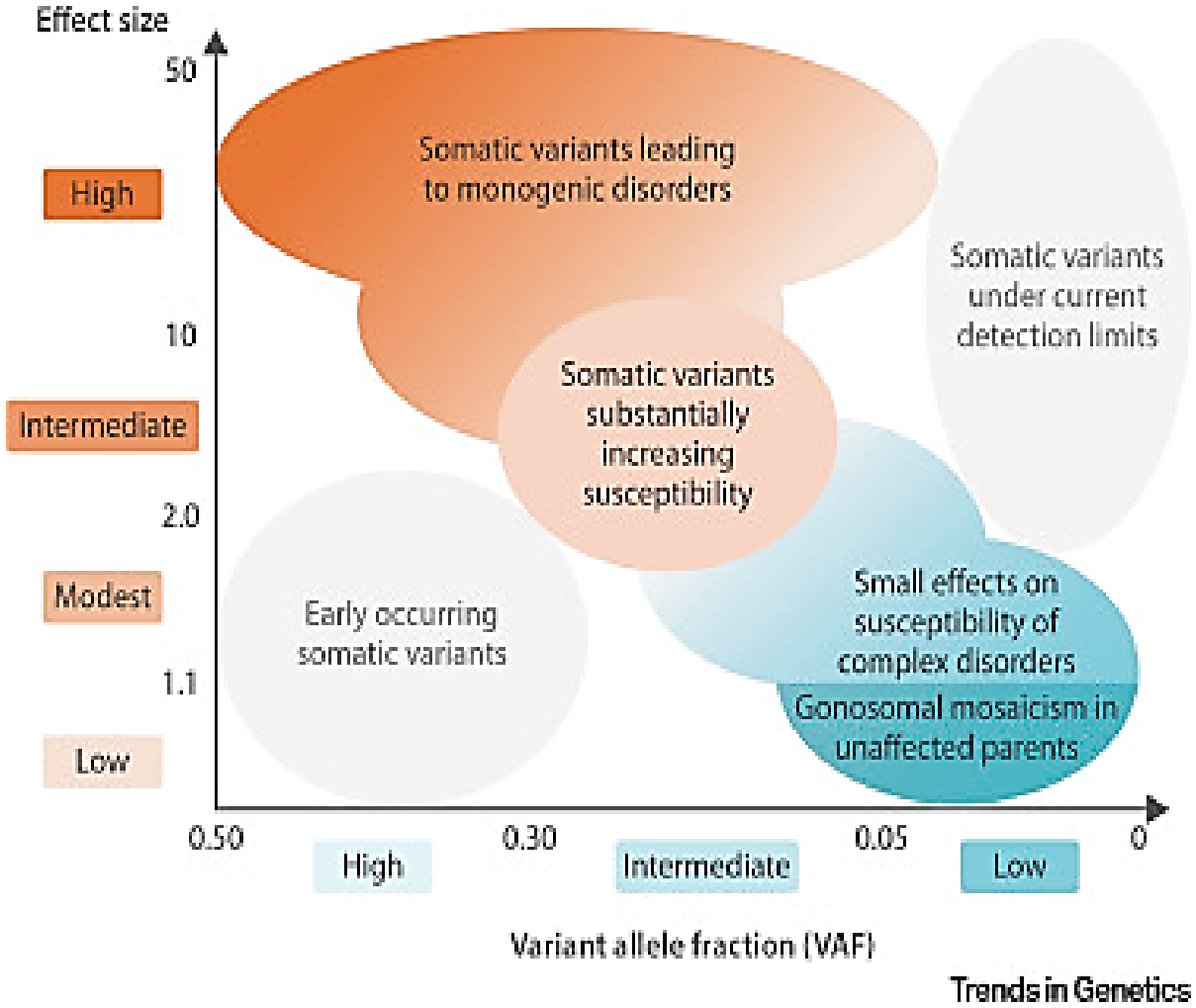


PACB, 2021

Somatic Variants – Location and Frequency



Cowley and Oakley, 2013

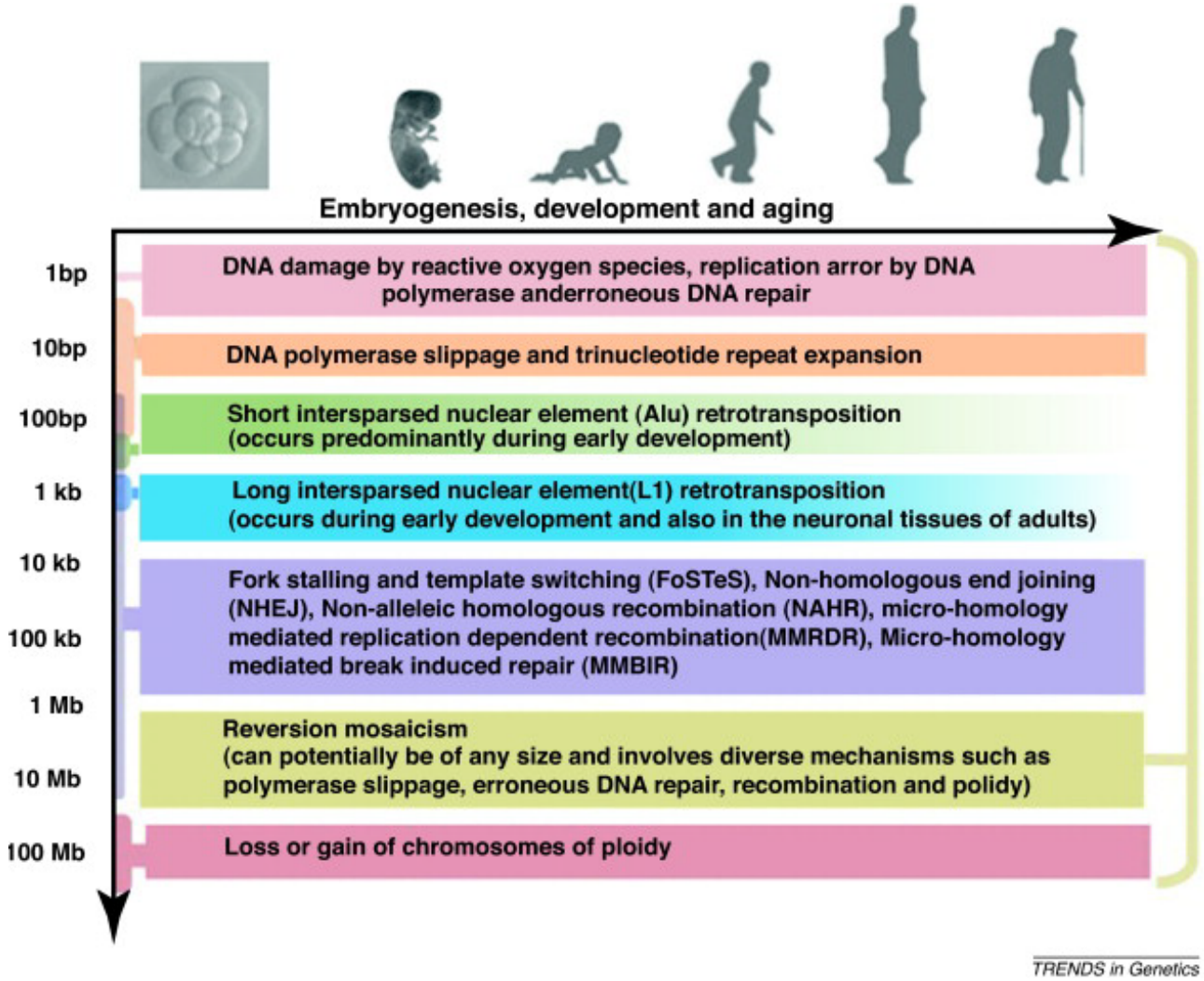


Van Horebeek et al., 2019

DNA Changes Accumulate Over the Lifespan

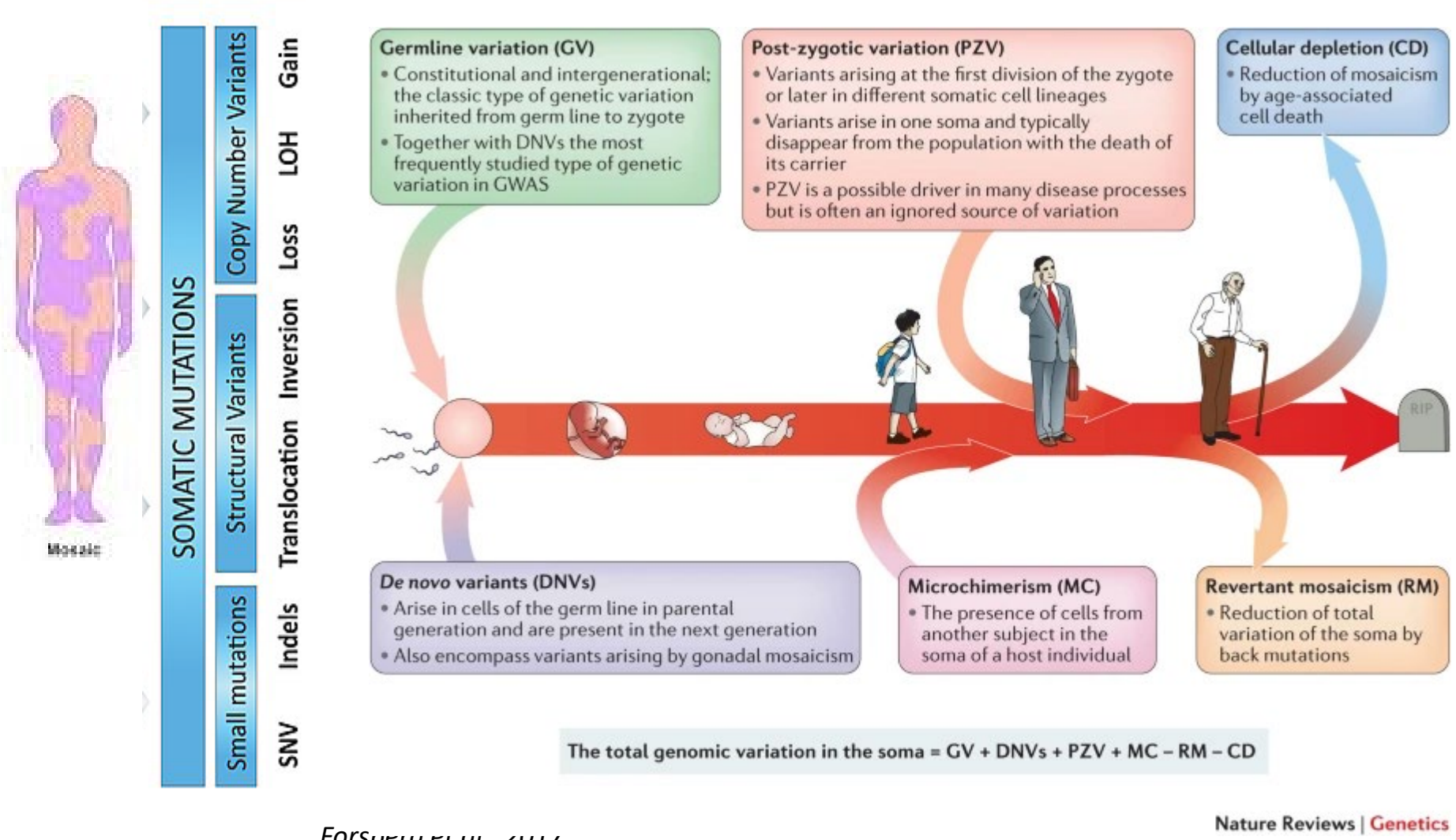
The Common Fund

- Depends on tissue, individual cell, and age: occur in different genomic regions in a cell-specific manner and frequency varies widely across tissues
- Total Somatic Variants: ~20-30k; Single Nucleotide Variants ~500 to >5000 per genome
- Different rates among tissues within an individual:
  - Colon ~ 50 SNVs/year
  - Blood ~18 SNVs/year
- Mosaic CNVs detected in 0.5% of young individuals and 2%–3% of older people (Laurie et al., 2012)
- ~40% of men older than 70 are missing the Y chromosome in a proportion of their white blood cells (UK Biobank; Thompson et al., 2019)



Dou et al., 2018

Somatic Mosaicism is part of a Larger Ecosystem of Genomic Variants



Forsberg et al., 2017

- Many sources of genomic variation
- Focus of this program will be on PZV
- Opportunity for collaboration with many programs studying human tissues – e.g. GTEx, BICCN, IGVF, HuBMAP, 4DN, HTAN etc.

