NEW YORK GENOME CENTER®

Collaboration
Research
Discovery

Computational Biology at the New York Genome Center

November 9th, 2022

Michael C. Zody, Ph.D.
Scientific Director, Computational Biology,
New York Genome Center

The Formation of the New York Genome Center

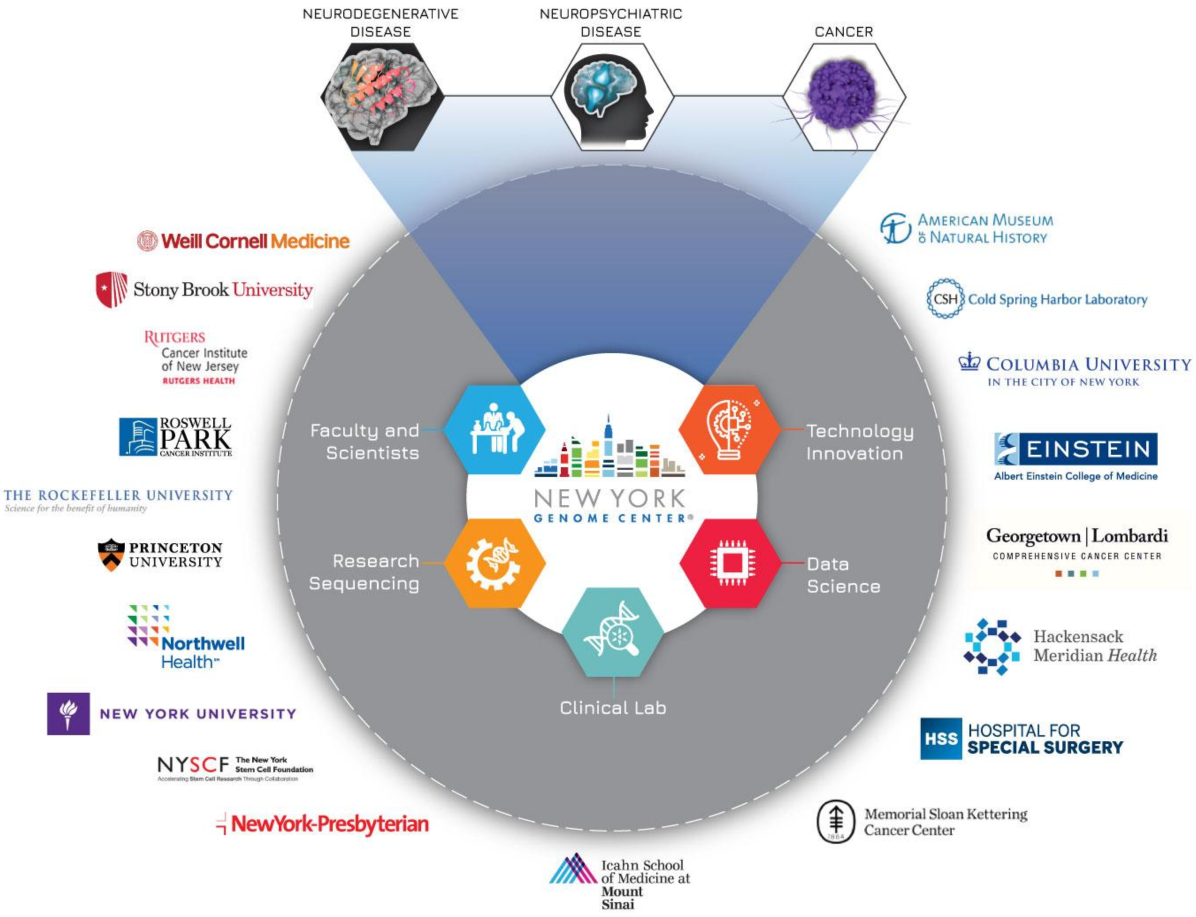
- Founded in 2011 to provide critical infrastructure and expertise in genomic research

- Located at 101 6th Ave. (SoHo)

- Partnership of academic researchers and civic minded philanthropists

- Serve as the convening nexus for collaborative genomic research

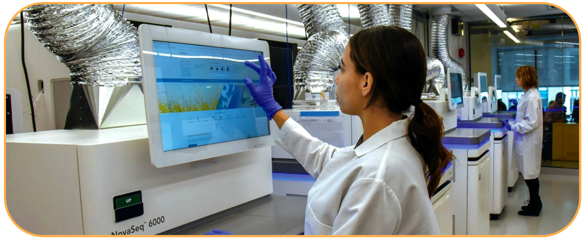
- Work to establish New York City as a biotech hub


NEURODEGENERATIVE DISEASE
NEUROPSYCHIATRIC DISEASE
CANCER
NEW YORK GENOME CENTER®
Faculty and Scientists
Technology Innovation
Research Sequencing
Data Science
Clinical Lab
Weill Cornell Medicine
Stony Brook University
RUTGERS Cancer Institute of New Jersey RUTGERS HEALTH
ROSWELL PARK CANCER INSTITUTE
THE ROCKEFELLER UNIVERSITY
Science for the benefit of humanity
PRINCETON UNIVERSITY
Northwell Health™
NEW YORK UNIVERSITY
NYSCF The New York Stem Cell Foundation
Accelerating Stem Cell Research Through Collaboration
NewYork-Presbyterian
Icahn School of Medicine at Mount Sinai
AMERICAN MUSEUM of NATURAL HISTORY
CSH Cold Spring Harbor Laboratory
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
EINSTEIN
Albert Einstein College of Medicine
Georgetown | Lombardi
COMPREHENSIVE CANCER CENTER
Hackensack Meridian Health
HSS HOSPITAL FOR SPECIAL SURGERY
Memorial Sloan Kettering Cancer Center
1884

Innovation & Technology Development at NYGC

- Fully Automated Production Sequencing Capacity
  - 5 NovaSeq 6000, 3 Illumina HiSeq X Ten, 2 Illumina HiSeq 2500 sequencers
  - 50,000 whole genomes per year
- Long read sequencing - Oxford Nanopore Technologies PromethION
- Low cost sequencing options - evaluating cost and quality for key applications:
  - Single cell and single nuclei genomics
  - Cell-free whole genome sequencing
  - FFPE tumor sequencing
  - Clinical WGS/WGTS and Precision Genomics Initiatives
- Single-cell genomics
  - scWGS (DLP+)
  - multimodal scRNA (CITE-Seq, etc)
- Spatial Transcriptomics

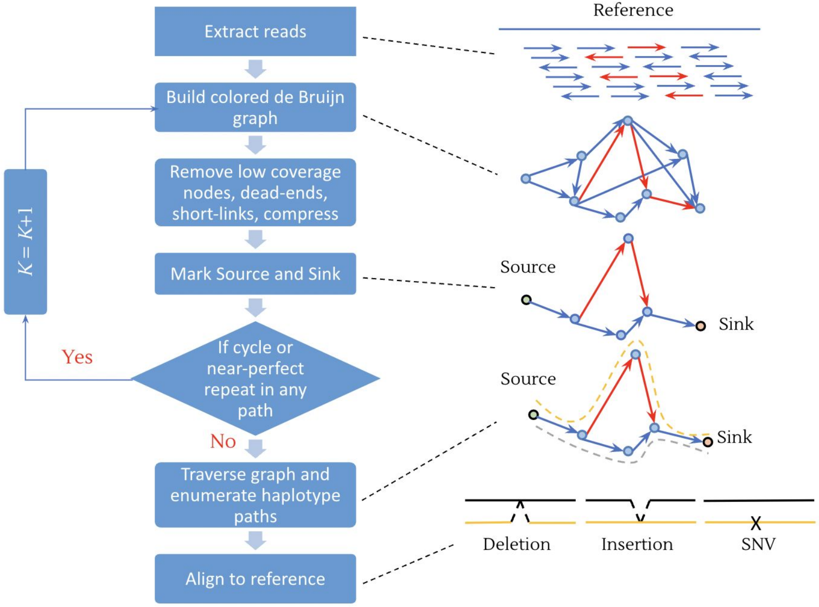
ULTIMA GENOMICS



Outline

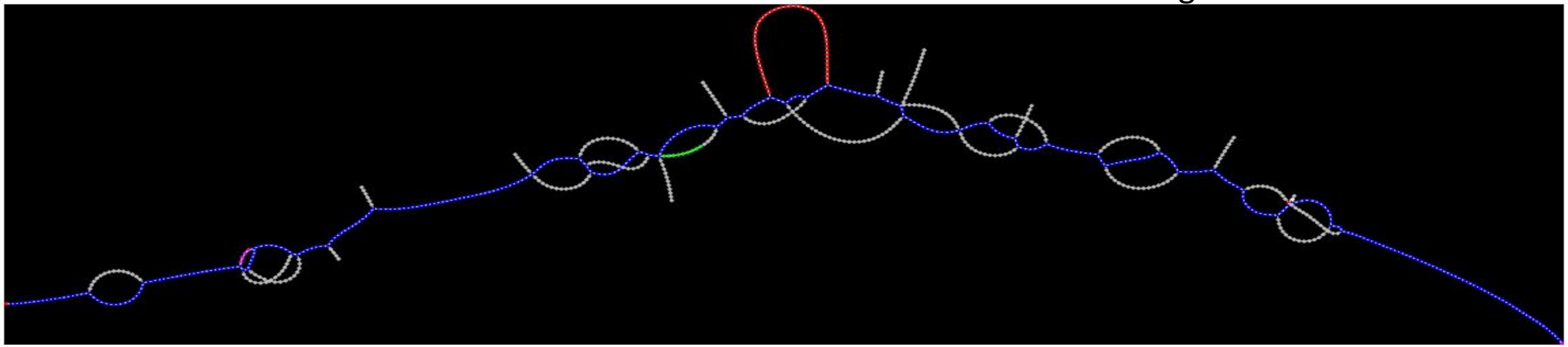
- Lancet cancer variant calling
- Polyethnic-1000 cancer project
- Absinthe insertion caller
- 1000 Genomes Project deep whole genome sequencing
- Structural Variant imputation

Lancet: somatic variant calling using colored de Bruijn graphs



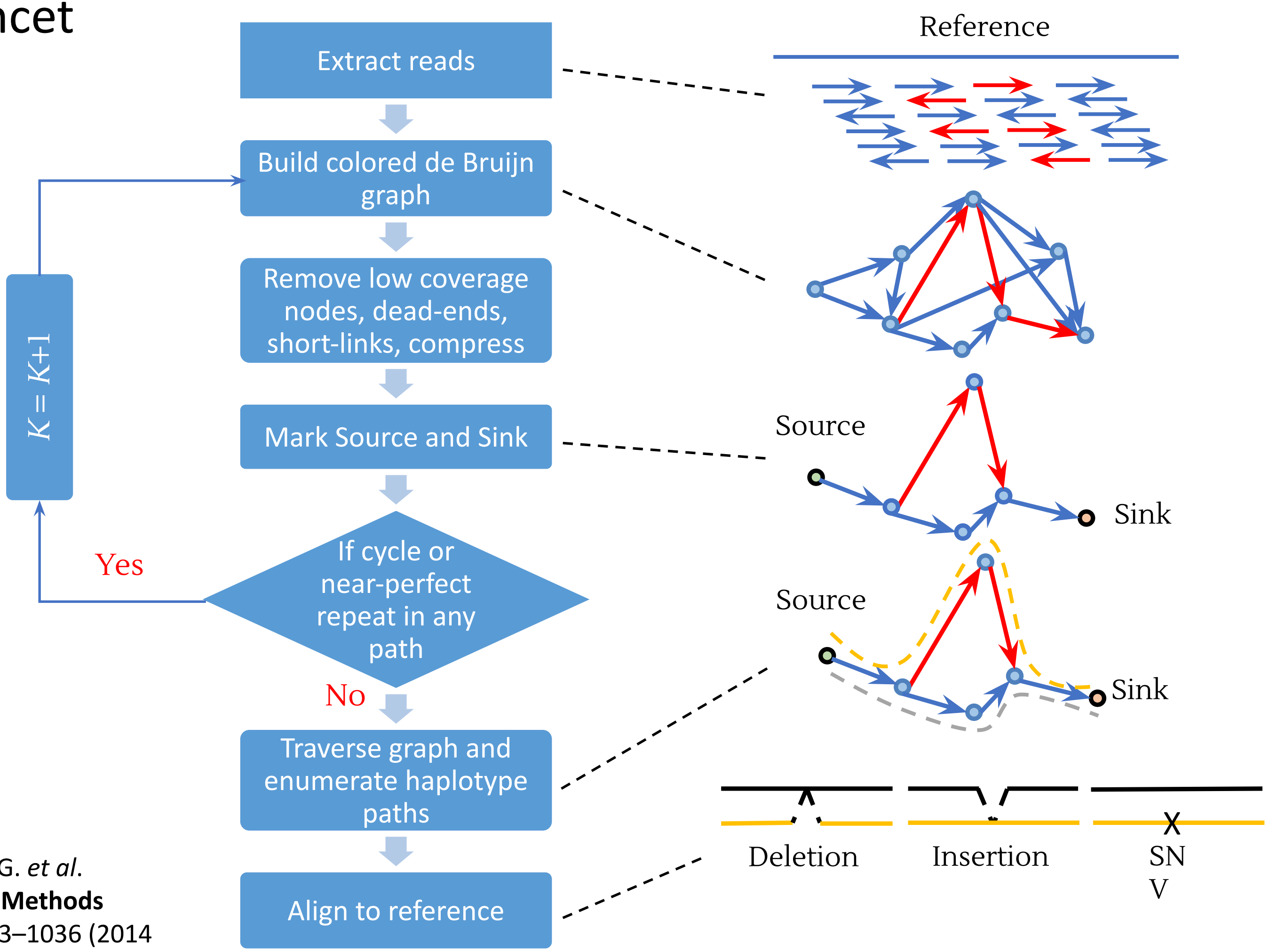
Extract reads
Build colored de Bruijn graph
Remove low coverage nodes, dead-ends, short-links, compress
Mark Source and Sink
If cycle or near-perfect repeat in any path
Yes
K = K+1
No
Traverse graph and enumerate haplotype paths
Align to reference
Reference
Source
Sink
Source
Sink
Deletion
Insertion
SNV

- *Joint assembly* of tumor and normal data
- *Reduced reference bias*: in regions of genomes that substantially differ from the reference sequence.
- *Increased power* to discover shared/private events across tumor and matched normal samples
- More accurate variant *allele fraction estimates*, critical to understanding sub-clonal structures.



red = tumor, green = normal, blue = shared, grey = low coverage & sequencing errors

Lancet



Extract reads

Build colored de Bruijn graph

Remove low coverage nodes, dead-ends, short-links, compress

Mark Source and Sink

If cycle or near-perfect repeat in any path

Traverse graph and enumerate haplotype paths

Align to reference

Colored DeBruijn graph augmented with:

- tumor/normal coverage by strand
- bp phred quality

Graph rendering and visualization

Network flow algorithm for graph exploration

On the fly STR analysis

Strongly connected components analysis

Active regions

Narzisi G. *et al.*
Nature Methods
11, 1033–1036 (2014
)

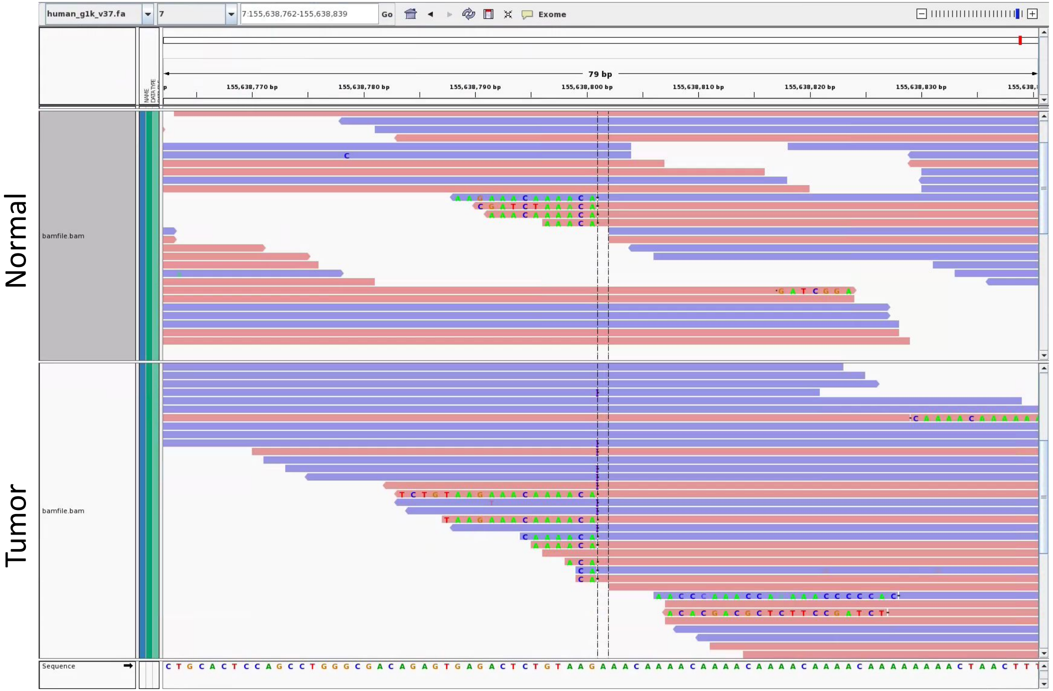
Example of variant with partial support

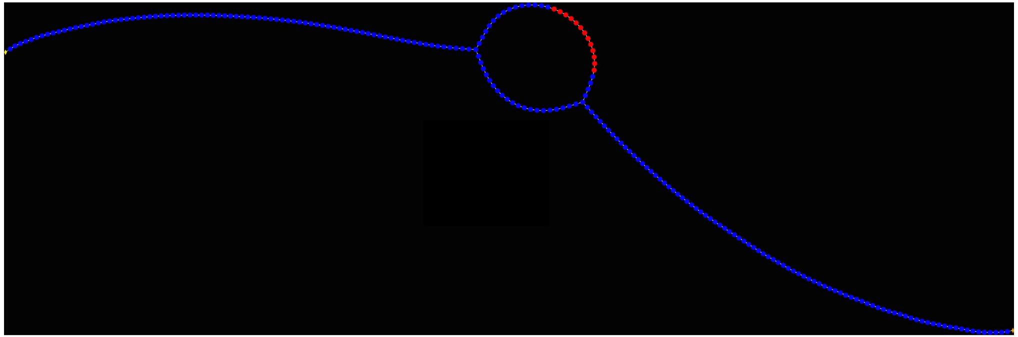
- Insertion is clearly present in the tumor, but it is partially supported in the normal

- Low support in normal (soft-clipping) + low complexity

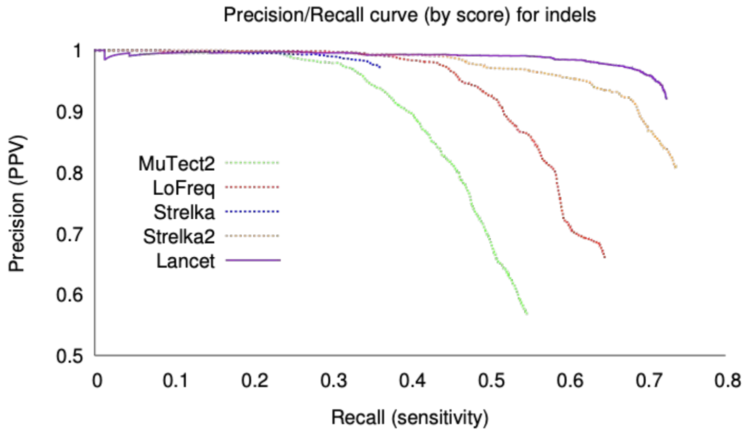
- The colored DeBruijn graph of the (tumor+normal) reads correctly characterizes the mutation