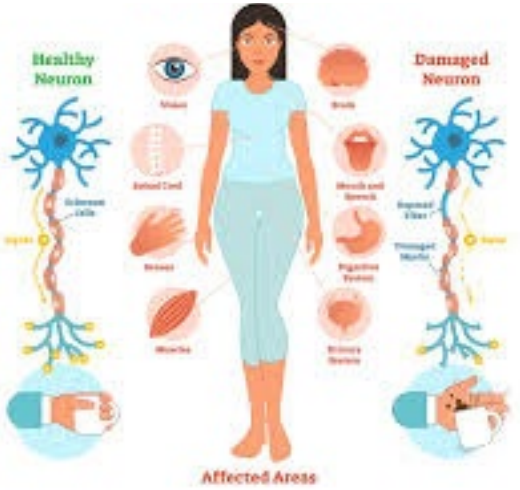
Role of Somatic Mosaicism in Disease

- Coding variants arising during development or in cancer are the most studied
- Variants in blood and brain correlate with pathogenesis
- Nascent understanding of the extent and impact of variants in most tissues across the lifespan

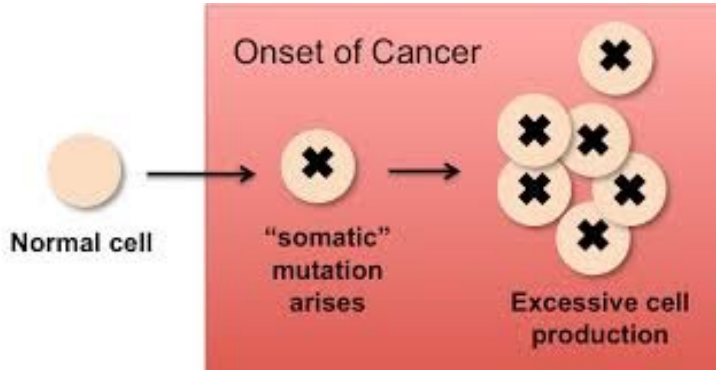
Brain

ALS, Rett, ASD, SUD, AD

Multiple Sclerosis



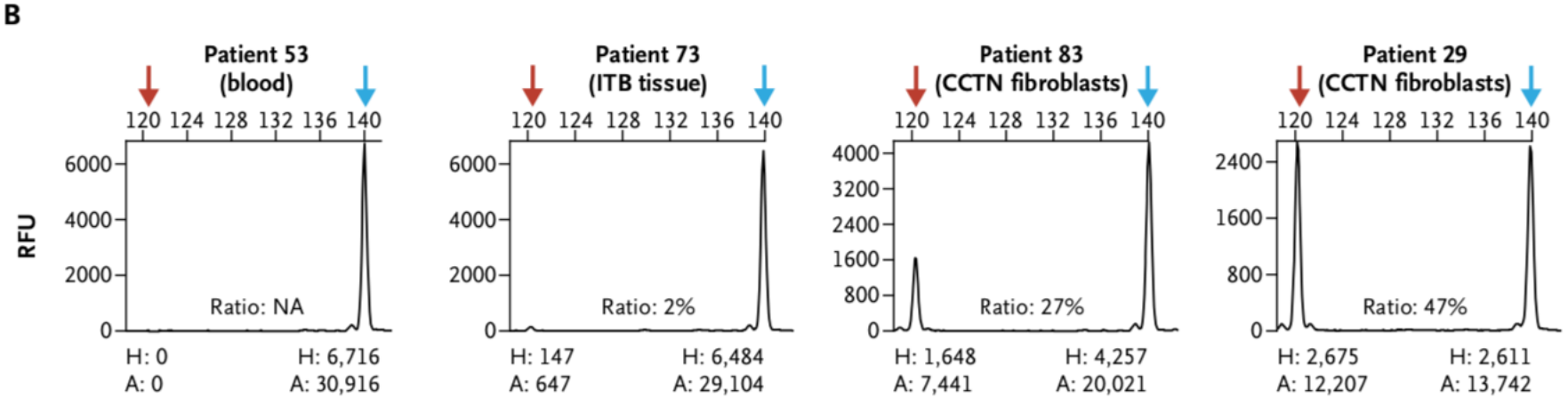
Systemic Lupus Erythematosus



Example: Activating Mutations in *AKT1* Can Cause Proteus Syndrome

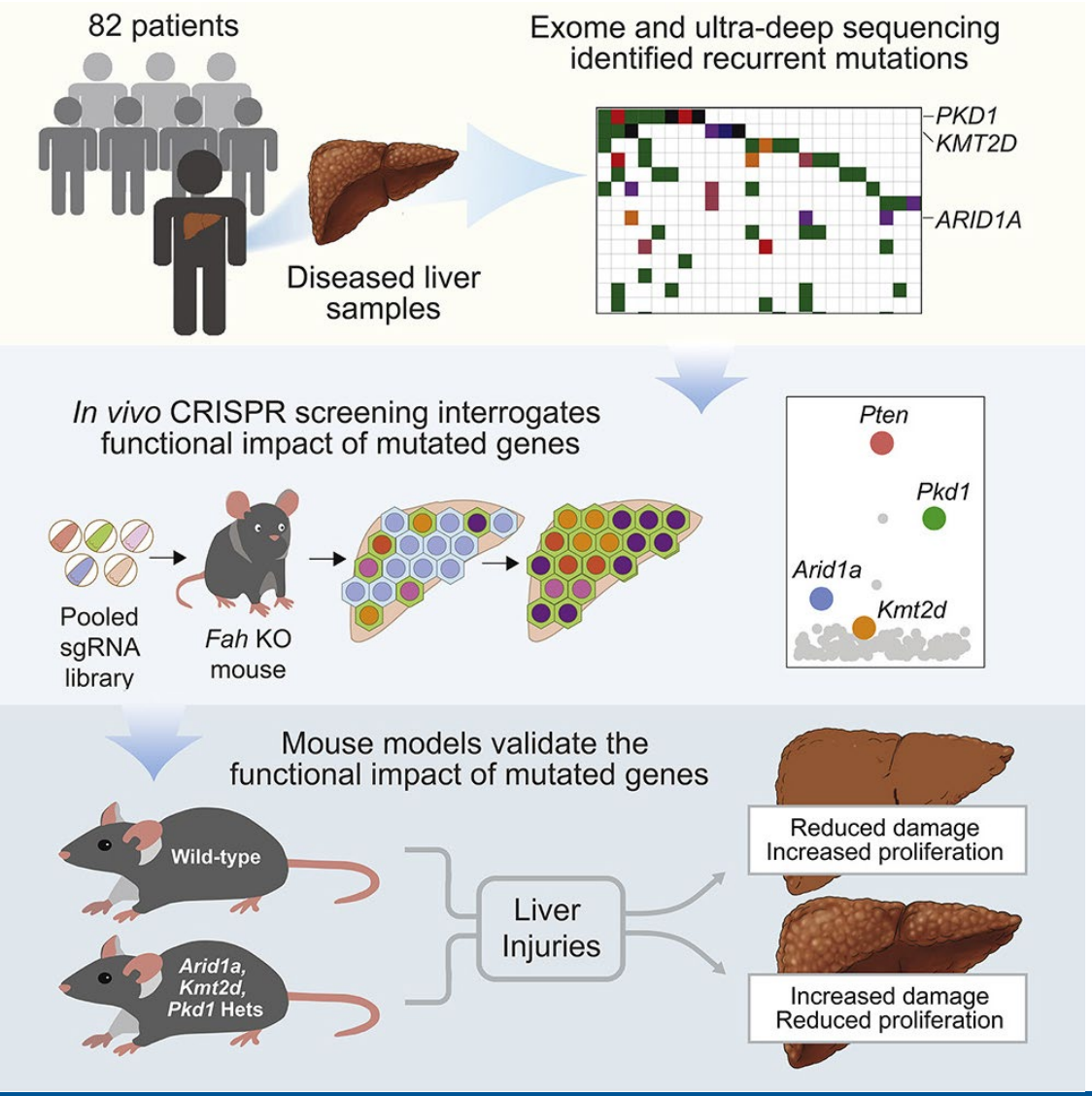
- Rare disorder characterized by segmental overgrowth and hyperplasia of multiple tissues and organs
- <1 in 1,000,000 individuals
- Observed in discordant monozygotic twins
- Occurs by somatic mutation of *AKT1*