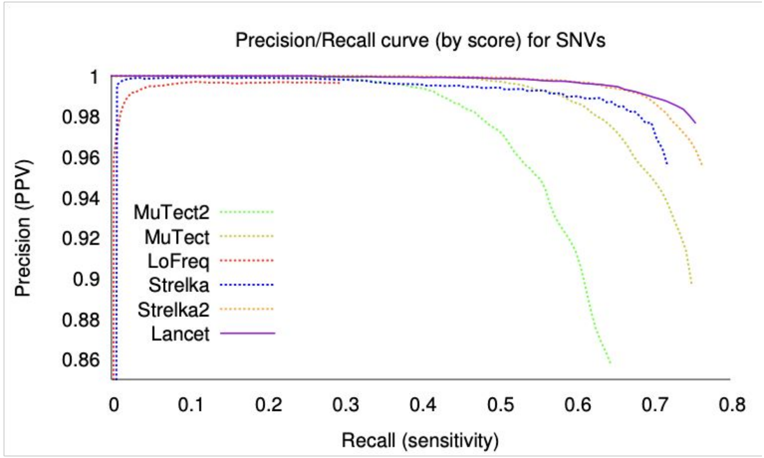
- More accurate estimation of variant allele fraction

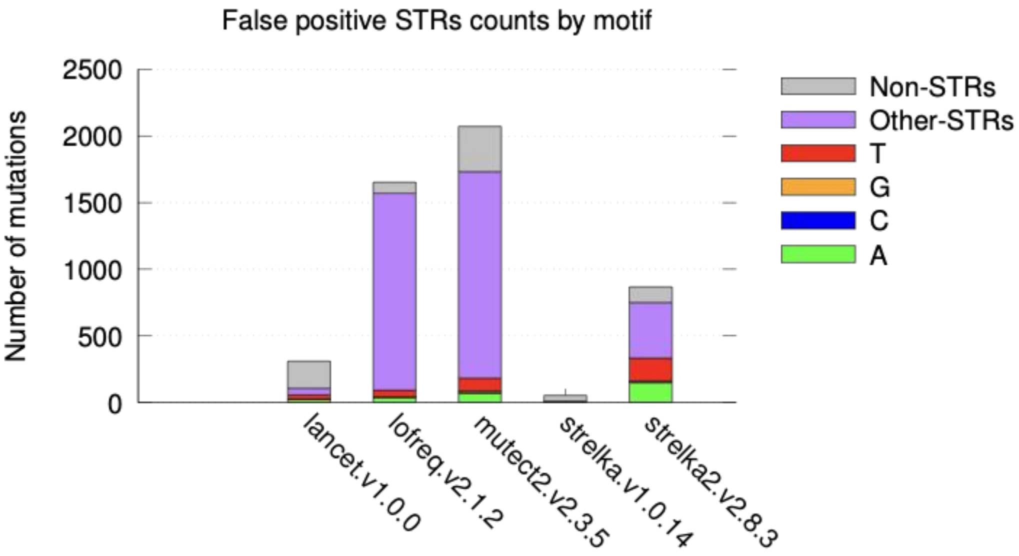




Somatic mutations performance comparison







Narzisi G, et al. *Genome-wide somatic variant calling using localized colored de Bruijn graphs.* **Communications Biology**, volume 1, Article number: 20 (2018)

Lancet2 - refactored code for speed

https://github.com/nygenome/Lancet2

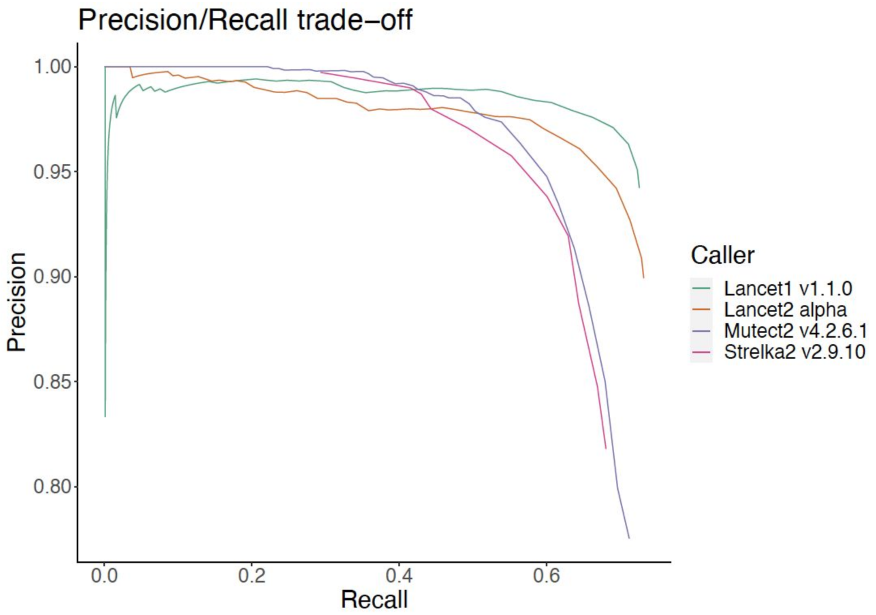
Rajeeva Musunuri
Bioinformatics
Data Scientist

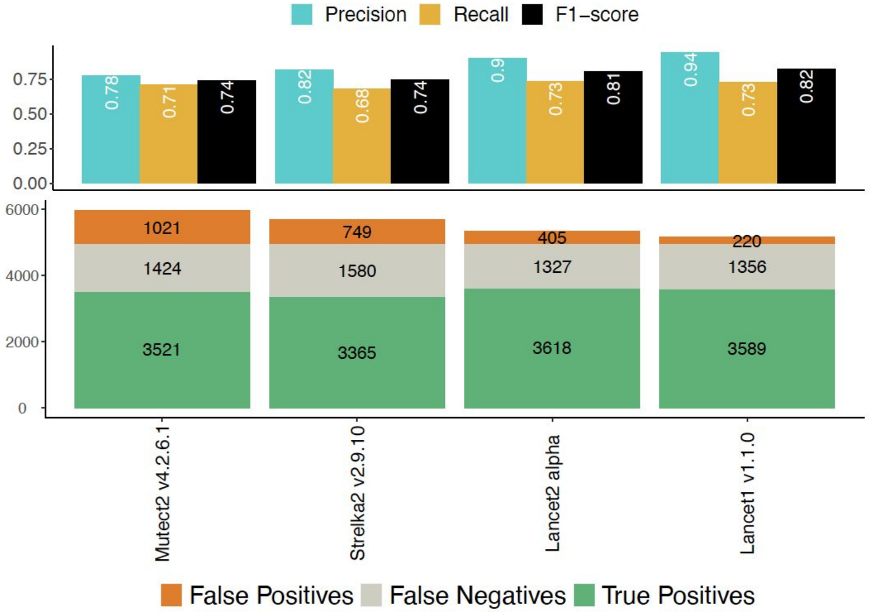
Genome-wide computational performance on the Virtual Tumor.

WGS	**Total Runtime (core hrs)**	**Max Memory Utilized (GB)**	**Avg CPU Utilization (%)**
Lancet v1.1.0	2902.7	38.69	94.1
Lancet2 alpha	**728.4**	**5.1**	**99.7**
Mutect2 v4.2.6.1	954.4	12.7	11.6
Strelka2 v2.9.10	81.4	3.2	45.8

1. Re-factor the source code using modern **C++17 features** for modularity and maintainability
2. Store the graph using a fast hash table (**Abseil's Swiss table**) to improve graph traversal performance
3. Efficient pull-based reactive multi-threading strategy for local assembly of windows using a **lock-free concurrent queue**
4. **Developer tool kit and APIs** to facilitate new feature development and integration with other bioinformatics tools.

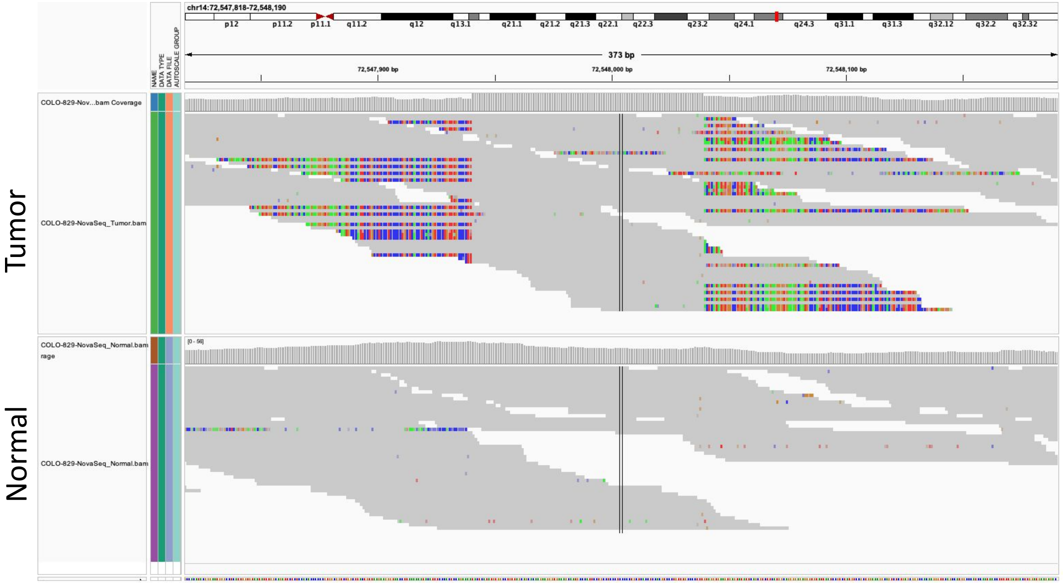
Somatic indel mutations performance comparison



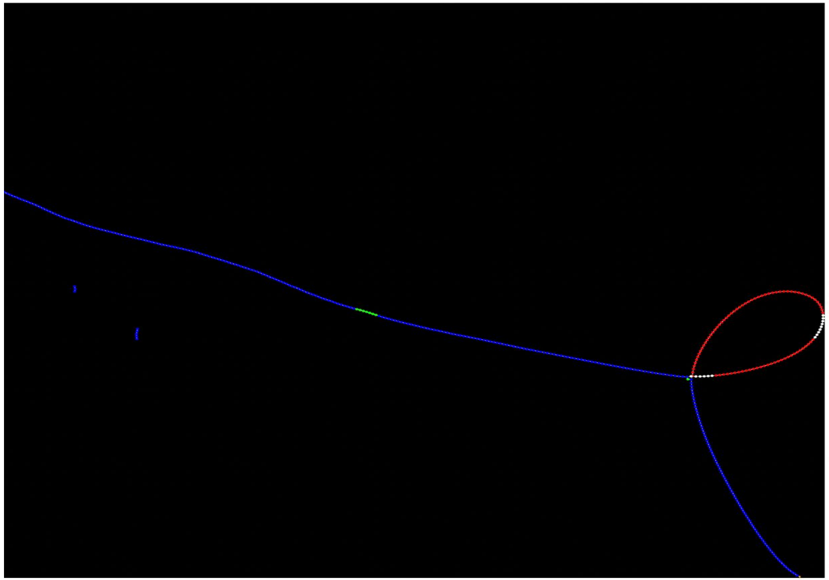


	Precision	Recall	F1−score	True Positives	False Positives	False Negatives
Lancet1 v1.1.0	0.94	0.73	0.82	3589	220	1356
Lancet2 alpha	0.9	0.73	0.81	3618	405	1327
Strelka2 v2.9.10	0.82	0.68	0.74	3365	749	1580
Mutect2 v4.2.6.1	0.78	0.71	0.74	3521	1021	1424

98 bp insertion in chr14 of COLO829 cancer cell-line



Clear pattern of soft-clipped sequences in the tumor reads indicating the challenge to map the reads to the reference.



Lancet colored de Bruijn graph for the same 98bp insertion in COLO829 (red = tumor; green = normal; blue = shared; white = sequencing errors).

Github repository

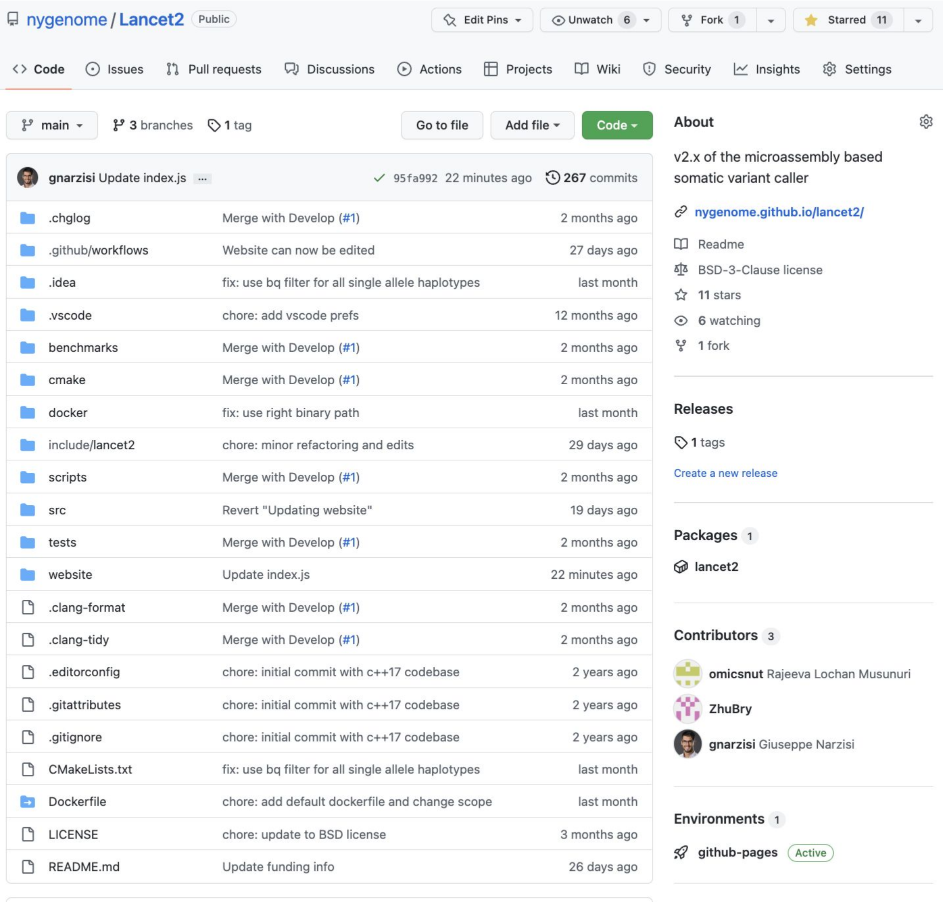
- Source code freely available (BSD-3-Clause) via NYGC github: https://github.com/nygenome/Lancet2

- 100% C/C++ code with native multi-threading parallelization.

- Interactive user interface similar to other bioinformatics utilities (e.g., samtools, bamtools, bedtools, etc.).

- Compilation:
  1. git clone https://github.com/nygenome/Lancet2.git
  2. cd Lancet2 && mkdir build && cd build
  3. cmake .. && make

- Pre-built docker images for Lancet2 are available on DockerHub: https://hub.docker.com/r/rmusunuri/lancet2



gnarzisi Update index.js ... ✓ 95fa992 22 minutes ago 267 commits

.chglog	Merge with Develop (#1)	2 months ago
.github/workflows	Website can now be edited	27 days ago
.idea	fix: use bq filter for all single allele haplotypes	last month
.vscode	chore: add vscode prefs	12 months ago
benchmarks	Merge with Develop (#1)	2 months ago
cmake	Merge with Develop (#1)	2 months ago
docker	fix: use right binary path	last month
include/lancet2	chore: minor refactoring and edits	29 days ago
scripts	Merge with Develop (#1)	2 months ago
src	Revert "Updating website"	19 days ago
tests	Merge with Develop (#1)	2 months ago
website	Update index.js	22 minutes ago
.clang-format	Merge with Develop (#1)	2 months ago
.clang-tidy	Merge with Develop (#1)	2 months ago
.editorconfig	chore: initial commit with c++17 codebase	2 years ago
.gitattributes	chore: initial commit with c++17 codebase	2 years ago
.gitignore	chore: initial commit with c++17 codebase	2 years ago
CMakeLists.txt	fix: use bq filter for all single allele haplotypes	last month
Dockerfile	chore: add default dockerfile and change scope	last month
LICENSE	chore: update to BSD license	3 months ago
README.md	Update funding info	26 days ago

About

v2.x of the microassembly based somatic variant caller

nygenome.github.io/lancet2/

Readme

BSD-3-Clause license

11 stars

6 watching

1 fork

Releases

1 tags

Create a new release

Packages 1

lancet2

Contributors 3

omicsnut Rajeeva Lochan Musunuri

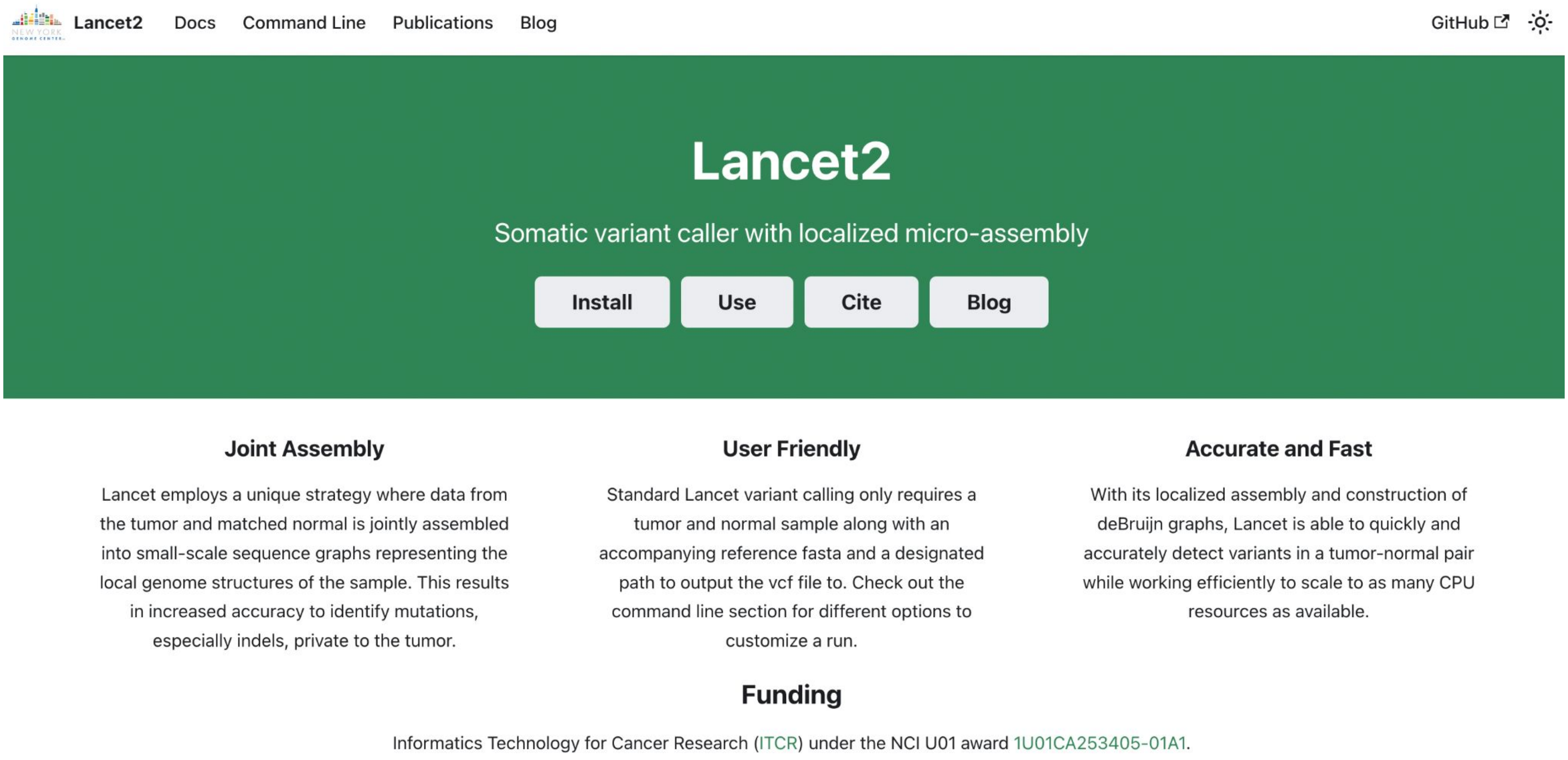
ZhuBry

gnarzisi Giuseppe Narzisi

Environments 1

github-pages Active

Documentation: https://nygenome.github.io/Lancet2/



Lancet2

Somatic variant caller with localized micro-assembly

Joint Assembly

Lancet employs a unique strategy where data from the tumor and matched normal is jointly assembled into small-scale sequence graphs representing the local genome structures of the sample. This results in increased accuracy to identify mutations, especially indels, private to the tumor.