Contribution of c.49G→A, p.Glu17Lys allele in people with Proteus Syndrome



Lindhurst et al., 2011

Modeling Liver Disease Caused by SM



- Accumulation of SM in chronic liver disease tissues
- *PKD1, PPARGC1B, KMT2D,* and *ARID1A* are recurrently mutated
- In vivo CRISPR screens validate functional relevance of *Pkd1, Kmt2d,* and *Arid1a*
- Mutations seen in liver tissues but not in cancer promote hepatocyte fitness

Zhu et al., Cell 2019

Portfolio Analysis

National Institutes of Health

Office of Strategic Coordination - The Common Fund

Portfolio Analysis - Overview

The Common Fund

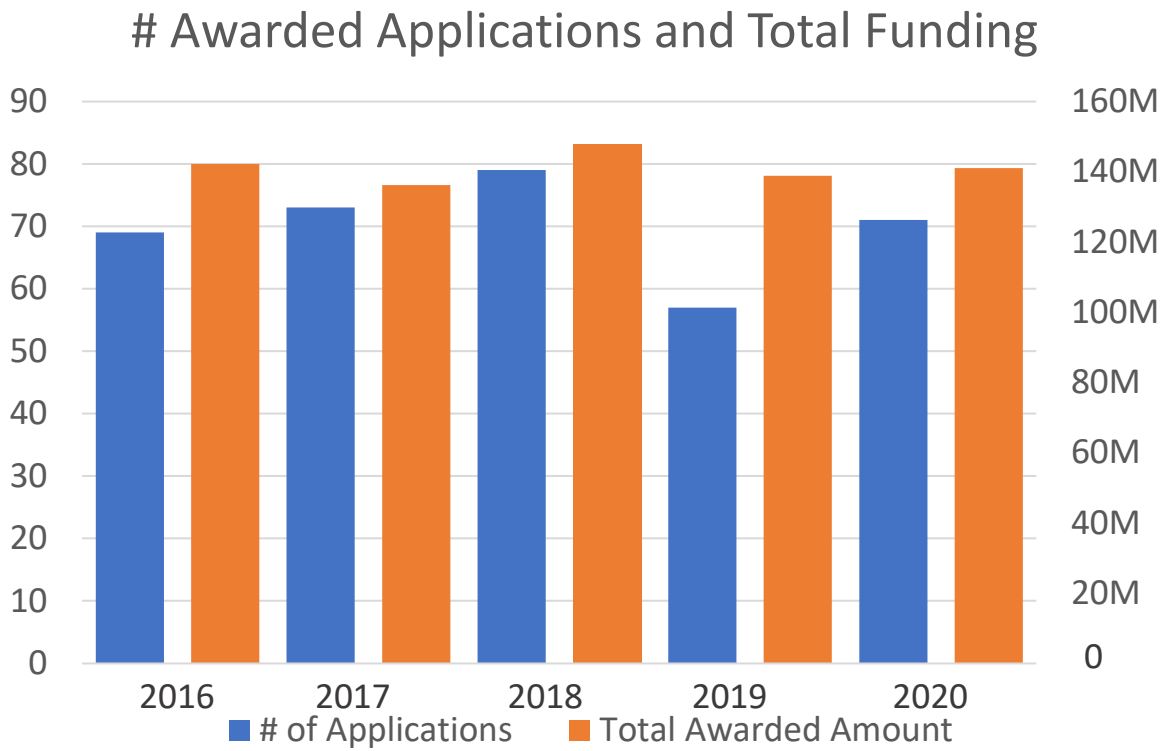
Methods

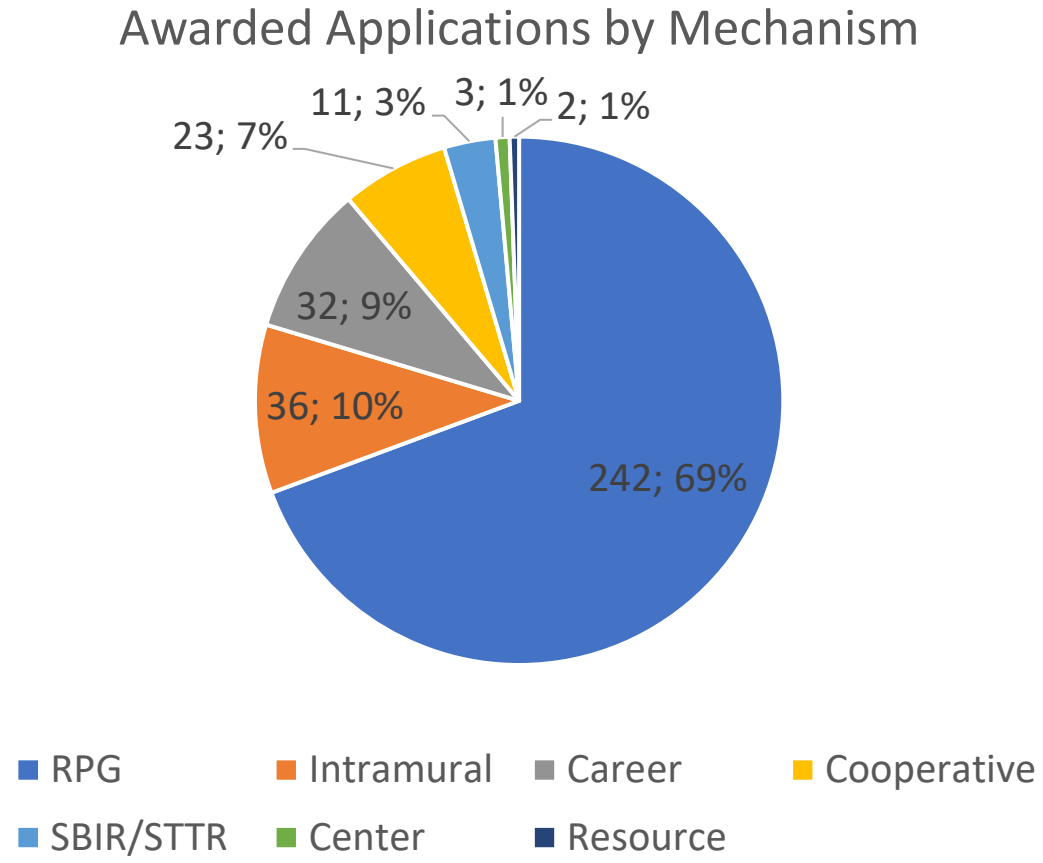
- A portfolio analysis was conducted using QVR to assess NIH support for grant applications relevant to somatic mosaicism
  - Grant applications were selected by combining three methods
    - 1) RCDC terms for genomic variation, tissue mosaicism, DNA transposable elements, transposable elements, retrotransposon
    - 2) Free text, wildcard search for somatic mosaic*
    - 3) Applications that were similar to research conducted by Dr. Peter Campbell
- Applications were excluded if focus was on non-mammalian models
- Data was collected for awarded and unawarded NIH grants for FYs 2016-2020
- The number of awards from each IC and total costs for all awards was determined
- World Report was used to analyze international awards for FYs 2016-2019
- Web of Science was used to analyze bibliometrics data for FYs 2000-2020

Summary of Analysis

- There were 1,510 applications and 349 awards for NIH between FYs 2016 and 2020, totaling over $700 million
- There was a generally consistent number of applications, awards, and total funding between FYs 2016 and 2020
- NCI had the highest number of applications, while NIGMS had the highest number of awarded applications
- R01s and R21s together represented over half (52%) of the awards
- After the US, the UK awarded the most grants

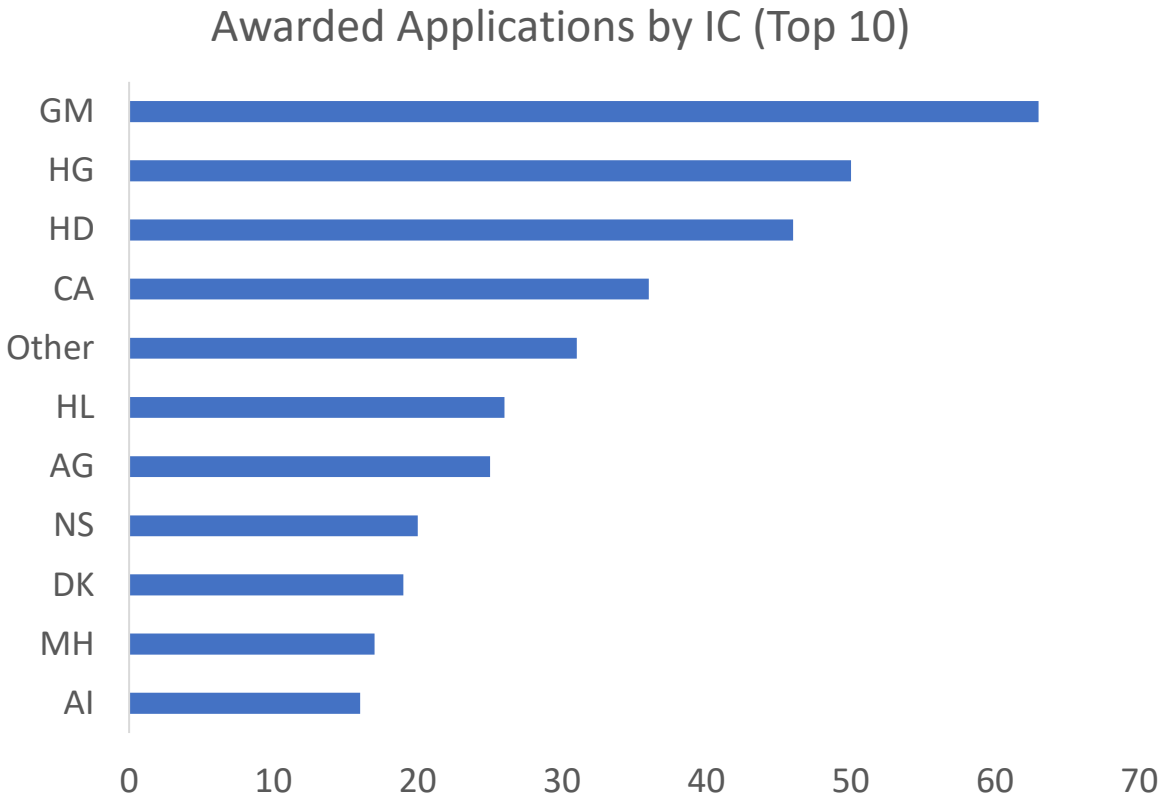
Portfolio Analysis – NIH Awards

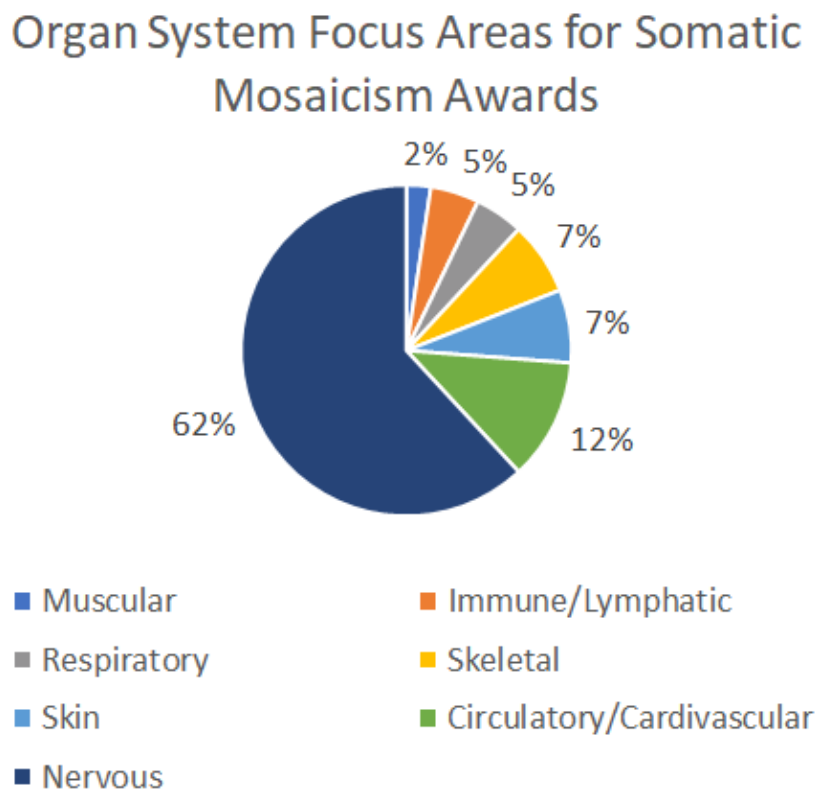




Portfolio Analysis – NIH Awards

The Common Fund