User Friendly

Standard Lancet variant calling only requires a tumor and normal sample along with an accompanying reference fasta and a designated path to output the vcf file to. Check out the command line section for different options to customize a run.

Accurate and Fast

With its localized assembly and construction of deBruijn graphs, Lancet is able to quickly and accurately detect variants in a tumor-normal pair while working efficiently to scale to as many CPU resources as available.

Funding

Informatics Technology for Cancer Research (ITCR) under the NCI U01 award 1U01CA253405-01A1.

Acknowledgements

NIH NATIONAL CANCER INSTITUTE
Informatics Technology for
Cancer Research

NEW YORK
GENOME CENTER SM

Giuseppe Narzisi

Rajeeva Musunuri

Bryan Zhu

Jennifer Shelton

Minita Shah

André Corvelo

Nicolas Robine

Michael Zody

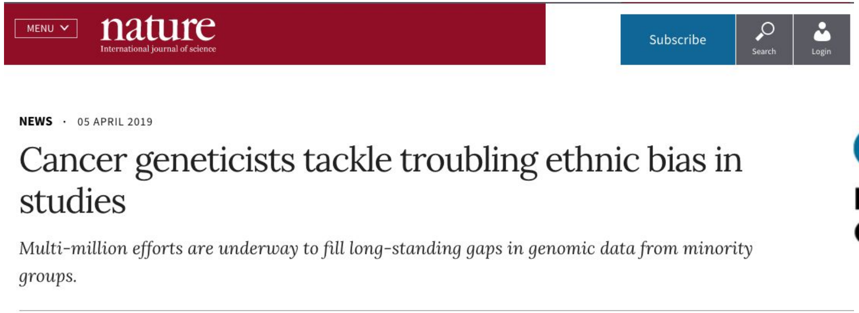
Kanika Arora, *MSKCC*

Ewa Bergmann, *Illumina*

Vladimir Vacic, *23andMe*

Anne-Katrin Emde, *Variant Bio*

Cancer Health Disparities in the news



MENU nature International journal of science Subscribe Search Login

NEWS · 05 APRIL 2019

Cancer geneticists tackle troubling ethnic bias in studies

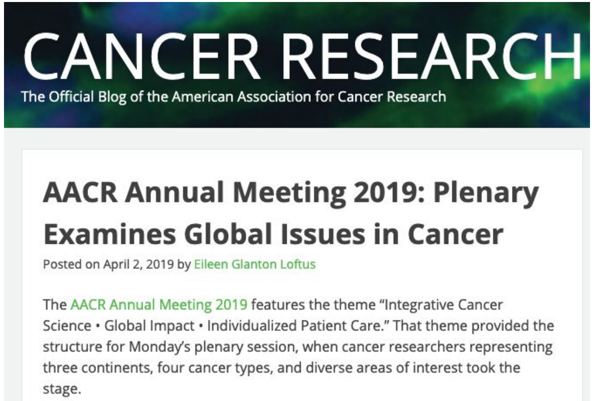
Multi-million efforts are underway to fill long-standing gaps in genomic data from minority groups.

VIEWPOINT

Ensuring Equity and Justice in the Care and Outcomes of Patients With Cancer

Cancer Cell

Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers



CANCER RESEARCH

The Official Blog of the American Association for Cancer Research

AACR Annual Meeting 2019: Plenary Examines Global Issues in Cancer

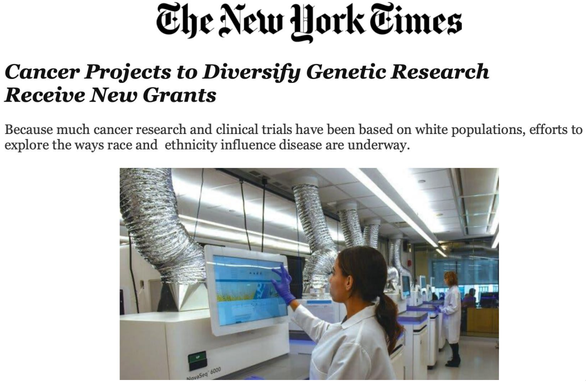
Posted on April 2, 2019 by Eileen Glanton Loftus

The AACR Annual Meeting 2019 features the theme "Integrative Cancer Science • Global Impact • Individualized Patient Care." That theme provided the structure for Monday's plenary session, when cancer researchers representing three continents, four cancer types, and diverse areas of interest took the stage.

NEWS FEATURE · 16 APRIL 2019

Facing up to injustice in genome science

Researchers from under-represented groups are making genomics more incl with communities that have been overlooked or abused.

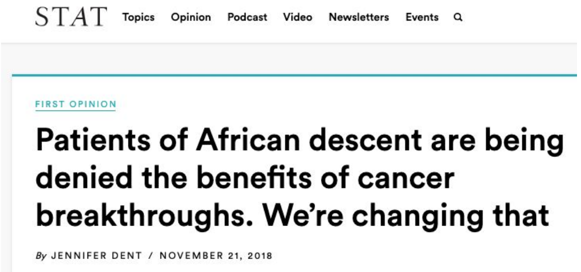


The New York Times

Cancer Projects to Diversify Genetic Research Receive New Grants

Because much cancer research and clinical trials have been based on white populations, efforts to explore the ways race and ethnicity influence disease are underway.

NYTimes, 9/11/2020



STAT Topics Opinion Podcast Video Newsletters Events

FIRST OPINION

Patients of African descent are being denied the benefits of cancer breakthroughs. We're changing that

By JENNIFER DENT / NOVEMBER 21, 2018

SCIENTIFIC AMERICAN.

We Need More Diversity in Our Genomic Databases

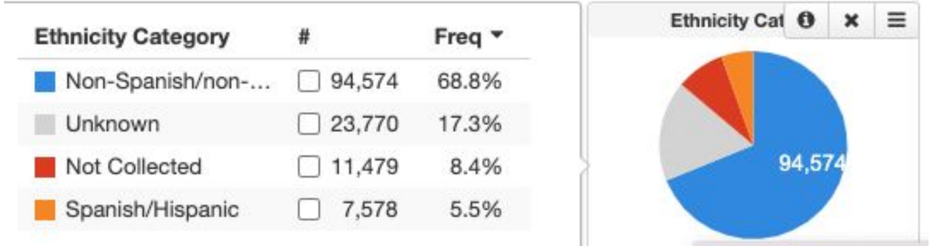
The ones we have now are too heavily skewed toward people of European descent

By Jonas Korlach on December 4, 2018

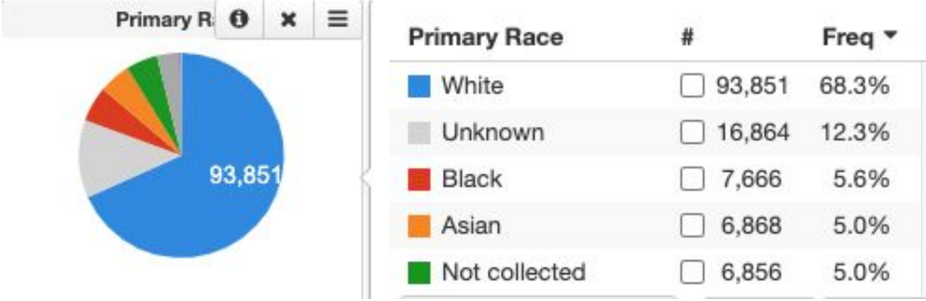
Public databases of cancer genomics

- A decade of tumor profiling
- Somatic landscape of the most prevalent cancer types
- Databases and interfaces, such as cBioPortal

However, 70-80% of the samples come from patients of European ancestry!



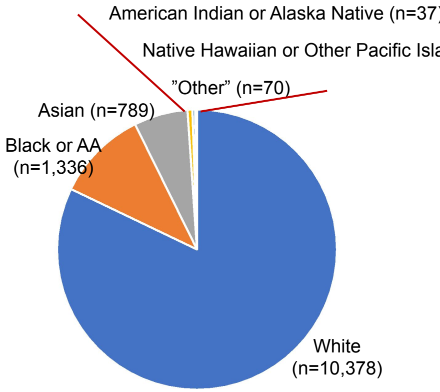
Ethnicity Category	#	Freq ▾
🟦 Non-Spanish/non-...	□ 94,574	68.8%
⬜ Unknown	□ 23,770	17.3%
🟥 Not Collected	□ 11,479	8.4%
🟧 Spanish/Hispanic	□ 7,578	5.5%

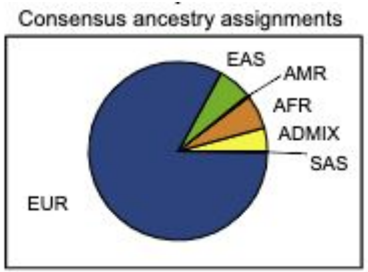


Primary Race	#	Freq ▾
🟦 White	□ 93,851	68.3%
⬜ Unknown	□ 16,864	12.3%
🟥 Black	□ 7,666	5.6%
🟧 Asian	□ 6,868	5.0%
🟩 Not collected	□ 6,856	5.0%

AACR Genie v12.0-public
(137401 patients)

TCGA Unreported: 19,606
Reported: 12,639





Carrot-Zhang, Chambwe et al. 2020

RALPH LAUREN CENTER FOR CANCER CARE

SUNY DOWNSTATE Medical Center

NewYork-Presbyterian Brooklyn Methodist Hospital

Montefiore HEALTH SYSTEM

U.S. Department of Veterans Affairs

NewYork-Presbyterian Queens

Northwell Health

Stony Brook University

COLUMBIA UNIVERSITY MEDICAL CENTER

Memorial Sloan Kettering Cancer Center

Weill Cornell Medicine

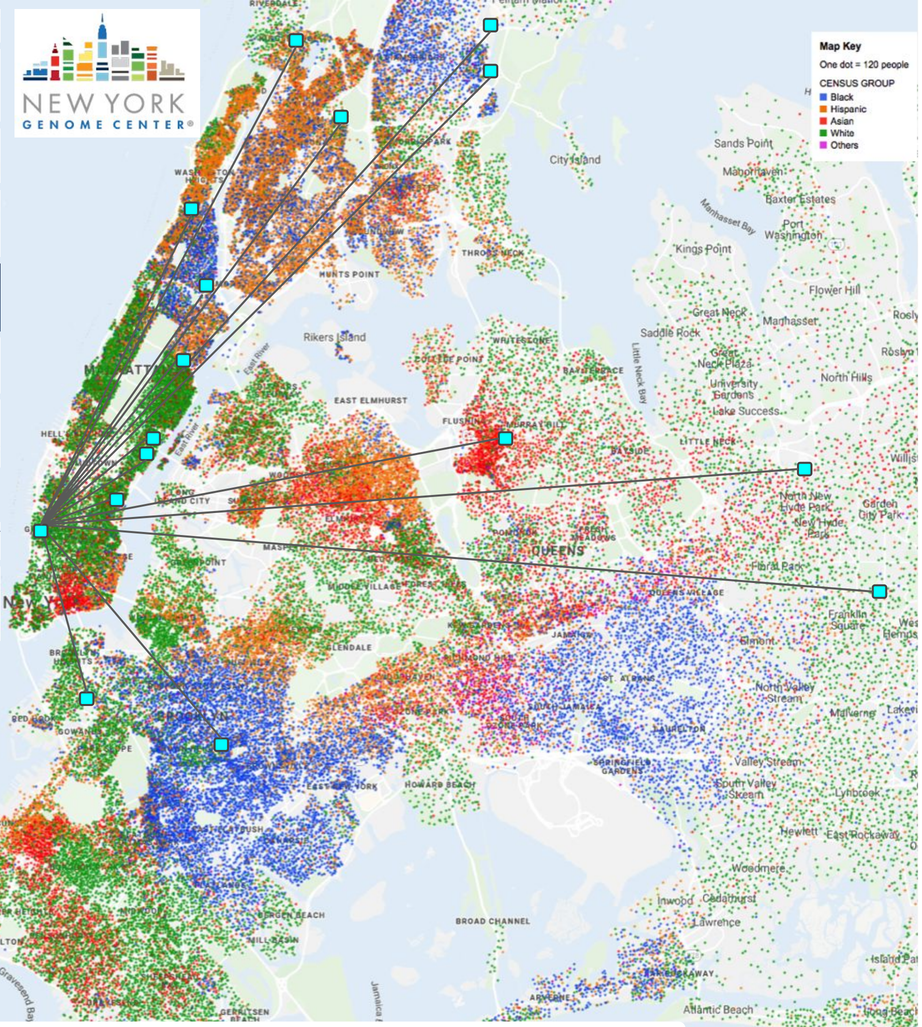
Icahn School of Medicine at Mount Sinai

NYU School of Medicine NYU LANGONE MEDICAL CENTER

James J. Peters Department of Veterans Affairs Medical Center Bronx, New York

CSH Cold Spring Harbor Laboratory

NEW YORK GENOME CENTER



Polyethnic-1000 Participating Sites

source https://www.nytimes.com/interactive/2015/07/08/us/census-race-map.html

P1000 Infrastructure

- 16 participating sites
- >40 collaborators
- 44 working group members
- 21 sites coordinators and pathologists
- Partners include: IRB, legal, technology transfer
- Supported by our scientific leads at the GCCG

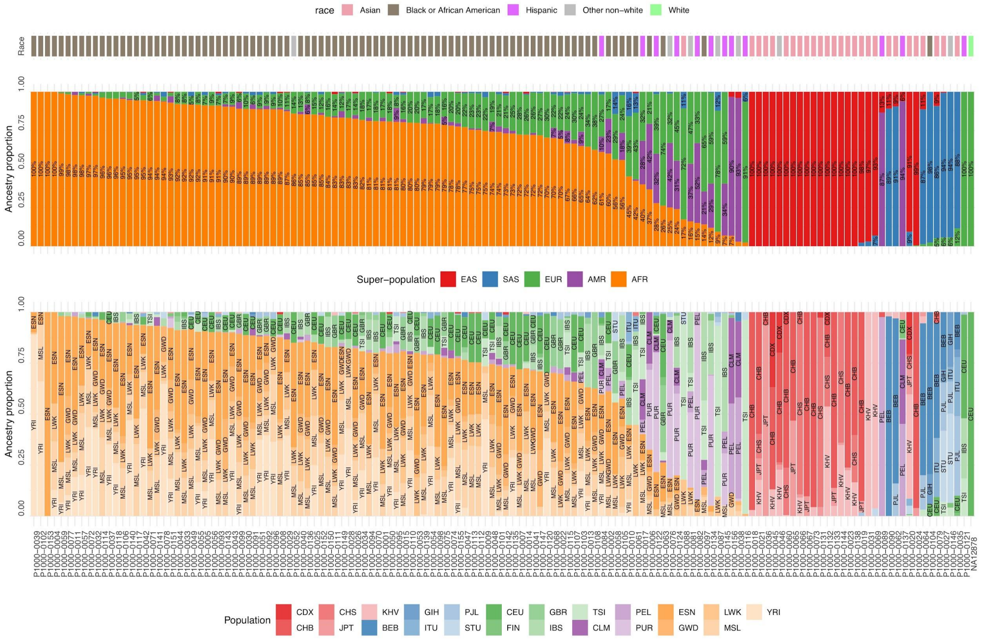
Harold Varmus, MD,
NYGC Senior Associate Core Member,
Weill Cornell Medicine Professor

Charles Sawyers, MD
Memorial Sloan Kettering Cancer Center

David Tuveson, MD
Cold Spring Harbor Laboratory

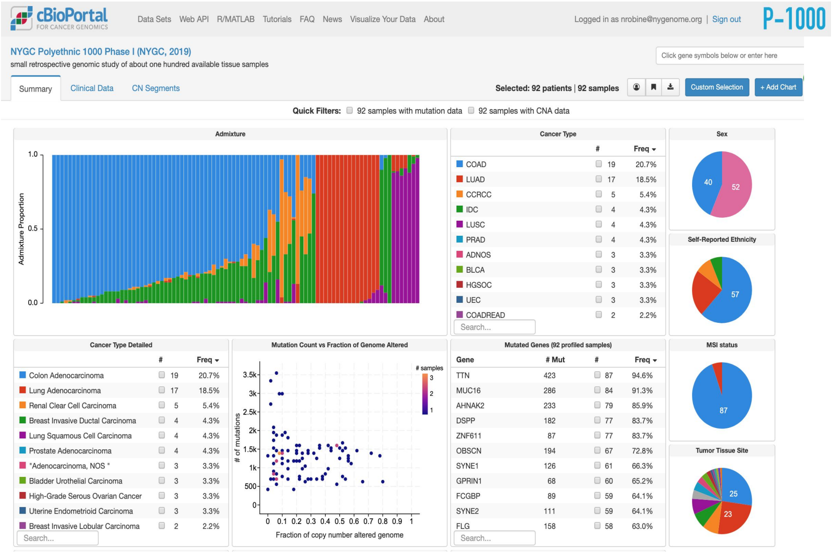
Polyethnic-1000 Phase 1

- 160 samples from 13 institutions
- “Non-white” patients
- WES+RNA (tumor-only)
- Genetic ancestry estimation



“Somatic” variants in local cBioportal

Data sharing within the consortium



Logged in as nrobine@nygenome.org

NYGC Polyethnic 1000 Phase I (NYGC, 2019)

small retrospective genomic study of about one hundred available tissue samples

Click gene symbols below or enter here

Selected: 92 patients | 92 samples

Quick Filters: □ 92 samples with mutation data □ 92 samples with CNA data

Cancer Type

	#	Freq ▾
COAD	□ 19	20.7%
LUAD	□ 17	18.5%
CCRCC	□ 5	5.4%
IDC	□ 4	4.3%
LUSC	□ 4	4.3%
PRAD	□ 4	4.3%
ADNOS	□ 3	3.3%
BLCA	□ 3	3.3%
HGSOC	□ 3	3.3%
UEC	□ 3	3.3%
COADREAD	□ 2	2.2%

Search...

Cancer Type Detailed

	#	Freq ▾
Colon Adenocarcinoma	□ 19	20.7%
Lung Adenocarcinoma	□ 17	18.5%
Renal Clear Cell Carcinoma	□ 5	5.4%
Breast Invasive Ductal Carcinoma	□ 4	4.3%
Lung Squamous Cell Carcinoma	□ 4	4.3%
Prostate Adenocarcinoma	□ 4	4.3%
"Adenocarcinoma, NOS "	□ 3	3.3%
Bladder Urothelial Carcinoma	□ 3	3.3%
High-Grade Serous Ovarian Cancer	□ 3	3.3%
Uterine Endometrioid Carcinoma	□ 3	3.3%
Breast Invasive Lobular Carcinoma	□ 2	2.2%

Search...