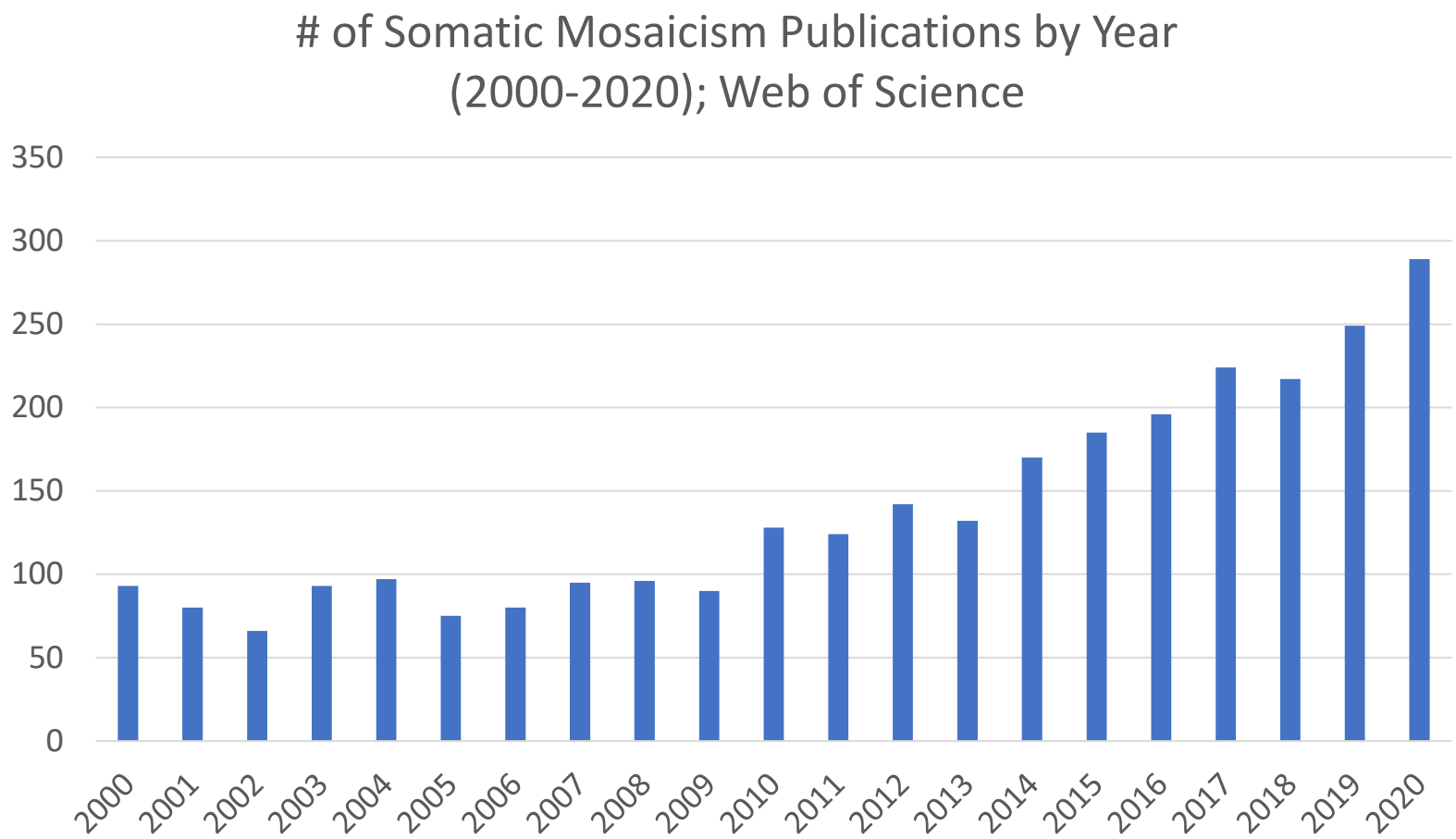


In preliminary analyses, the majority of somatic mosaicism awards (62%) focus on diseases of the nervous system, including schizophrenia and Alzheimer's.