Mutated Genes (92 profiled samples)

Gene	# Mut	#	Freq ▾
TTN	423	□ 87	94.6%
MUC16	286	□ 84	91.3%
AHNAK2	233	□ 79	85.9%
DSPP	182	□ 77	83.7%
ZNF611	87	□ 77	83.7%
OBSCN	194	□ 67	72.8%
SYNE1	126	□ 61	66.3%
GPRIN1	68	□ 60	65.2%
FCGBP	89	□ 59	64.1%
SYNE2	111	□ 59	64.1%
FLG	158	□ 58	63.0%

Search...

Polyethnic-1000.

"Phase 2"

- 1000 samples collected in 2021-2022
- Retrospective and prospective samples
- Tumor-normal Whole-Genome Sequencing
- Tumor RNA-seq
- Research samples, consented for data sharing

7 Projects

- Bladder
- Breast/Prostate
- Pancreas
- Multiple Myeloma
- Lung
- Colon
- Endometrial

CSH Cold Spring Harbor Laboratory

Weill Cornell Medicine

NYP

NYU

Memorial Sloan Kettering Cancer Center

COLUMBIA
COLUMBIA UNIVERSITY IRVING MEDICAL CENTER

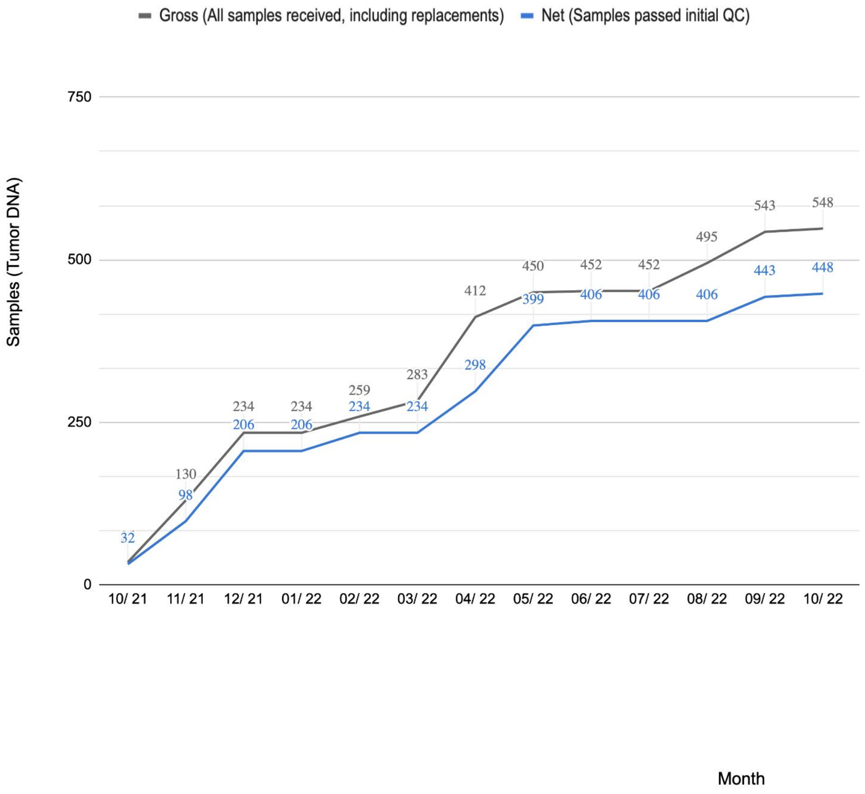
Northwell Health®

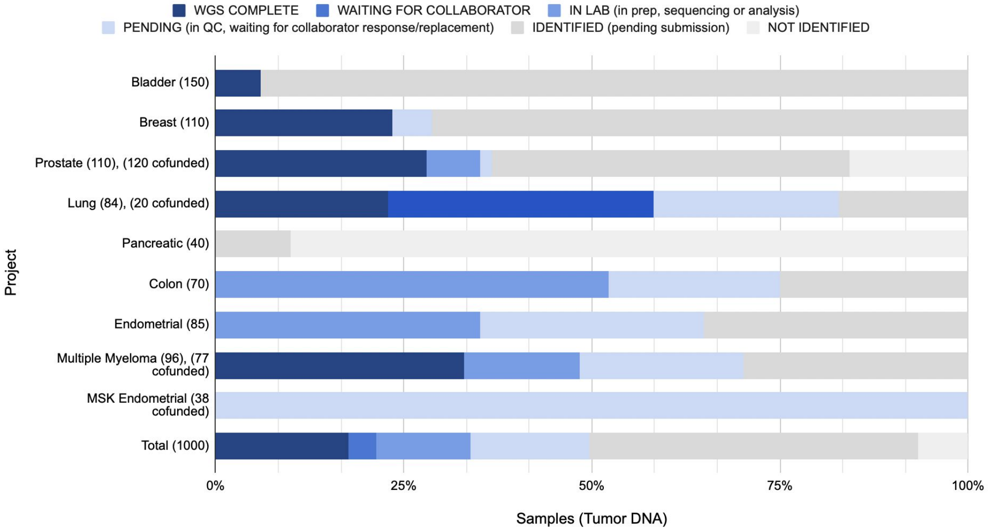
SUNY DOWNSTATE Health Sciences University

Mount Sinai Hospital

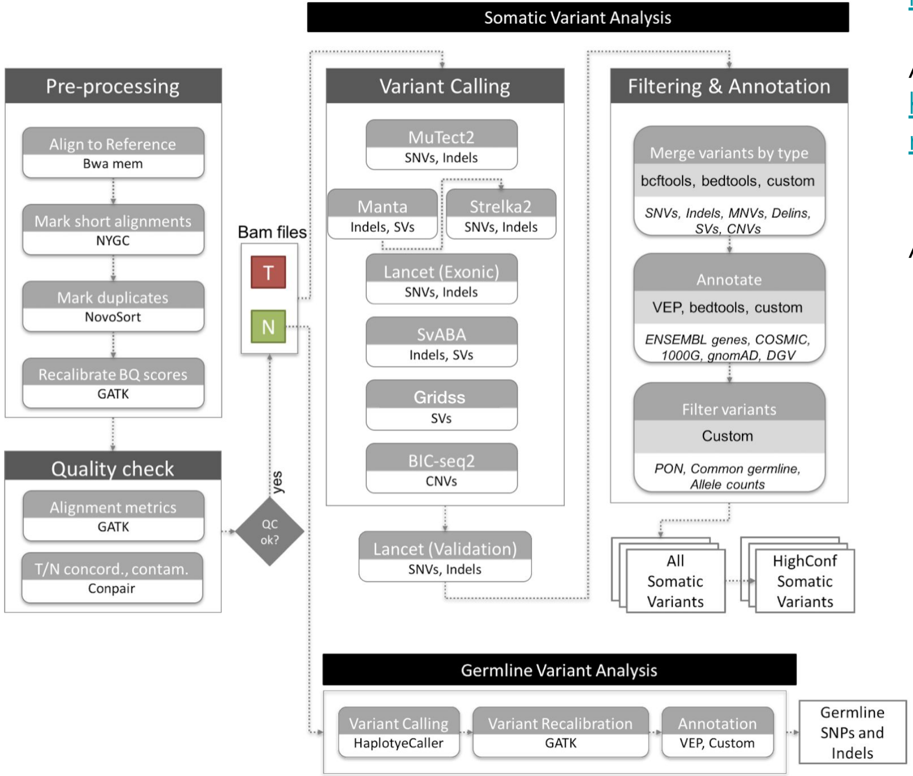
UNIVEF OF MIAI

Samples received to date





WGS pipeline



NYGC Somatic Pipeline v7 (Arora et al. 2019)

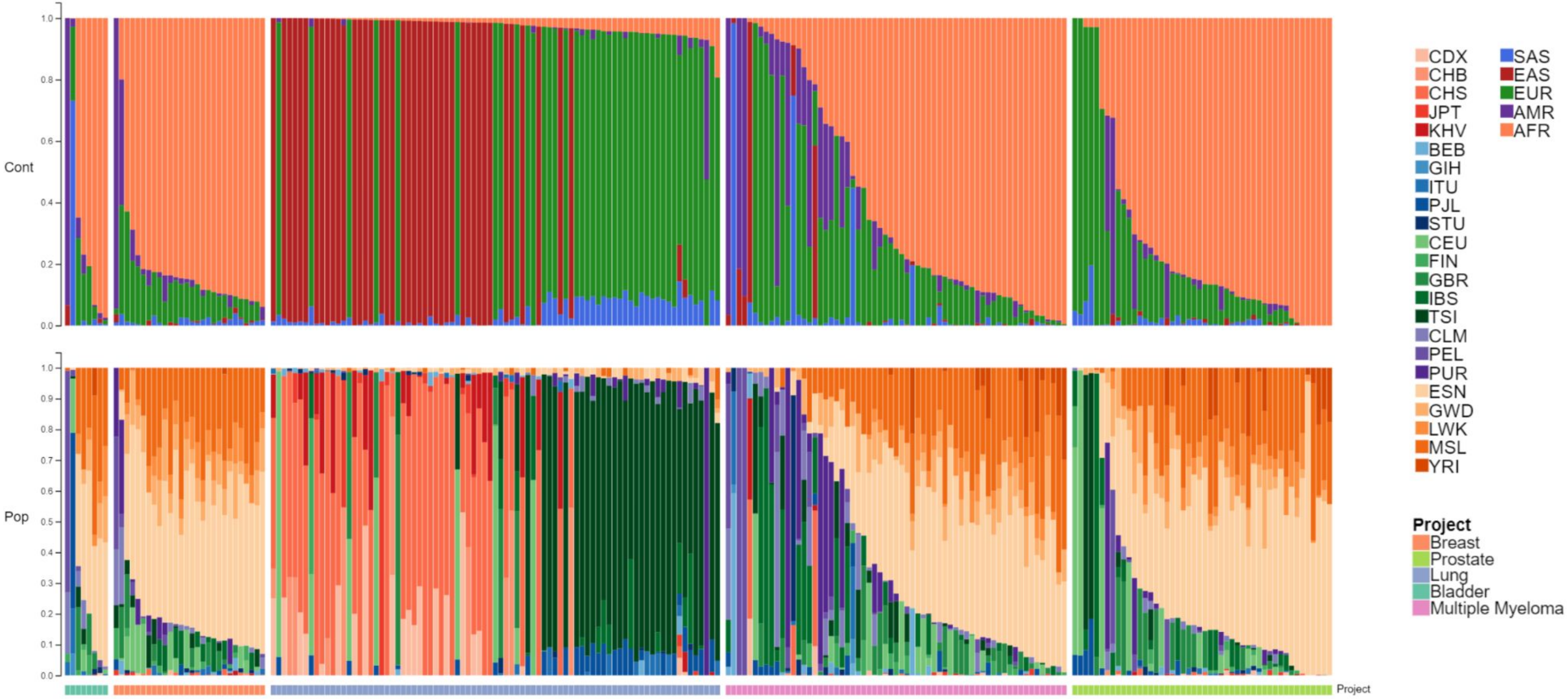
Code:
https://bitbucket.nygenome.org/scm/compbio/wdl_port.git

Additional documentation:
https://www.nygenome.org/bioinformatics/software/nygc-cancer-pipeline/

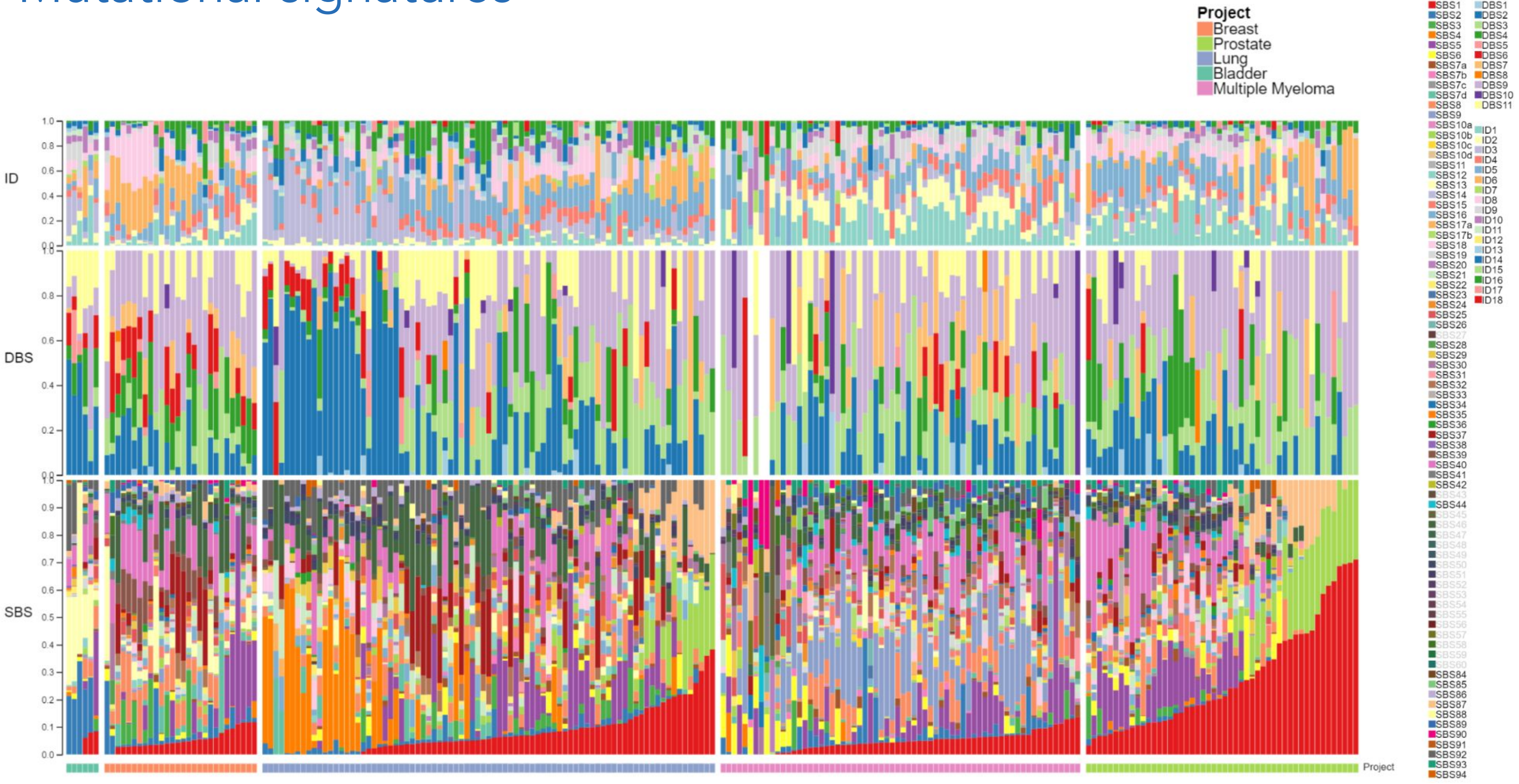
Additional analyses:

- Mutational signatures (COSMIC v3.3) with deconstructSig
- MicroSatellite Instability with MANTIS
- HLA Typing with Kourami
- Ancestry estimation with fastNGSAdmix
- Homologuous Recombination Deficiency with HRDetect
- Purity/ploidy estimation
- JaBba (Complex Structural Variants)
- Recurrence analysis with FishHook, GISTIC, etc
- RNA-DNA integration
- Batch effect correction
- Immune infiltration deconvolution with CIBERSORT

Genetic ancestry estimation



Mutational signatures



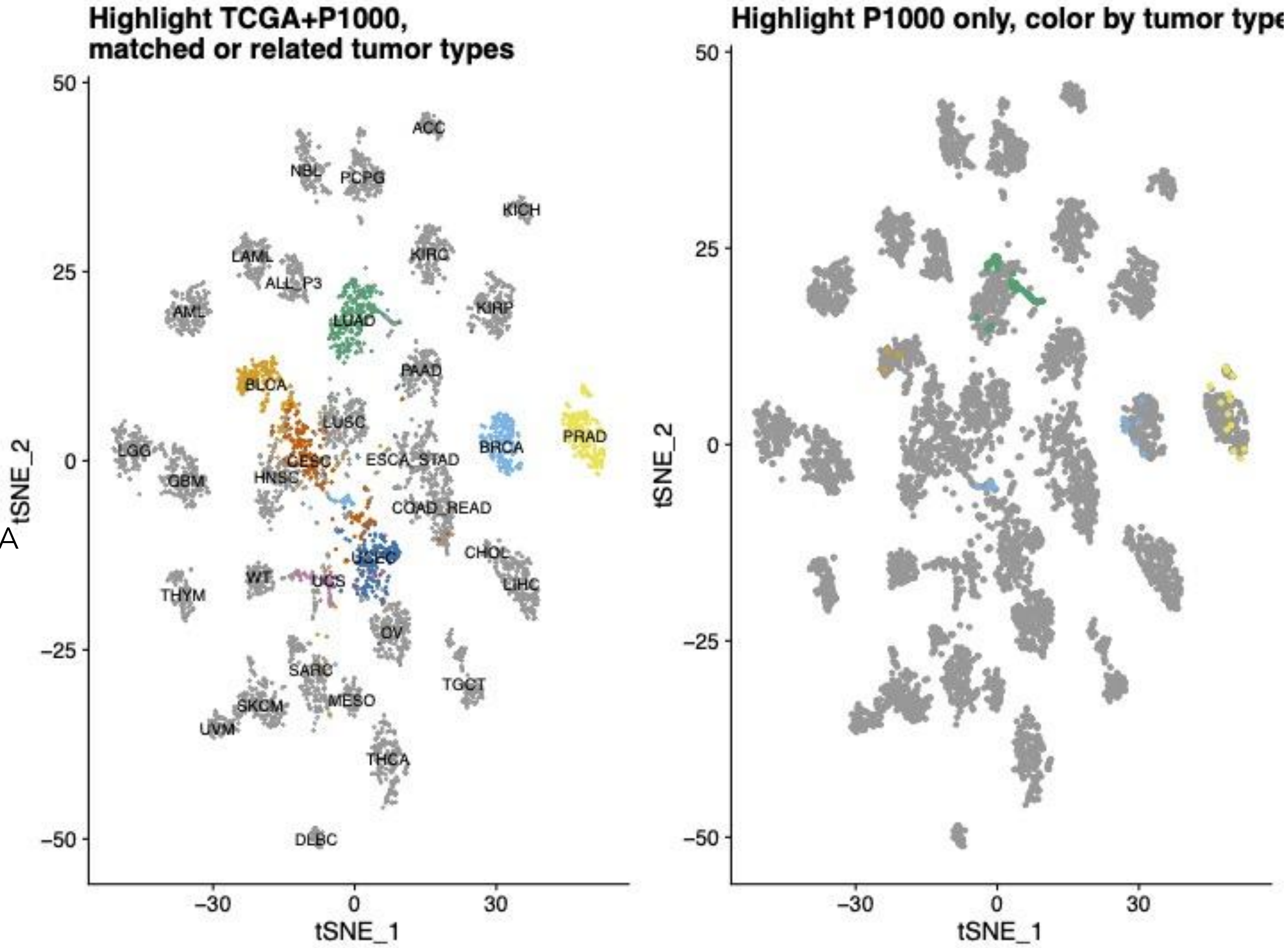
RNA-seq

Alignment with STAR
Gene quantification with featureCounts
Differential expression with DESeq2
Fusion discovery with FusionCatcher and STAR-Fusion

Unsupervised clustering of TCGA expression profiles.

Clustering by tumor types.

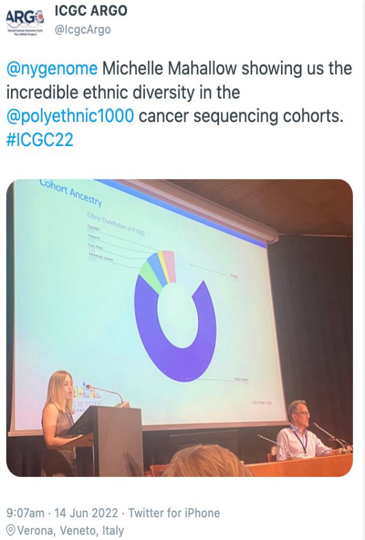
Overlay of P1000 samples.



Polyethnic-1000 Next Steps

- Data analysis and data sharing
- Patient and Community Outreach
- Clinical Sequencing and return of results to patients
- More minority populations
- Additional cancer types

AACR Health Disparity conference, Sept 2022



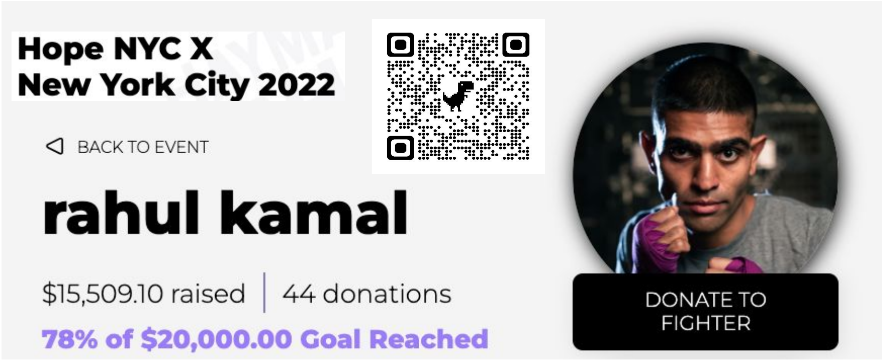
ICGC ARGO
@IcgcArgo

@nygenome Michelle Mahallow showing us the incredible ethnic diversity in the @polyethnic1000 cancer sequencing cohorts. #ICGC22

9:07am · 14 Jun 2022 · Twitter for iPhone
Verona, Veneto, Italy

Onyinye
Balogun

Melissa
Davis



Hope NYC X
New York City 2022

BACK TO EVENT

rahul kamal

$15,509.10 raised | 44 donations

78% of $20,000.00 Goal Reached

Acknowledgments

NYGC Project Management

- Lara Winterkorn
- Michelle Mehallow
- Cat Reeves

NYGC Ethnicity and Cancer Scholars

- Melissa Davis
- Onyinye Balogun

NYGC Development Office, Sweng, ResComp, CompBio, Seq lab.

Contact

- nrobine@nygenome.org
- lwinterkorn@nygenome.org
- polyethnic1000@nygenome.org

All patients and participants to the Polyethnic-1000 studies

Polyethnic-1000

A collaborative effort in cancer research organized through the New York Genome Center and led by Harold Varmus, MD, and Charles Sawyers, MD, the Polyethnic-1000 initiative supports research that explores the contribution of genetic ancestry to the understanding of cancer incidence, progression and clinical care. Working with the exceptionally diverse population of New York City, Polyethnic-1000 researchers are exploring the molecular landscapes of cancer in ethnically diverse populations of patients with the goal of bringing advanced clinical genomics to cancer care to historically underserved communities.

We gratefully acknowledge the following individuals and organizations for their generous support.

ILLUMINA INC.

WESLIE JANEWAY

THE MARK FOUNDATION FOR CANCER RESEARCH

MORTIMER B. ZUCKERMAN FAMILY FUND

THE NEW YORK COMMUNITY TRUST

DONNA AND BEN ROSEN

P1000 Steering Committee

- Charles Sawyers
- David Tuveson
- Harold Varmus
- Sam Aparicio

Support

- Mark Foundation
- Illumina
- Zuckerman Family Fund
- New York Community Trust
- Weslie Janeway
- Ben and Donna Rosen
- CSHL-Northwell
- Columbia
- Weill-Cornell

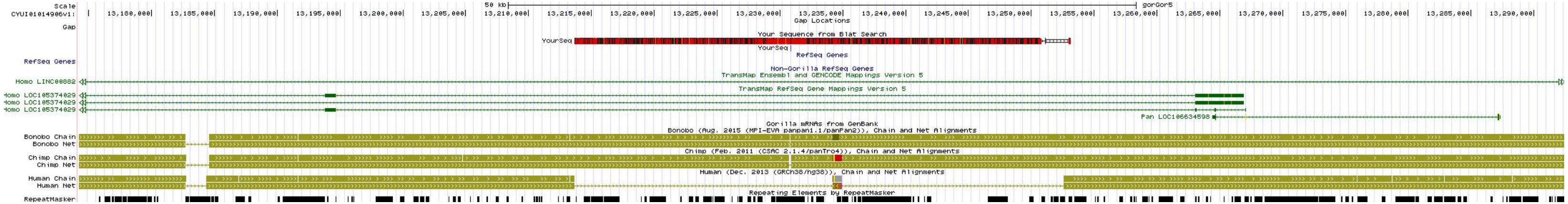
ABSINTHE INSERTION CALLING

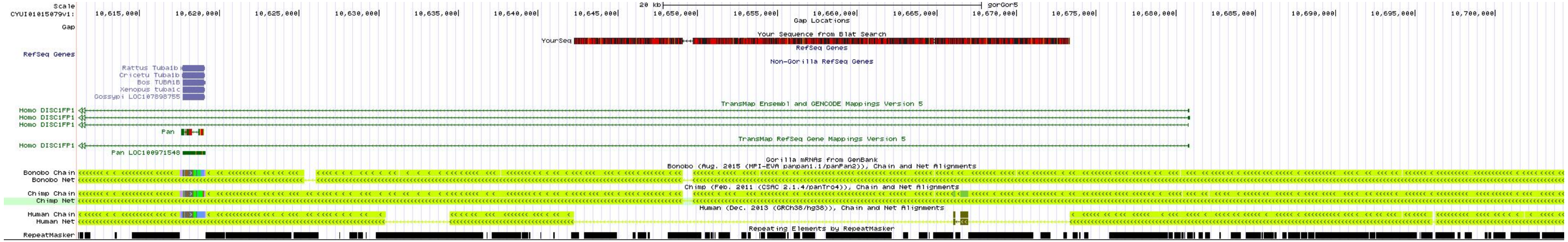
- Calling “insertions” from short reads has traditionally been difficult
- Absinthe identifies reads that don’t map or mismap and assembles them
- The resulting contigs can then be placed back on the reference

- Used to call variants from several projects including:
  - TOPMed (Taliun et al., *Nature*, 2021)
  - 1000 Genomes (Byrska-Bishop et al., *Cell*, 2022)
  - HGSVC analysis of 1000 Genomes (Ebert et al., *Science*, 2021)

- Recently run in the cloud on CCDG Freeze 3
- Working on call set for Alzheimer’s Disease
- Work of André Corvelo at NYGC

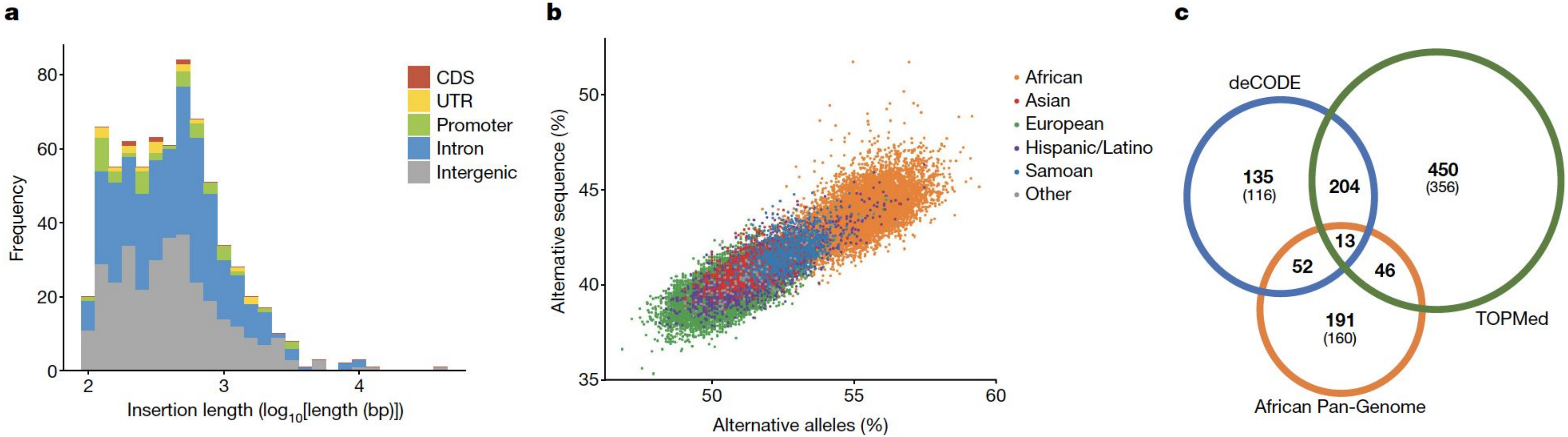
EXAMPLES OF ASSEMBLED INSERTIONS





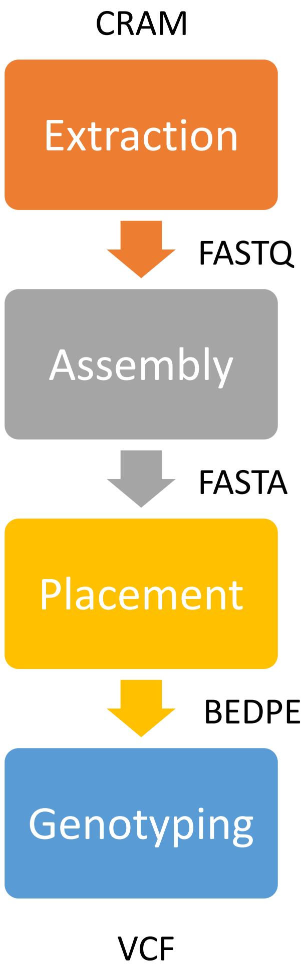
RESULTS FROM TOPMED

From the TOPMed 53,831 analysis:



Taliun et al., *Nature*, 2021

ABSINTHE PIPELINE



CRAM

Extraction

FASTQ

Assembly

FASTA

Placement

BEDPE

Genotyping

VCF

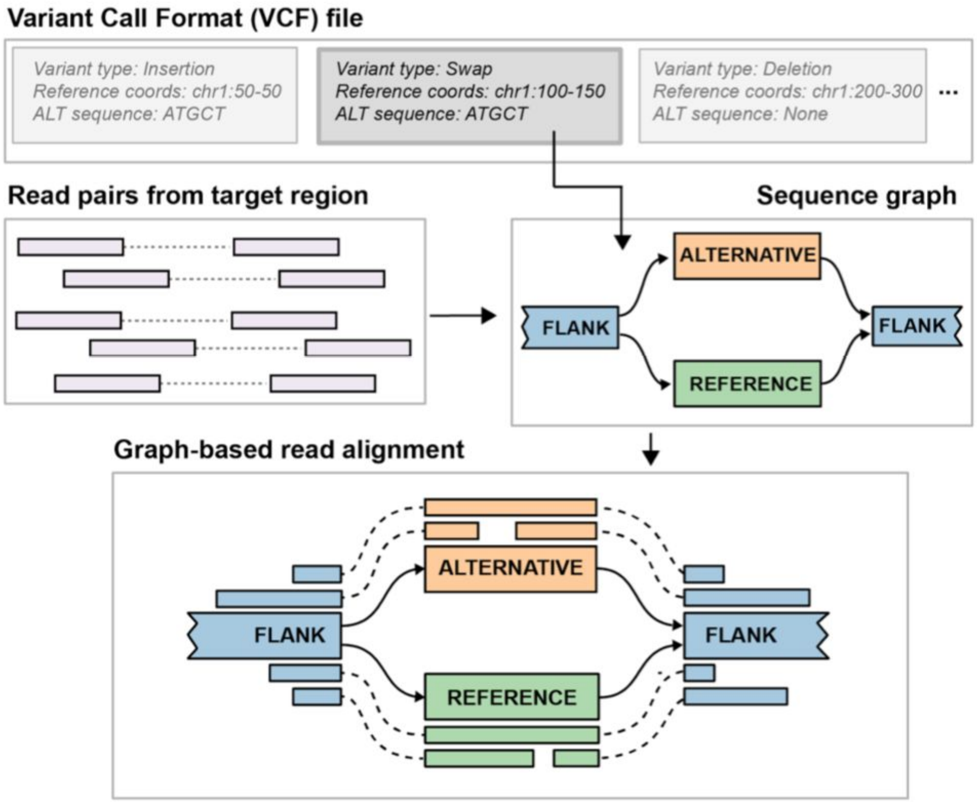
- Not properly mapped read-pairs
- phiX removal, adapter clipping, low quality base trimming

- *de novo*
- ABySS v2.0.2
- k = 77

- *ab initio*:
  - Flank maximal best hit pairs to GRCh38
  - Alignment with gap excision
- LiftOver:
  - Hominid alignment and reference-based scaffolding
  - Coordinate transposition to GRCh38
  - Alignment with gap excision

- Merging
- Paragraph v2.4b

PARAGRAPH GENOTYPING



Variant Call Format (VCF) file

Variant type: Insertion
Reference coords: chr1:50-50
ALT sequence: ATGCT

Variant type: Swap
Reference coords: chr1:100-150
ALT sequence: ATGCT

Variant type: Deletion
Reference coords: chr1:200-300
ALT sequence: None

...

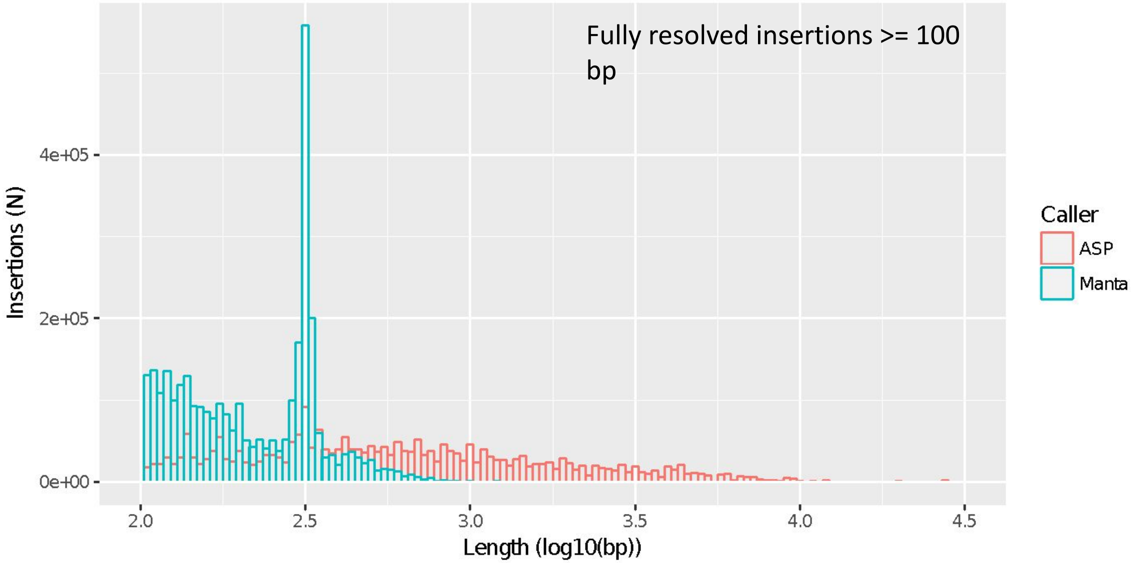
Read pairs from target region

Sequence graph

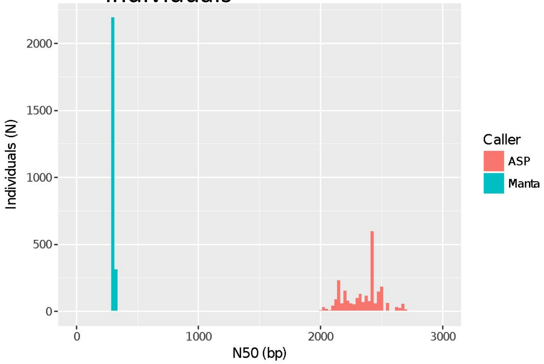
Graph-based read alignment

Chen *et al.*, *Genome Biology*, 2019

INSERTION LENGTH DISTRIBUTION



- Consistent across individuals



- Absinthe calls are a good complement to Manta’s as they extend well into the range of 1Kb – 10Kbp
- Several fully resolved insertions are longer than 10Kbp

1KG - UNIFIED CALLSET ACROSS 3202 SAMPLES

Merging:

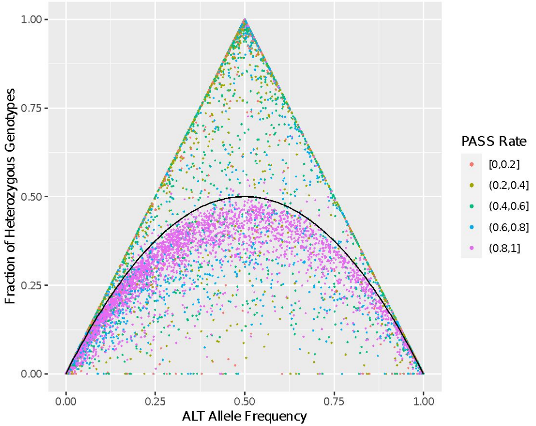
- MSA-based
- Input:
  - 3,583,674 per-sample calls
    - Self-genotyped (1, 0/1, 1/1)
  - 657,757 distinct
  - 12,222 loci
- Output:
  - 12,704 insertions

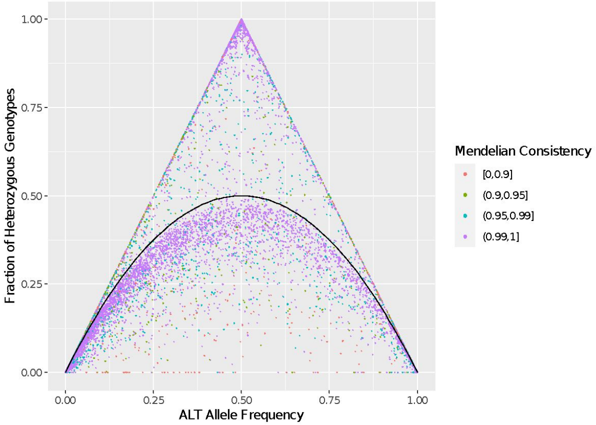
Genotyping:

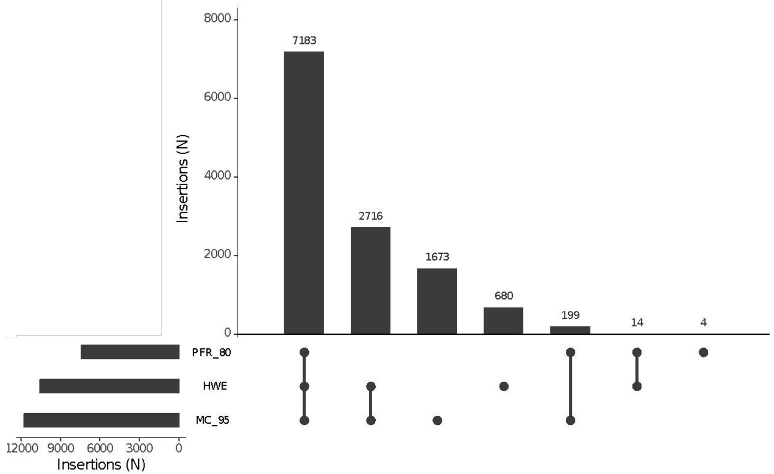
- Paragraph (Chen et al, 2019)

Filters:

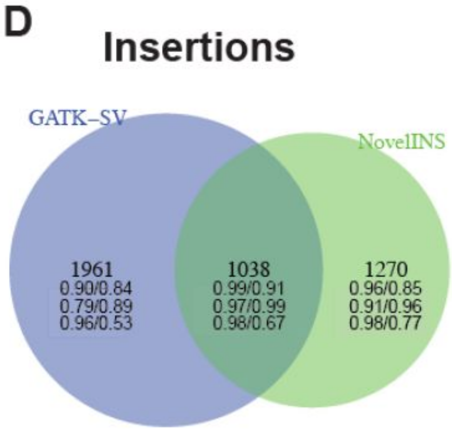
- Super population PASS-filter rate [all >= 0.8]
- Super population HWE [any > 10^{-6}]
- Mendelian Consistency based on 602 trios [>= 0.95]
- Output:
  - 7,183 HQ genotyped insertions





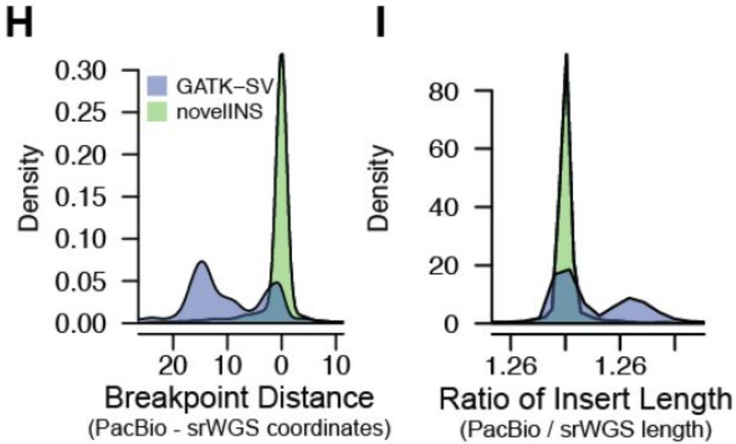


RESULTS FROM 1000 GENOMES



Insertions detected per sample.