Note: awards focusing on somatic mutations in cancer may not be retrieved from the "somatic mosaicism" key word search used for this analysis

Portfolio Analysis – Publication Trends

The Common Fund



Analysis of Related Studies

National Institutes of Health

Office of Strategic Coordination - The Common Fund

Challenges in Identifying Somatic Mosaicism and Structural Variation from Selected Studies

The Common Fund

- Two sets of studies were analyzed to identify challenges related to
  - detecting structural variation in human genomes (Set 1)
  - characterizing somatic mosaicism/mutations in human tissues (Set 2)
- A PubMed keyword search revealed over 500 studies of potential relevance for Set 1 an over 1,000 studies for Set 2
- A select group of studies was analyzed to represent the relevant implications and challenges for each set of studies

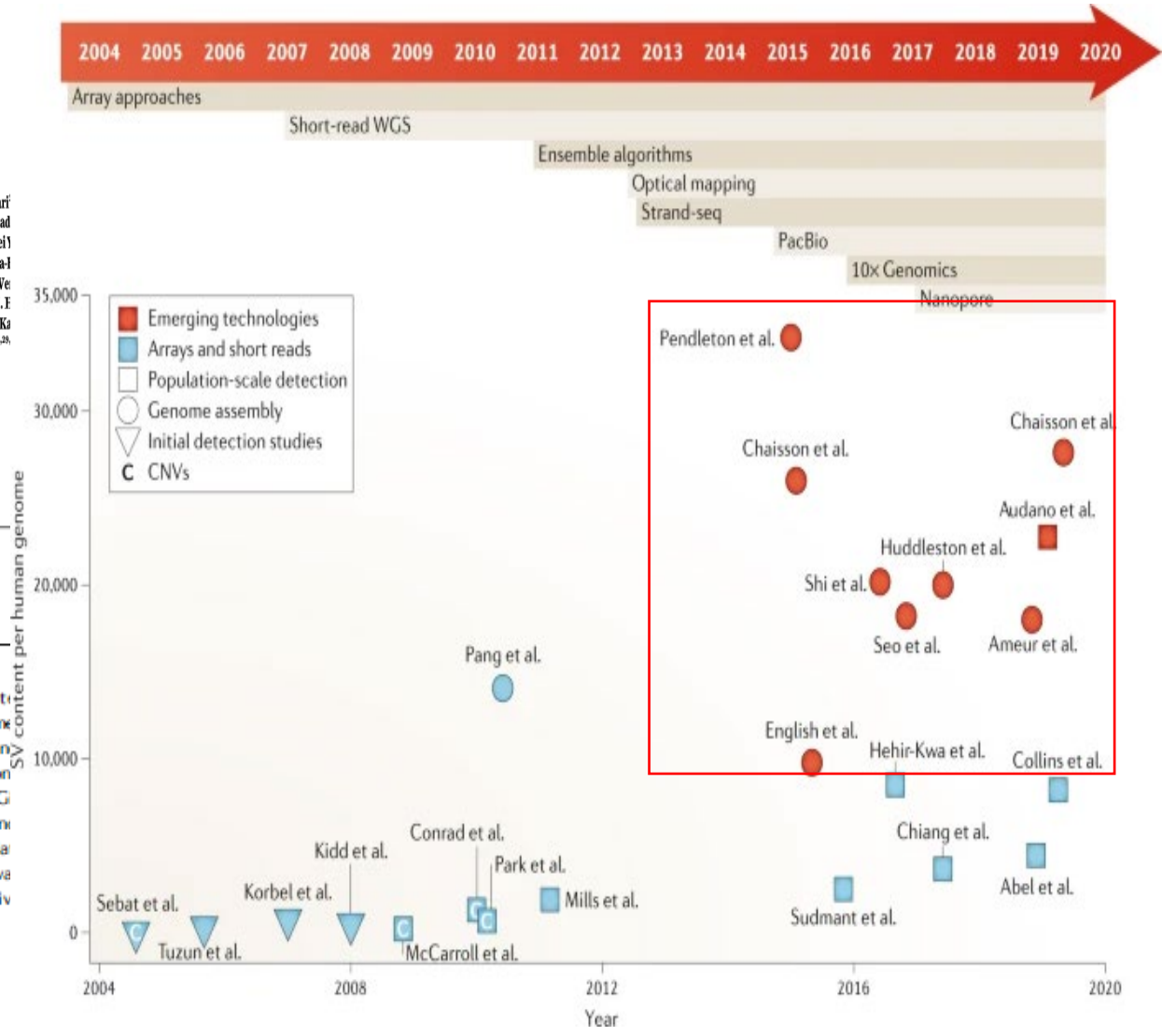
Haplotype-resolved diverse human genomes and integrated analysis of structural variation

Peter Ebert[1*], Peter A. Audano[2*], Qihui Zhu[2*], Bernardo Rodriguez-Martin[1*], David Porubsky[2], Marc Jan Bonder[4,5], Arvis Sulovari[6], Jana Ebler[1], Weichen Zhou[6], Rebecca Serra Mari[1], Feyza Yilmaz[2], Xuefang Zhao[7,8], PingHsun Hsieh[2], Joyce Lee[9], Sushant Kumar[10], Jiadong Lin[1], Tobias Rausch[1], Yu Chen[11], Jingwen Ren[12], Martin Santamarina[14,15], Wolfram Höps[1], Hufsah Ashraf[1], Nelson T. Chuang[16], Xiaofei Yang[17], Katherine M. Munson[2], Alexandra P. Lewis[2], Susan Fairley[18], Luke J. Tallon[19], Wayne E. Clarke[19], Anna O. Basile[19], Marta Byrska-Bishop[20], André Corvelo[19], Uday S. Evani[19], Tsung-Yu Lu[12], Mark J.P. Chaisson[13], Junjie Chen[20], Chong Li[20], Harrison Brand[7,8], Aaron M. Wenger[21], Maryam Ghareghani[22,23,1], William T. Harvey[2], Benjamin Raeder[4], Patrick Hasenfeld[4], Allison A. Regier[24], Haley J. Abel[24], Ira M. Hall[24], Paul Flicek[18], Oliver Stegle[4,5], Mark B. Gerstein[10], Jose M.C. Tubio[14,15], Zepeng Mu[26], Yang I. Li[27], Xinghua Shi[20], Alex R. Hastie[9], Karyn Meltz Steinberg[9], Zechen Chong[12], Ashley D. Sanders[4], Michael C. Zody[19], Michael E. Talkowski[7,8], Ryan E. Mills[6,28], Scott E. Devine[16], Charles Lee[3,29], Jan O. Korbel[1,30†‡], Tobias Marschall[1†‡], Evan E. Eichler[2,31†‡]

Structural variation in the sequencing era

Steve S. Ho[1], Alexander E. Urban[2,3] and Ryan E. Mills[1,4*]

Abstract | Identifying structural variation (SV) is essential for genome int… been historically difficult due to limitations inherent to available genom… Detection methods that use ensemble algorithms and emerging sequen… have enabled the discovery of thousands of SVs, uncovering information… relationship to disease and possible effects on biological mechanisms. G… in SV type and size, along with unique detection biases of emerging gen… multiplatform discovery is necessary to resolve the full spectrum of varia… modern approaches for investigating SVs and proffer that, moving forwa… biological information with detection will be necessary to comprehensiv… impact of SV in the human genome.