Underneath are validation rates and fraction overlapping for three orthogonal methods.



Accuracy of breakpoints and insertion length by comparison to long read sequencing on the same samples.

Byrska-Bishop *et al.*, *Cell*, 2022

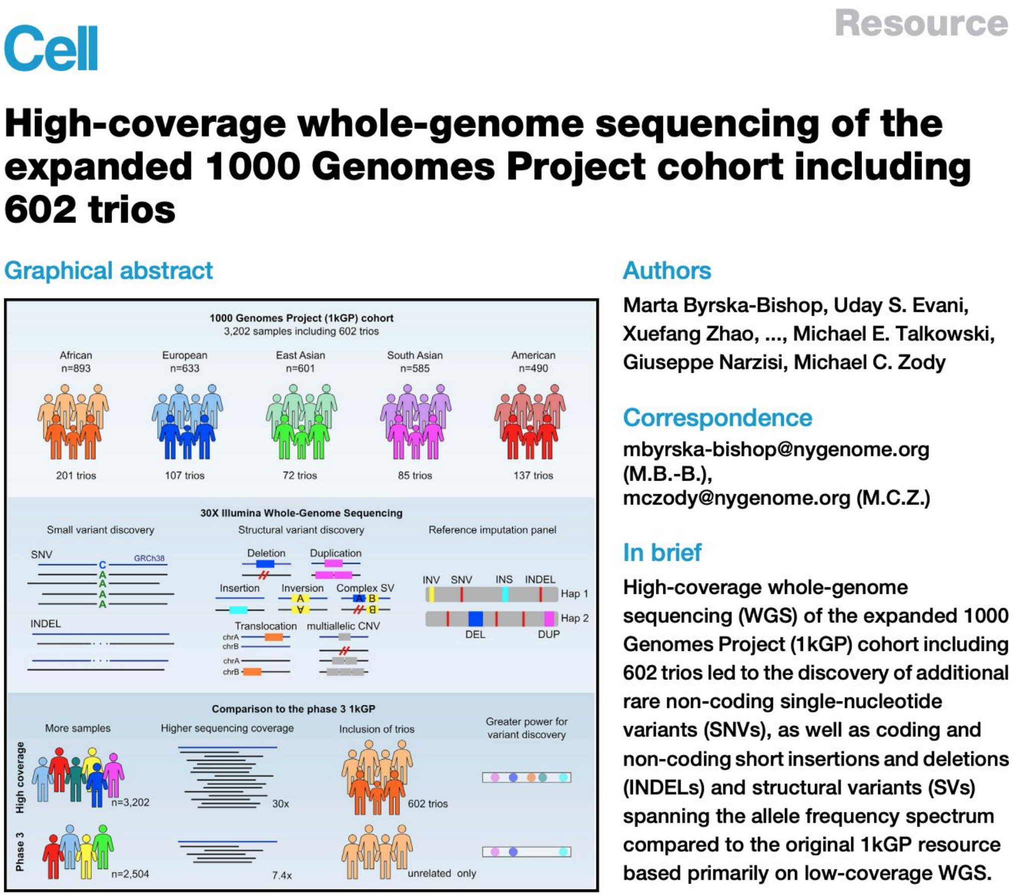
1000 Genomes Project (1kGP)

- International research effort launched in 2008 to establish an **open-access catalog of human genetic variation**.
- Culminated in 2015 with the release of the final, phase 3 variant call set based on **2,504 unrelated samples** collected from 26 populations across 5 continental regions of the world.
- **Phase 3** was based primarily on low-coverage whole-genome sequencing (WGS), deep coverage whole-exome sequencing (WES), and genotyping chip data.
- Discovered 84.7 mln SNVs, 3.6 mln INDELs, and 68.8 thousand SVs.
- 1kGP resources utilized for **foundational applications** such as genotype imputation, expression quantitative trait loci (eQTL) mapping, variant pathogenicity prioritization, population history, and evolutionary genetics studies.

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
END OF THE BEGINNING
Final phase of 1000 Genomes Project maps human genetic variation in open-access resource PAGES 52, 68 & 75
REVIEWS
AUTUMN BOOKS SPECIAL
Fizzwars, geopoetry, mind and matter
PAGE 34
MAUNA KEA OBSERVATORY
MOUNTAIN DIFFICULTIES
Is the Thirty Meter Telescope a step too far?
PAGE 24
THE HUMAN GENOME
25 YEARS OF BIG BIOLOGY
Three major players reflect on lessons learned
PAGE 29
NATURE.COM/NATURE
1 October 2015 £10
Vol 526, No. 7571
TIFF Preview

The 1000 Genomes Project Consortium. 2015. Nature; Sudmant et al. 2015. Nature

Expansion and upgrade of the 1kGP resource



Cell

Resource

High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios

Graphical abstract

Authors

Marta Byrska-Bishop, Uday S. Evani, Xuefang Zhao, ..., Michael E. Talkowski, Giuseppe Narzisi, Michael C. Zody

Correspondence

mbyrska-bishop@nygenome.org (M.B.-B.), mczody@nygenome.org (M.C.Z.)

In brief

High-coverage whole-genome sequencing (WGS) of the expanded 1000 Genomes Project (1kGP) cohort including 602 trios led to the discovery of additional rare non-coding single-nucleotide variants (SNVs), as well as coding and non-coding short insertions and deletions (INDELs) and structural variants (SVs) spanning the allele frequency spectrum compared to the original 1kGP resource based primarily on low-coverage WGS.

Highlights

- Expansion of the 1000 Genomes Project (1kGP) resource to include 602 trios.
- High-coverage whole-genome sequencing of the expanded 1kGP cohort.
- Discovery of more rare SNVs as well as INDELs and SVs across the frequency spectrum.
- Generation of an improved and accessible reference imputation panel.

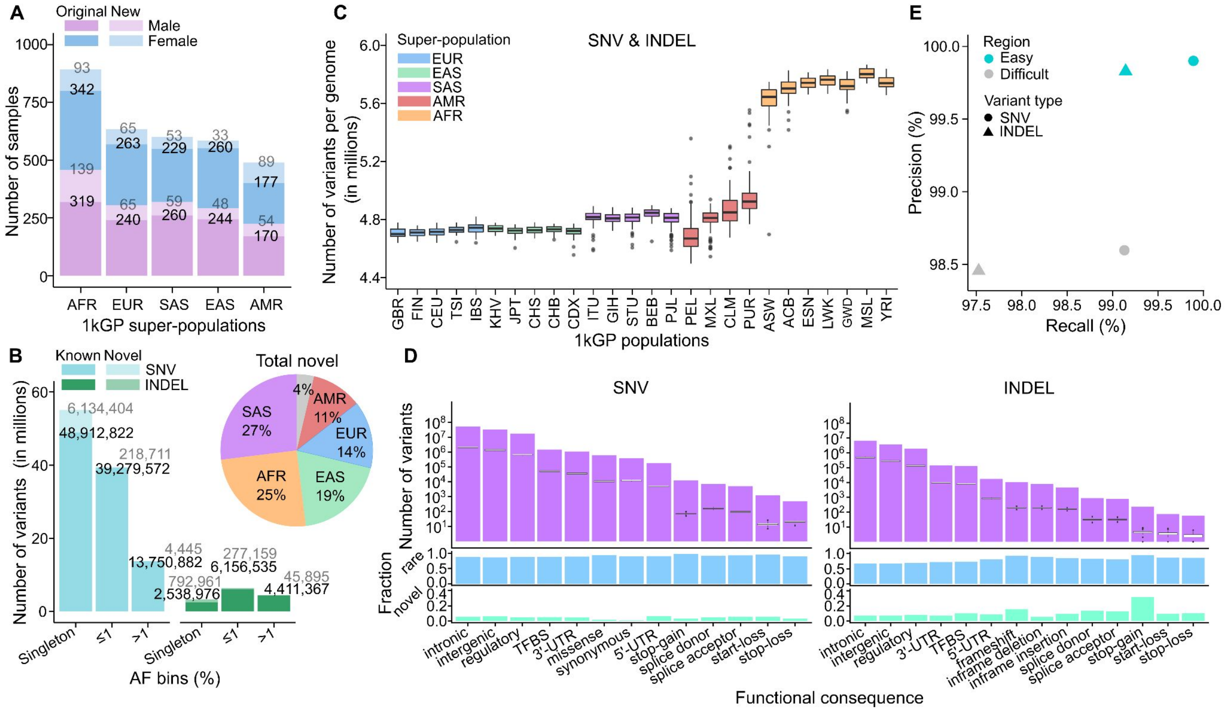
Outline

1. Small variant discovery.
2. Structural variant discovery.
3. Generation of an integrated reference imputation panel.

Outline

1. Small variant discovery.
2. Structural variant discovery.
3. Generation of an integrated reference imputation panel.

Over 111 million SNVs and 14 million INDELs discovered across 3,202 1kGP samples



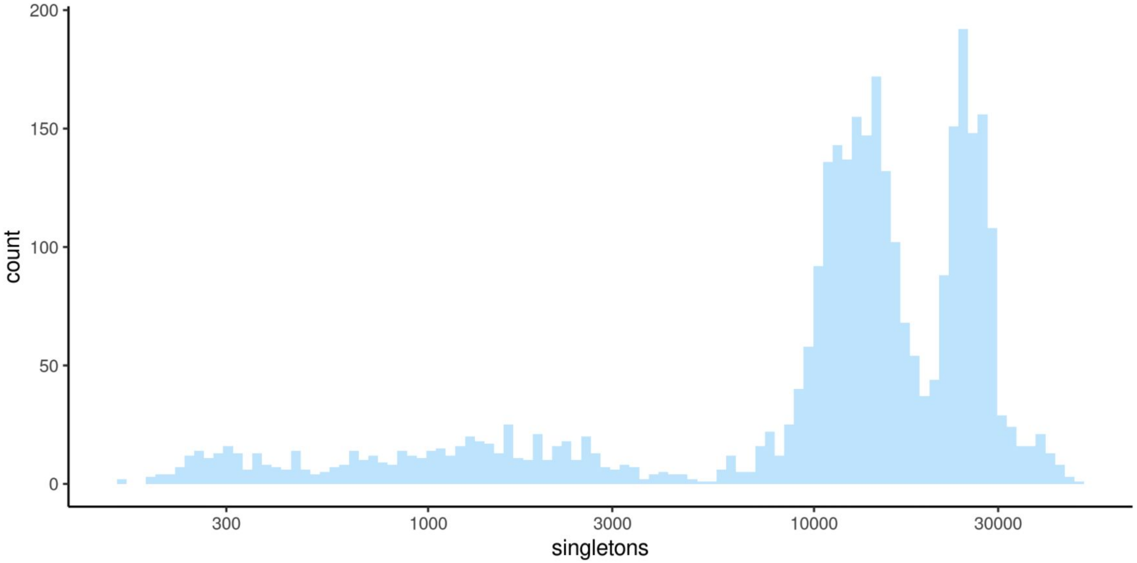
	Cohort level		Per sample (mean)	
	SNV	**INDEL**	**SNV**	**INDEL**
Total	111,048,944	14,435,076	4,080,992	871,923
Singletons	55,047,226	3,331,937	23,197 (unrelated)	
Novel	6,357,560	1,116,015		

Small variant discovery:

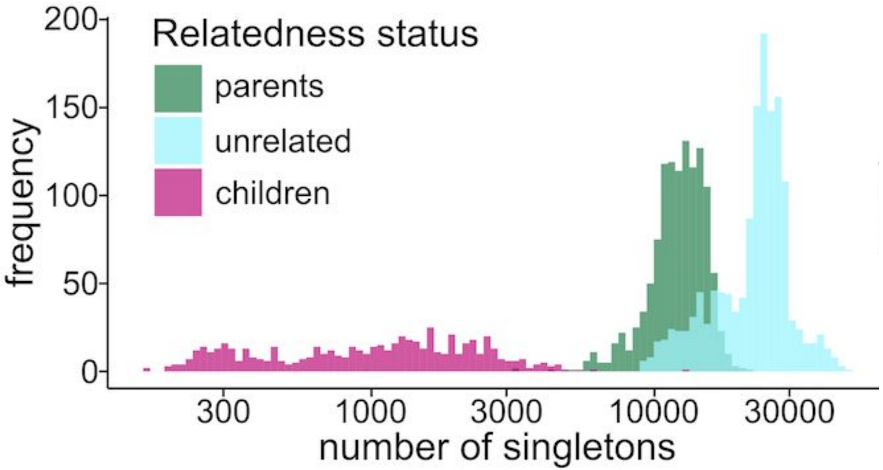
- 117,175,809 small variant loci, which represent 125,484,020 distinct alternate alleles.
- 4,952,915 small variants per sample on average.
- Functional predictions:
  - 605,896 missense,
  - 384,451 synonymous,
  - 36,520 pLoF mutations.
- At MAF <=1%, each sample carries on average:
  - 11 stop-gain,
  - 18 essential splice,
  - 14 frameshift mutations.
- FDR:
  - 0.3% for SNVs
  - 1.15% for INDELs

Taking a closer look at the singletons in the 1kGP cohort

Singletons: variants with allele count (AC)=1 across the 3,202 samples



The number of singletons per genome varies depending on the sample's relatedness status

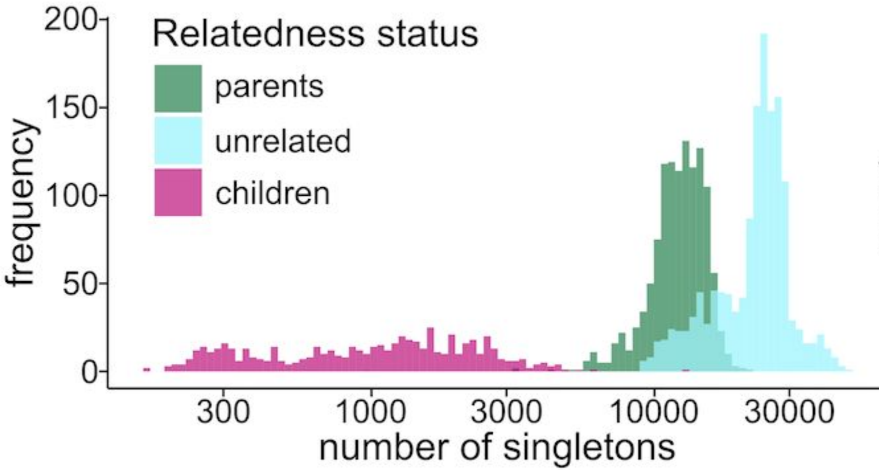


“Private” variants (~20,000 per genome): inherited variants private to one family.

- **Children**: 100% of them are shared with parents (i.e. are not counted as singletons).
- **Parents**: 50% shared with children (i.e. 50% are counted as singletons).
- **Unrelated**: all counted as singletons.

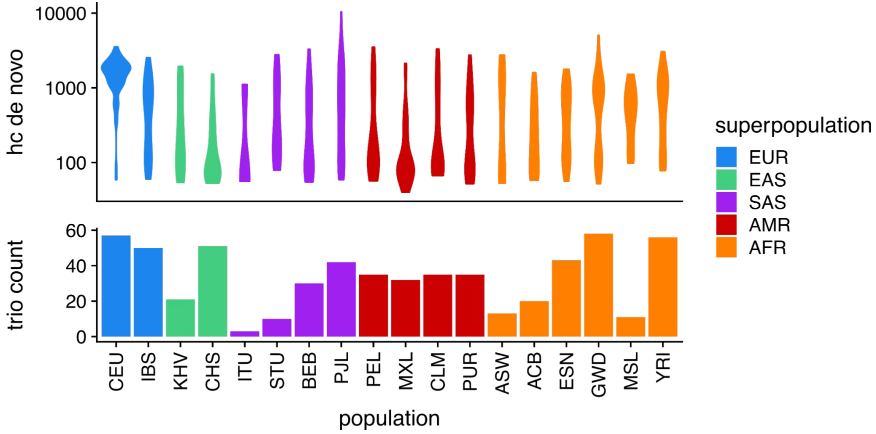
Zook et al., 2019, Wagner et al., 2021.

The number of singletons per genome varies depending on the sample's relatedness status



“Private” variants (~20,000 per genome): inherited variants private to one family.

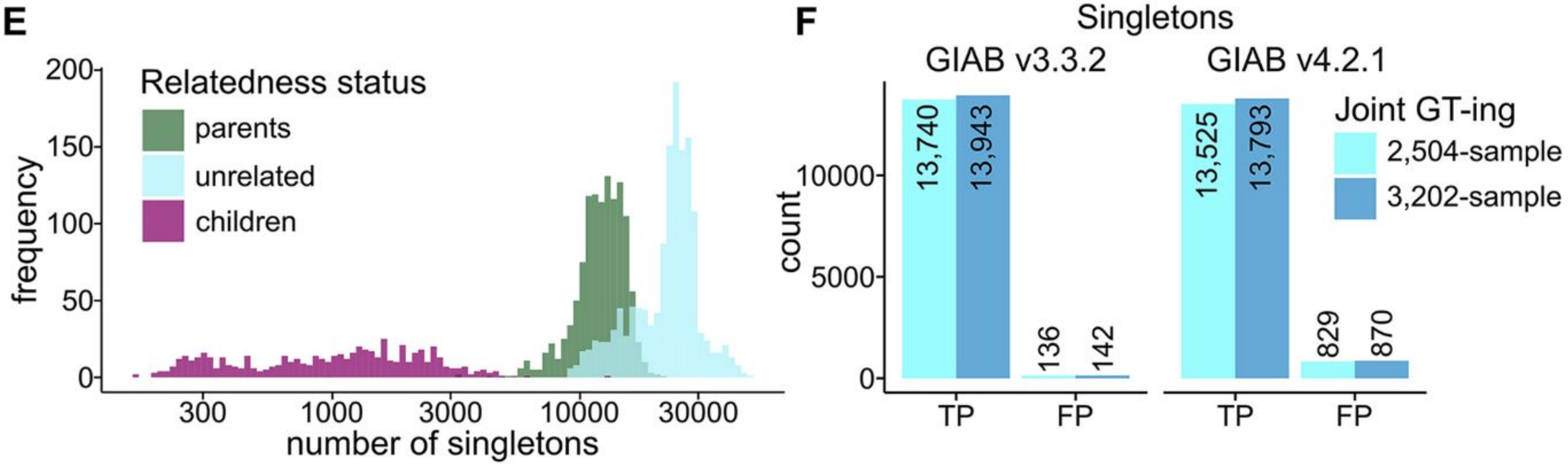
- **Children**: 100% of them are shared with parents (i.e. are not counted as singletons).
- **Parents**: 50% shared with children (i.e. 50% are counted as singletons).
- **Unrelated**: all counted as singletons.



Accumulation of somatic *de novos:* variability across cell lines likely dependent on age of the cell line.

Zook et al., 2019, Wagner et al., 2021.

~5% of singleton calls appear to be truly present in the cell lines but may not represent true population variants or even real DNMs in the original donors



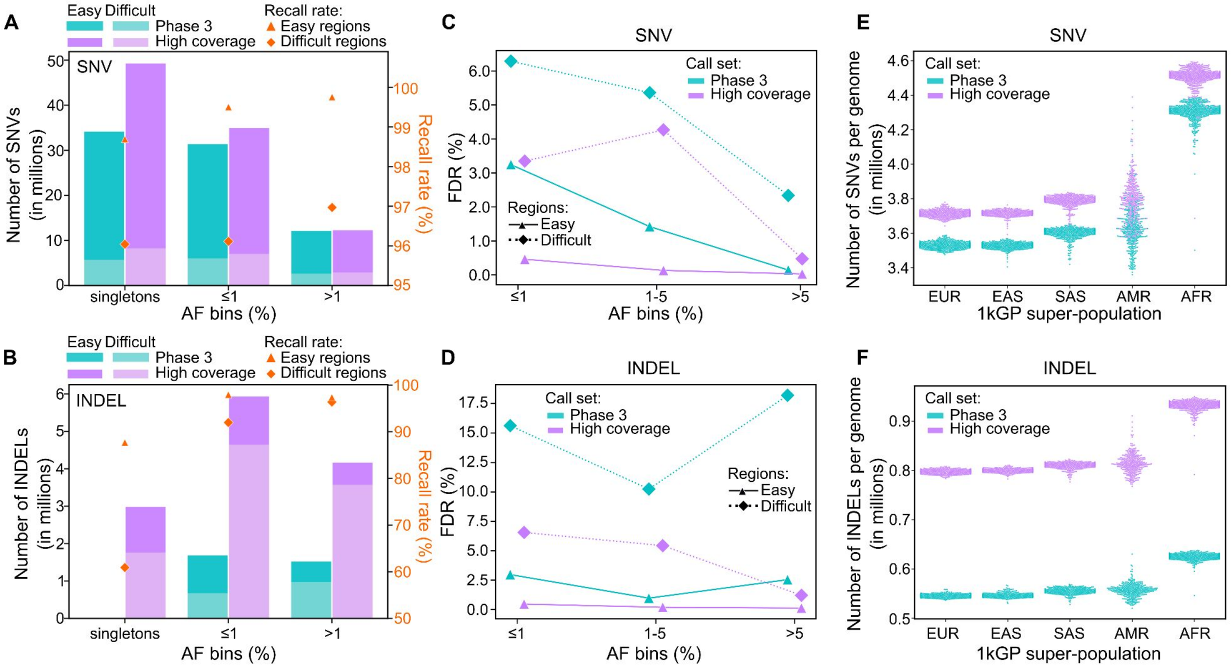
FDR among singletons:

- **1.01%** (GIAB v3.3.2);
- **5.93%** (GIAB v4.2.1, which excludes some of the mosaic variants).

Zook et al., 2019, Wagner et al., 2021.

Discovered more rare SNVs and more INDELs across the frequency spectrum

- **1.24-fold cohort-level increase** in the number of SNVs and **4.05-fold increase** in the number of INDELs compared to the phase 3 call set across the 2,504 shared samples.
- **1.05-fold average per-sample increase** in the number of SNVs and **1.47-fold increase** in the number of INDELs in the high-coverage call set.
- Discovered **more non-coding/regulatory SNVs** as well as **coding & non-coding INDELs**.



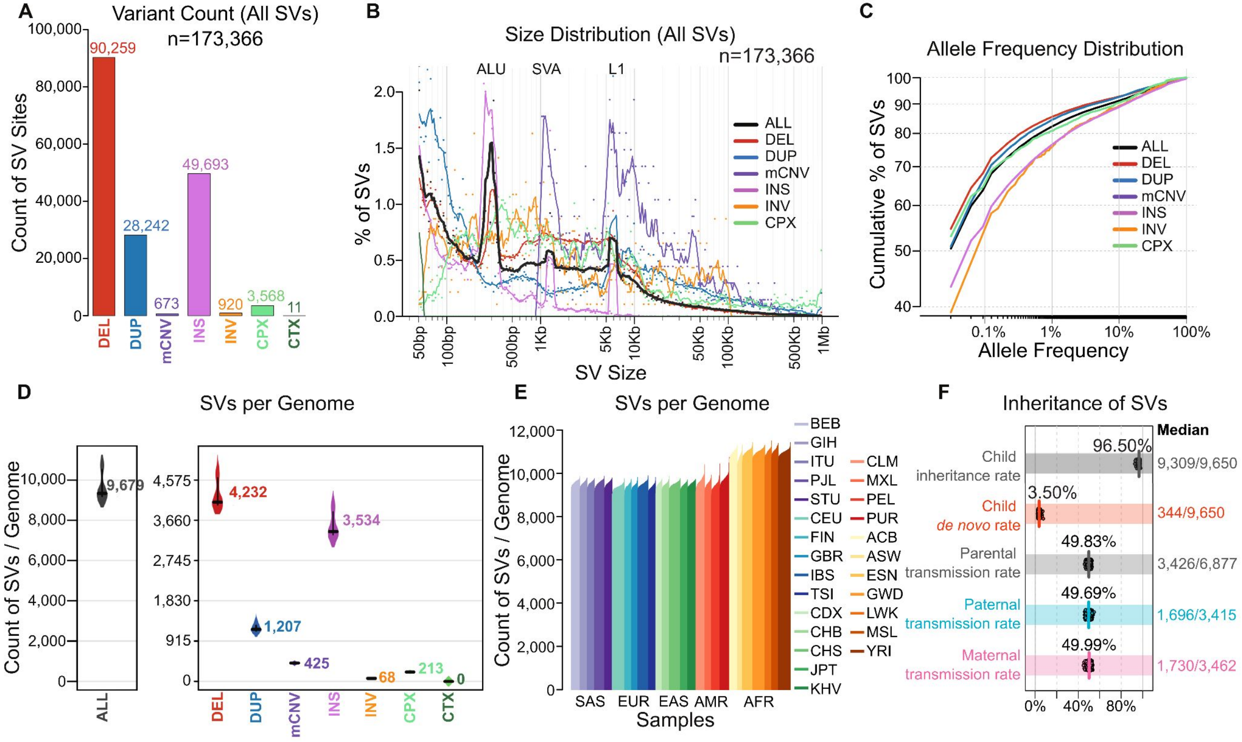
FDR (%):

Variant type	Phase 3	High coverage
SNV	0.60	0.10
INDEL	12.40	1.10

Outline

1. Small variant discovery.
2. Structural variant discovery.
3. Generation of an integrated reference imputation panel.

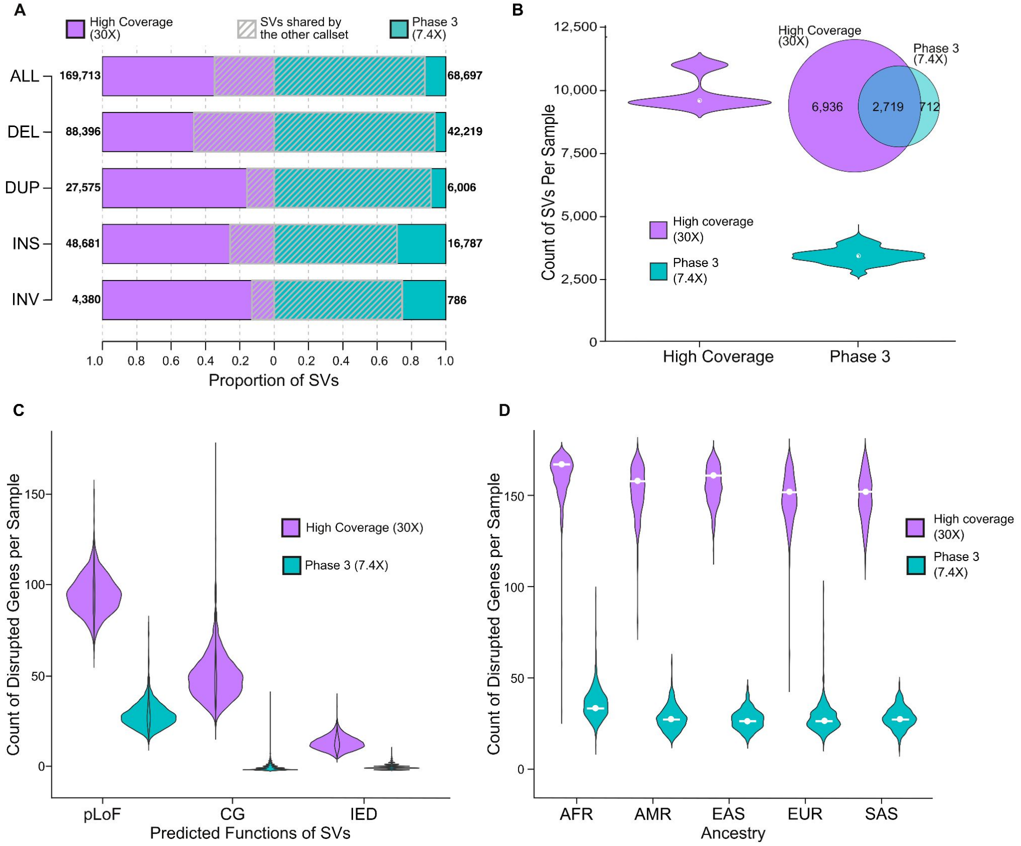
SV discovery using multiple algorithms and analytic pipelines



SV call set integrated from GATK-SV, svtools, and Absinthe:

- A total of 173,366 SV sites across 3,202 samples in the high-coverage call set.
- An average of 9,679 SVs per genome.
- More SVs are observed in African ancestry group.

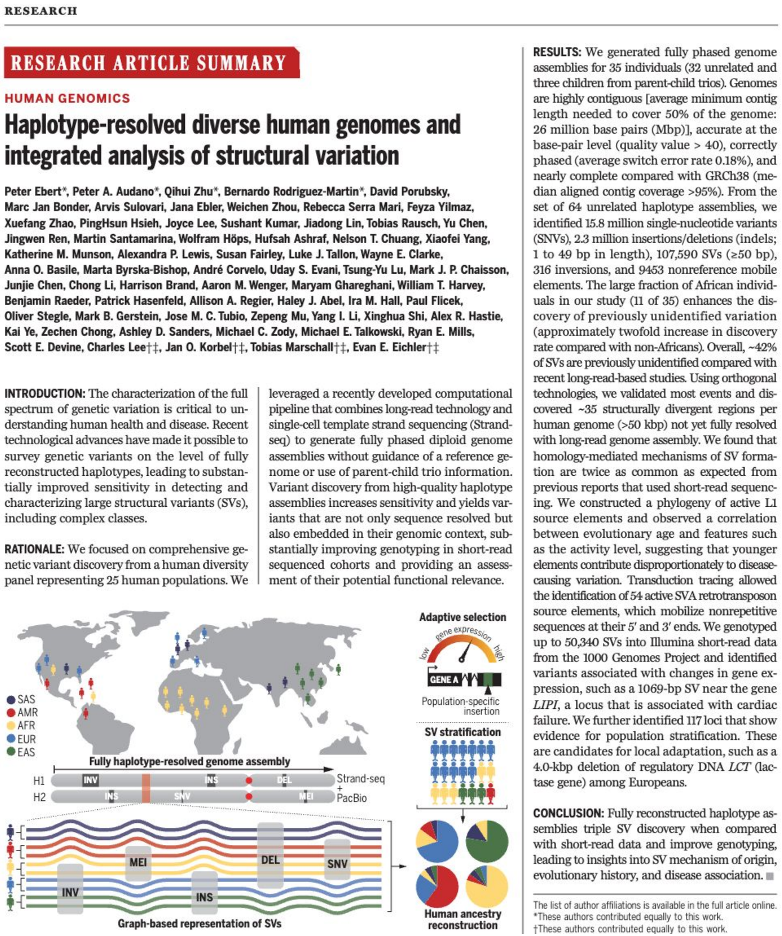
> 2-fold greater power for SV discovery compared to phase 3



- **2.5-fold increase in SV sites at the cohort-level** in the high-coverage vs. phase 3 call set (169,713 vs. 68,697).
- **2.8-fold average increase in SVs per sample** (9,655 vs. 3,431).
- **5.0-fold average increase in genes altered by SVs** in the high-coverage call set than phase 3 (162 vs. 32).
- More genes are altered in AFR population than others.

CG: complete copy gain; IED: duplication of intragenic exons.

How much are we still missing? Comparison to long-read data



RESEARCH

RESEARCH ARTICLE SUMMARY

HUMAN GENOMICS

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

Peter Ebert*, Peter A. Audano*, Qihui Zhu*, Bernardo Rodriguez-Martin*, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, Feyza Yilmaz, Xuefang Zhao, PingHsun Hsieh, Joyce Lee, Sushant Kumar, Jiadong Lin, Tobias Rausch, Yu Chen, Jingwen Ren, Martin Santamarina, Wolfram Höps, Hufsah Ashraf, Nelson T. Chuang, Xiaofei Yang, Katherine M. Munson, Alexandra P. Lewis, Susan Fairley, Luke J. Tallon, Wayne E. Clarke, Anna O. Basile, Marta Byrska-Bishop, André Corvelo, Uday S. Evani, Tsung-Yu Lu, Mark J. P. Chaisson, Junjie Chen, Chong Li, Harrison Brand, Aaron M. Wenger, Maryam Ghareghani, William T. Harvey, Benjamin Raeder, Patrick Hasenfeld, Allison A. Regier, Haley J. Abel, Ira M. Hall, Paul Flicek, Oliver Stegle, Mark B. Gerstein, Jose M. C. Tubio, Zepeng Mu, Yang I. Li, Xinghua Shi, Alex R. Hastie, Kai Ye, Zechen Chong, Ashley D. Sanders, Michael C. Zody, Michael E. Talkowski, Ryan E. Mills, Scott E. Devine, Charles Lee†‡, Jan O. Korbel†‡, Tobias Marschall†‡, Evan E. Eichler†‡

INTRODUCTION: The characterization of the full spectrum of genetic variation is critical to understanding human health and disease. Recent technological advances have made it possible to survey genetic variants on the level of fully reconstructed haplotypes, leading to substantially improved sensitivity in detecting and characterizing large structural variants (SVs), including complex classes.

RATIONALE: We focused on comprehensive genetic variant discovery from a human diversity panel representing 25 human populations. We leveraged a recently developed computational pipeline that combines long-read technology and single-cell template strand sequencing (Strand-seq) to generate fully phased diploid genome assemblies without guidance of a reference genome or use of parent-child trio information. Variant discovery from high-quality haplotype assemblies increases sensitivity and yields variants that are not only sequence resolved but also embedded in their genomic context, substantially improving genotyping in short-read sequenced cohorts and providing an assessment of their potential functional relevance.

RESULTS: We generated fully phased genome assemblies for 35 individuals (32 unrelated and three children from parent-child trios). Genomes are highly contiguous [average minimum contig length needed to cover 50% of the genome: 26 million base pairs (Mbp)], accurate at the base-pair level (quality value > 40), correctly phased (average switch error rate 0.18%), and nearly complete compared with GRCh38 (median aligned contig coverage >95%). From the set of 64 unrelated haplotype assemblies, we identified 15.8 million single-nucleotide variants (SNVs), 2.3 million insertions/deletions (indels; 1 to 49 bp in length), 107,590 SVs (≥50 bp), 316 inversions, and 9453 nonreference mobile elements. The large fraction of African individuals in our study (11 of 35) enhances the discovery of previously unidentified variation (approximately twofold increase in discovery rate compared with non-Africans). Overall, ~42% of SVs are previously unidentified compared with recent long-read-based studies. Using orthogonal technologies, we validated most events and discovered ~35 structurally divergent regions per human genome (>50 kbp) not yet fully resolved with long-read genome assembly. We found that homology-mediated mechanisms of SV formation are twice as common as expected from previous reports that used short-read sequencing. We constructed a phylogeny of active L1 source elements and observed a correlation between evolutionary age and features such as the activity level, suggesting that younger elements contribute disproportionately to disease-causing variation. Transduction tracing allowed the identification of 54 active SVA retrotransposon source elements, which mobilize nonrepetitive sequences at their 5′ and 3′ ends. We genotyped up to 50,340 SVs into Illumina short-read data from the 1000 Genomes Project and identified variants associated with changes in gene expression, such as a 1069-bp SV near the gene *LIPI*, a locus that is associated with cardiac failure. We further identified 117 loci that show evidence for population stratification. These are candidates for local adaptation, such as a 4.0-kbp deletion of regulatory DNA *LCT* (lactase gene) among Europeans.

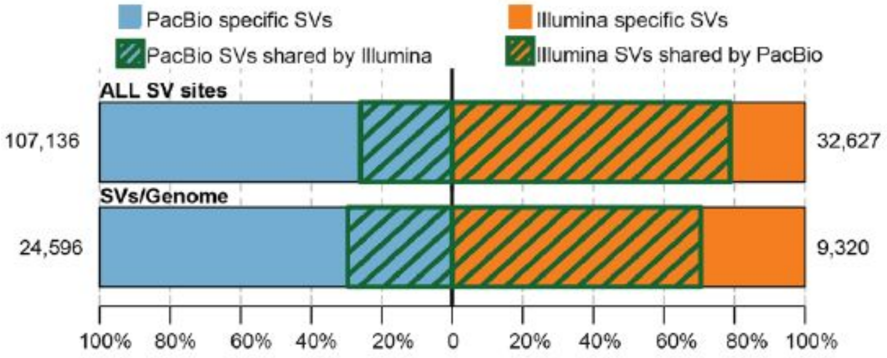
CONCLUSION: Fully reconstructed haplotype assemblies triple SV discovery when compared with short-read data and improve genotyping, leading to insights into SV mechanism of origin, evolutionary history, and disease association. ■

SAS
AMR
AFR
EUR
EAS
Fully haplotype-resolved genome assembly
H1
H2
INV
INS
DEL
MEI
SNV
Strand-seq + PacBio
Graph-based representation of SVs
Adaptive selection
gene expression
low
high
GENE A
Population-specific insertion
SV stratification
Human ancestry reconstruction

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†These authors contributed equally to this work.

Comparing 31 Illumina genomes to the same genomes done with PacBio:

- < 30% of PacBio discovered events are found by Illumina overall and by genome
- > 70% of Illumina discovered events are found by PacBio overall and by genome

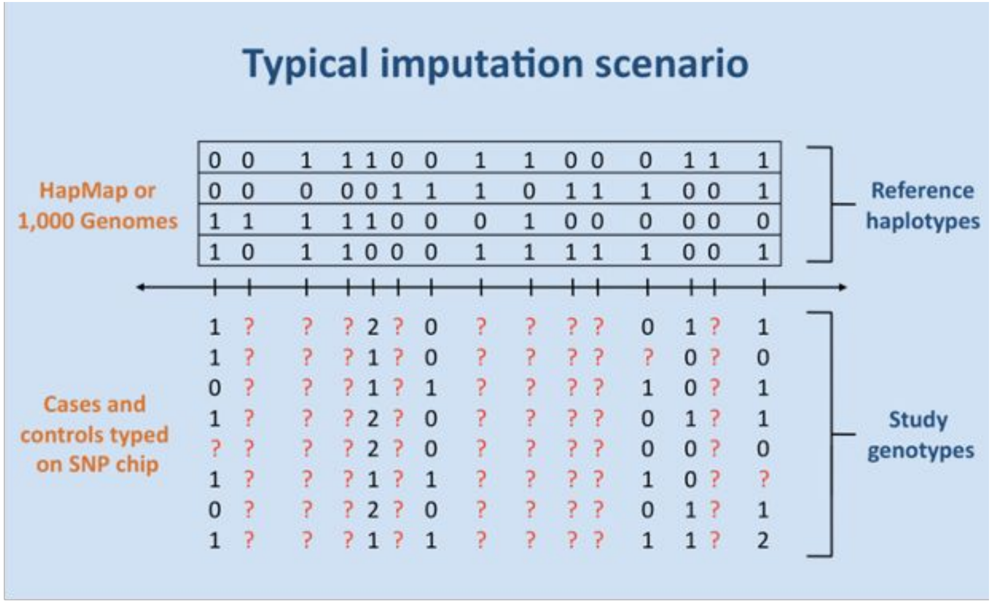


Ebert et al. Science, 2021.