Measures of SVs from Selected Studies

The Common Fund

Study	Pendleton 2015	English 2015	Chaisson 2015	Huddleston 2017	Shi 2016	Seo 2016	Ameur 2018	Chaisson 2019	Audano 2019	Ebert 2021
# SVs identified	23,180	9,777	26,079	20,470-20,602	20,175	18,210	17,687-17,936	27,622	22,755	24,653
Samples	NA12878 cell line	1 genome	1 hydatidiform mole	2 hydatidiform mole	1 sample (Chinese); blood	1 sample (Korean); LCL	2 samples (Swedish); blood	3 diverse samples	15 diverse samples	32 diverse samples
Approach	All the studies employed long-read and short-read sequencing of sample genomes and alignment/comparison with a human reference genome to identify SVs									
Other Approaches		Combined paired-end and aCGH data with long-read, long-insert, and whole-genome architecture data	Comparison to BAC and fosmid clones, Sanger-based BAC-end sequence	SMRT-SV; tile across euchromatic genome in 60-kbp windows; validation with clones, BACs, Sanger sequencing	Long-read RNA sequencing (Iso-Seq)	SMRT sequencing microfluidics-based linked reads, and BAC sequencing approaches		Long-read, short-read, linked-reads, strand-specific sequencing technologies; variant discovery algorithms	SMRT-SV and Sequel sequencing platforms (STAR Methods)	Continuous long-read or high-fidelity sequencing; Strand-seq; graph-based genotyping; QTL
Optical Maps	Yes				Yes	Yes	Yes	Yes		Yes
Coverage	~22x-80x	~90x	41x	62.4x-66.3x	103x	101x	~78x	39.6	40x-98x	20x-40x
Results	• In three studies, >80% of identified SVs were previously unreported • Most novel sequences were between 100 bp and 5kb (there was a 5-fold increase in discovery of SVs <1kb) • SVs involving transposable elements and regions rich with repeats (simple repeats, long tandem repeats, high GC content) were resolved • Of identified SVs, most were insertions (46-64%) or deletions (36-53%), fewer complex variants (4%) or inversions (0.2-2.8%) • Mean length for deletions (442-460 bp), insertions (435-477 bp), inversions (6,087-6,449 bp) • Samples from African individuals contained more SVs found in a single sample than in non-African samples									
Implications	• Combination of long-read and short-read approaches yielded more SVs than short-read approaches alone • Multiple SV detection algorithm use and validation with targeted sequencing increased sensitivity of SV calls									
Gaps/ Challenges	• Difficult to resolve: segmental duplication, CNV in highly duplicated regions, inversions > 20 kbp, regions with long repeats, centromeric and acrocentric regions • Scale of long-read sequencing is limited to tens of kilobases • Small insertion and deletion (1-2 bp) errors with long-read approach									

Selected Studies of SM in Human Tissues

The Common Fund

Study	Martincorena 2015	Lee-Six 2018	Martincorena 2018	Brunner 2019	Yoshida 2020
Tissue Type	Normal skin (eyelid epidermis)	Normal colorectal epithelial cells	Normal esophageal epithelium	Normal and cirrhotic liver	Bronchial epithelium
# of patients and samples	234 samples from 4 individuals	2,035 colorectal crypts from 42 individuals	844 samples from 9 deceased organ donors	482 dissections from 14 individuals	632 colonies from 14 individuals
Patient age	55 to 73 years	11 to 78 years	20 to 75 years	49 to 77 years	11 months to 81 years
Methods	Small biopsies and algorithms to detect mutations in a small fraction (as few as 1%) of cells; sequenced 74 cancer genes	Laser-capture microdissection to isolate colorectal crypts and WGS; estimated contribution of mutational signatures to burden	Ultradeep targeted sequencing of small samples; WGS of 21 samples with large clones to assess SV	WGS of laser-capture microdissections of hepatocytes; targeted deep sequencing of cancer genes	WGS of colonies derived from single epithelial cells
Coverage	500x (targeted seq)	15x (WGS)	870x (targeted seq); 37x (WGS)	30-70x per sample (WGS)	16x (WGS)
Results	18-32% of skin cells had positively selected driver mutations; 3,760 somatic mutations identified across 234 biopsies; many mutations found in 1 to 2% of cells, some mutations found in most of the cells; 2 – 6 somatic mutations/Mb/cell	1% of normal colorectal epithelial cells had driver mutations; significant variation in mutation burdens between crypts: burden in older individuals ranged from ~1,500 to ~15,000; ~1/2 of mutational signatures were ubiquitous, some correlated with age	Median allele frequency of mutations – 1.6%; 8,919 somatic mutations across 844 samples; several hundred mutations per cell for individuals in 20s to >2,000 later in life	<5% of clones had driver mutations/SVs in non-malignant liver; mutation accumulation rate 33/year; 13/year variation between individuals; some mutational signatures ubiquitous; substantial intra-individual variation	4-14% of cells in NS had driver mutations; Substitutions increased by age: 22/cell/year (2,330/5,300 per cell for ex/CS. Intra-individual variation: 140 per cell in children, 290 adult NS, 2,100 CS; inter-individual: ~1,200/cell SD for ex/CS, 90 NS
CNV/SV	One gene most frequently displayed CNV; ability to detect CNVs variable	18% of crypts had CNVs/SVs; SVs: 48 deletions, 18 tandem duplications, 4 translocations	CNV detected, particularly for the NOTCH1 gene	SVs and CNVs moderate in patients with cirrhosis, rare in normal liver	Normal bronchial epithelial cells had few CNVs or SVs

Gaps/Challenges	• Detecting and assessing inter-individual differences in mutation landscape
	  • What is contribution of environmental exposure versus genetic background to inter-individual variation?
	  • Studies would benefit from ancestral diversity of participants
	• Determining contribution of heterogeneity in mutational burden among competing cells to clonal evolution/disease development
	• Detecting and characterizing intermediate stages of disease progression				

Opportunities for Collaboration

National Institutes of Health

Office of Strategic Coordination - The Common Fund

Related Programs Studying Human Tissues

The Common Fund

ENCODE

PsychENCODE

COSMIC
Catalogue Of Somatic Mutations In Cance

4D Nucleome

GTEx

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM

AnVIL

BRAIN
The BRAIN Initiative seeks to deepen understanding of the inner workings of the human mind and to improve how we treat, prevent, and cure disorders of the brain.

IHEC
International Human Epigenome Consortiur