Outline

1. Small variant discovery.
2. Structural variant discovery.
3. Generation of an integrated reference imputation panel.

Imputation increases discovery power of genome-wide association studies (GWAS)



Typical imputation scenario

HapMap or 1,000 Genomes

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1

Reference haplotypes

Cases and controls typed on SNP chip

1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

Study genotypes

http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

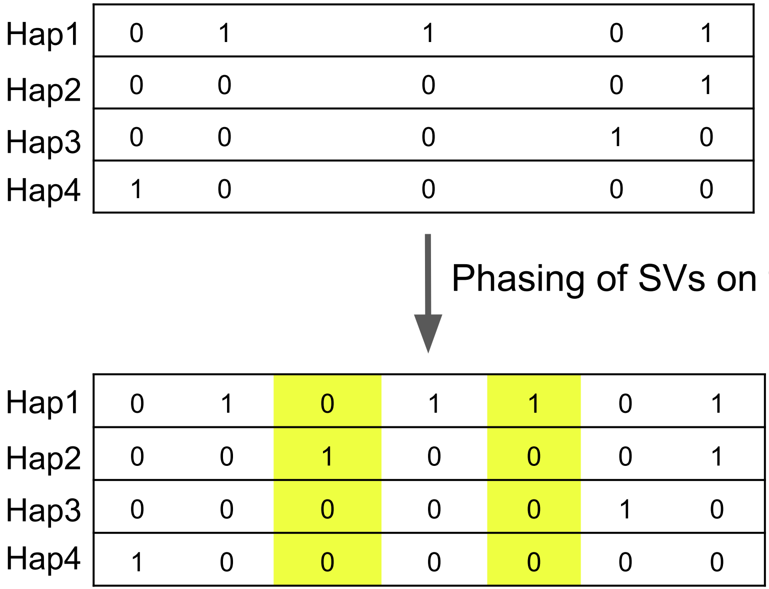
Imputation = statistical inference of unobserved genotypes in sparse genotyping array data using a reference panel based typically on WGS

Challenges associated with inclusion of SVs in the reference panel

- Most existing reference panels, such as HRC or TOPMed, do not include SVs due to challenges with SV calling and GT-ing.
- Lack of well-established truth sets for SV genotyping and phasing accuracy evaluations.
    - Haplotype-resolved LR data now available on 34 1kGP samples from Ebert et al. 2021.
    - Inclusion of trios allows us to use inheritance patterns to evaluate quality of GT-ing and phasing.

2-step process of haplotype phasing

- **73,452,337 SNV/INDELs and 102,459 SVs (DELs, INSs, DUPs, and INVs)** included in the phased panel (filtering criteria: PASS, missingness < 5%, HWE PASS, MER ≤ 5%, MAC ≥ 2).
- **STEP 1**: Phasing of SNVs/INDELs was performed using statistical phasing with pedigree-based correction (SHAPEIT2-duohmm) across autosomes (chrX was phased using Eagle2).
- **STEP 2**: SVs were phased on top of the SNV/INDEL haplotype scaffold using SHAPEIT4 v4.2.2.



Hap1	0	1	1	0	1
Hap2	0	0	0	0	1
Hap3	0	0	0	1	0
Hap4	1	0	0	0	0

SNV/INDEL scaffold built using SHAPEIT2-duohmm

Phasing of SVs on top of the SNV/INDEL scaffold with SHAPEIT4

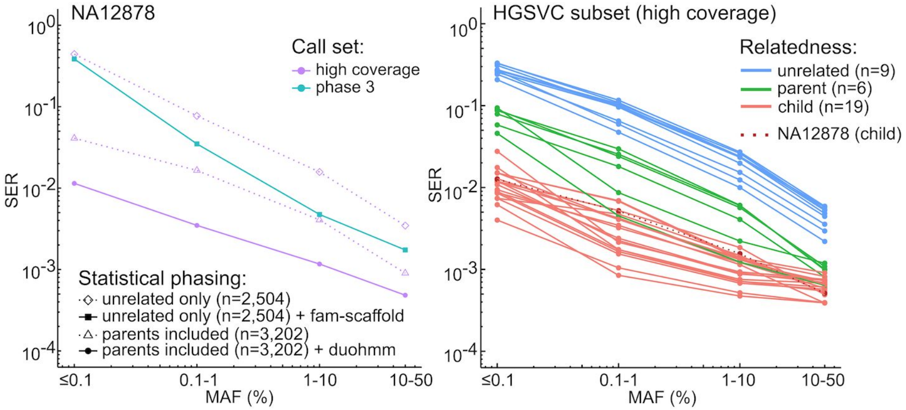
Hap1	0	1	0	1	1	0	1
Hap2	0	0	1	0	0	0	1
Hap3	0	0	0	0	0	1	0
Hap4	1	0	0	0	0	0	0

Reference imputation panel consisting of phased SNVs, INDELs, and SVs.

Delaneau et al. 2019, Delaneau et al. 2011; O'Connell et al. 2014; Loh et al. 2016.

Superior SNV/INDEL phasing accuracy & imputation performance of the high-coverage panel compared to phase 3

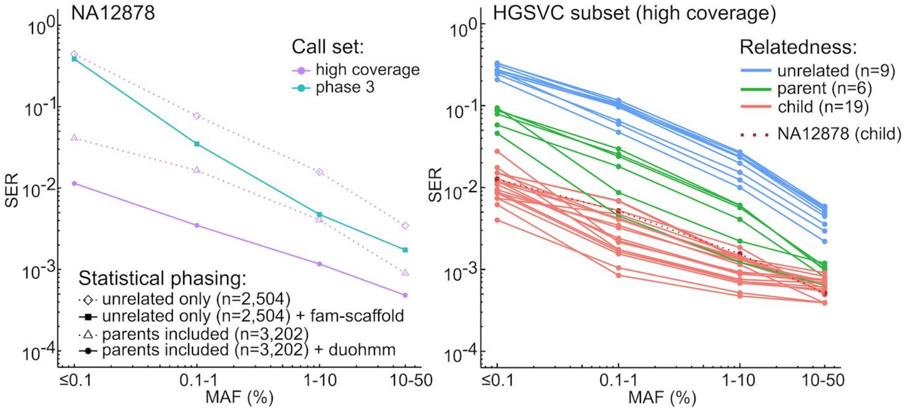
Phasing accuracy evaluation:



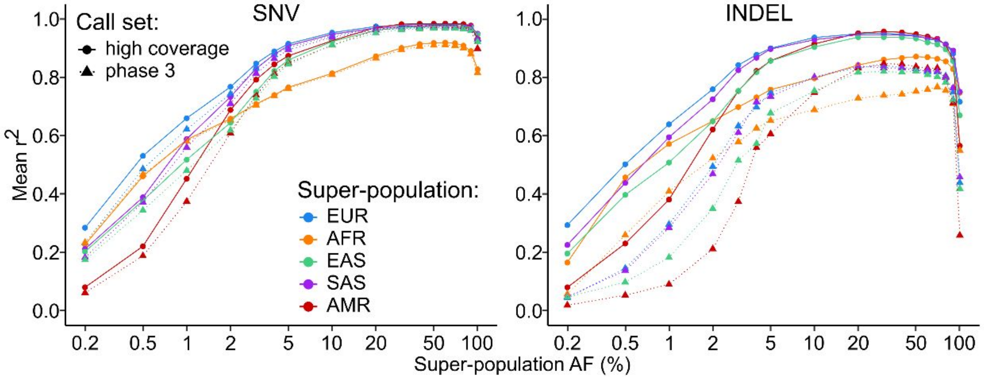
- Up to 10-fold higher SNV/INDEL phasing accuracy in the high-coverage vs. phase 3 panel (autosomal SER=0.07% vs. 0.76%).
- Average autosomal SER in the high coverage panel:
  - Children: 0.09%
  - Parents: 0.22%
  - Unrelated: 0.79%
- Parental and unrelated samples showed 2.2-fold and 1.3-fold average improvement, respectively, relative to phase 3.

Superior SNV/INDEL phasing accuracy & imputation performance of the high-coverage panel compared to phase 3

Phasing accuracy evaluation:



Imputation accuracy evaluation:



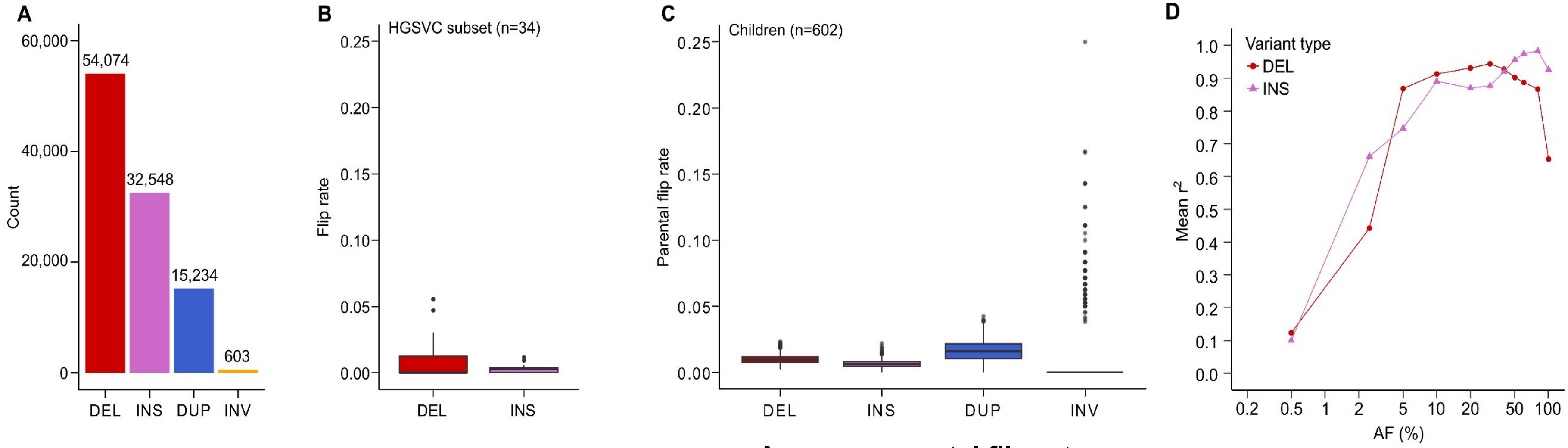
- Up to 10-fold higher SNV/INDEL phasing accuracy in the high-coverage vs. phase 3 panel (autosomal SER=0.07% vs. 0.76%).
- Average autosomal SER in the high coverage panel:
  - Children: 0.09%
  - Parents: 0.22%
  - Unrelated: 0.79%
- Parental and unrelated samples showed 2.2-fold and 1.3-fold average improvement, respectively, relative to phase 3.

- SNV imputation performance was comparable across the panels.
- Imputation of INDELs with the high-coverage panel displayed superior accuracy across all five super-population ancestry groups across the entire AF spectrum.

SVs show high phasing accuracy and imputation accuracy comparable to small variants at MAF > 5% but lower at rarer MAF bins

Phasing accuracy evaluation using 2 approaches:

Imputation accuracy evaluation:



Average flip rate:

- 0.89% for DELs,
- 0.24% for INSs.

Average parental flip rate:

- 0.99% for DELs,
- 0.65% for INS,
- 1.63% for DUPs,
- 1.20% for INV.

Conclusions

- Expanded the 1kGP cohort to include 602 trios.
- Upgraded the sequencing to high-coverage WGS.
- Discovered more rare non-coding SNVs and substantially more coding and non-coding INDELs and SVs across the frequency spectrum.
- Generated an improved reference imputation panel which makes variants discovered here accessible for association studies.
- All data publicly available without restriction at IGSR FTP, EBI-EMBL, dbSNP, dbVAR.

Acknowledgements

Marta Byrska-Bishop*
Uday S. Evani*
Xuefang Zhao*
Anna O. Basile
Haley J. Abel
Allison A. Regier
André Corvelo
Wayne E. Clarke
Rajeeva Musunuri
Kshithija Nagulapalli
Susan Fairley
Alexi Runnels
Lara Winterkorn
Ernesto Lowy

HGSVC (Charles Lee, Evan E. Eichler, Jan O. Korbel et al.)
Paul Flicek
Soren Germer
Harrison Brand
Ira M. Hall
Michael E. Talkowski
Giuseppe Narzisi

Yale University
School of Medicine

EMBL-EBI

MGH
1811

HGSVC

NEW YORK
GENOME CENTER®

BROAD
INSTITUTE

Generation of the 1kGP reference imputation panel including PanGenie SV and INDEL calls

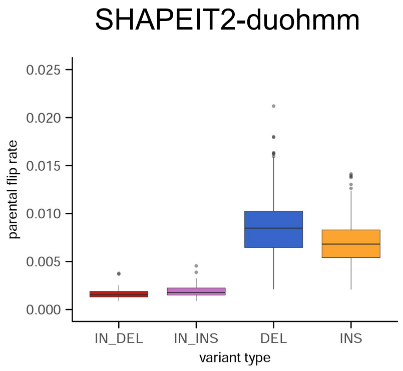
Preliminary analysis:

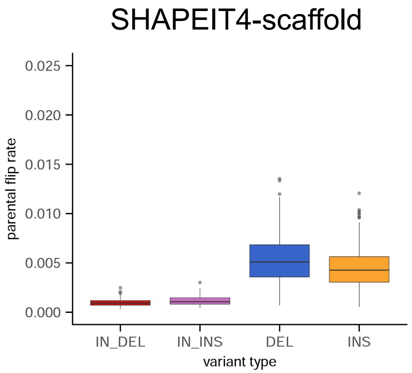
- Integrated the 1kGP PanGenie strict call set (DELs, INSs, and INDELs) from Ebert et al. 2021 with the non-singleton high-quality SNV subset of the high-coverage 1kGP call set from Byrska-Bishop et al. 2022.
- Performed haplotype phasing of SNVs, SVs, and INDELs using statistical phasing with pedigree-based correction (SHAPEIT2-duohmm) and evaluated phasing accuracy by computing parental flip rate of phased HET GTs across 602 children samples (see table below).

Future plans:

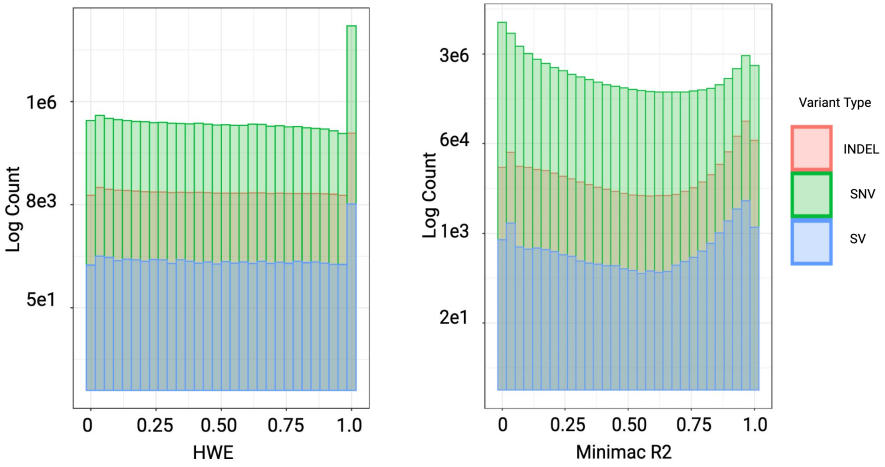
- Switch to a 2-step phasing approach, in which SVs and INDELs are phased on top of the previously-phased SNV scaffold (SHAPEIT4-scaffold), which results in a slightly better phasing accuracy and substantially lower computational cost (~5-10-fold faster run time).

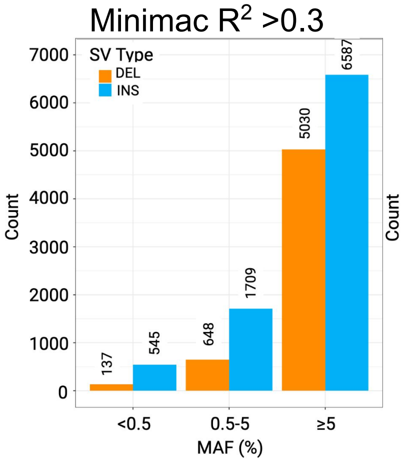
Variant type	Approach for phasing accuracy estimation	SHAPEIT2 duohmm	SHAPEIT4 scaffold
SNV	SER (n=1; truth set: PG NA12878)	0.0008	0.0008
IN-DEL	Mean parental flip rate (n=602)	0.0016	0.0010
IN-INS	Mean parental flip rate (n=602)	0.0019	0.0011
DEL	Mean parental flip rate (n=602)	0.0086	0.0054
INS	Mean parental flip rate (n=602)	0.0069	0.0044

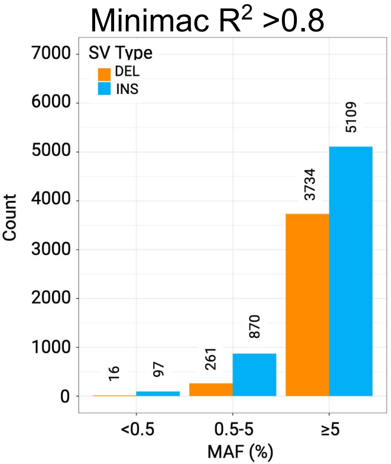

SHAPEIT2-duohmm


SHAPEIT4-scaffold

SV Imputation in UK Biobank using the Integrated Reference Panel




Minimac R^2 >0.3

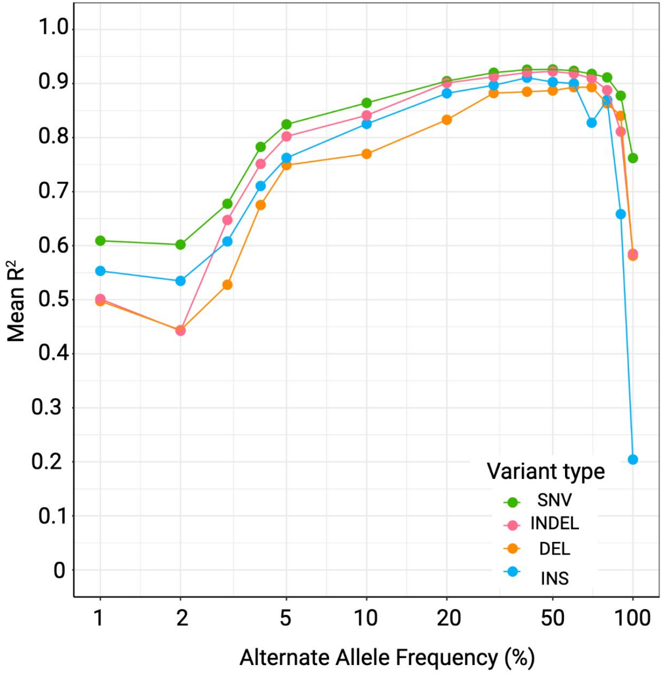

Minimac R^2 >0.8

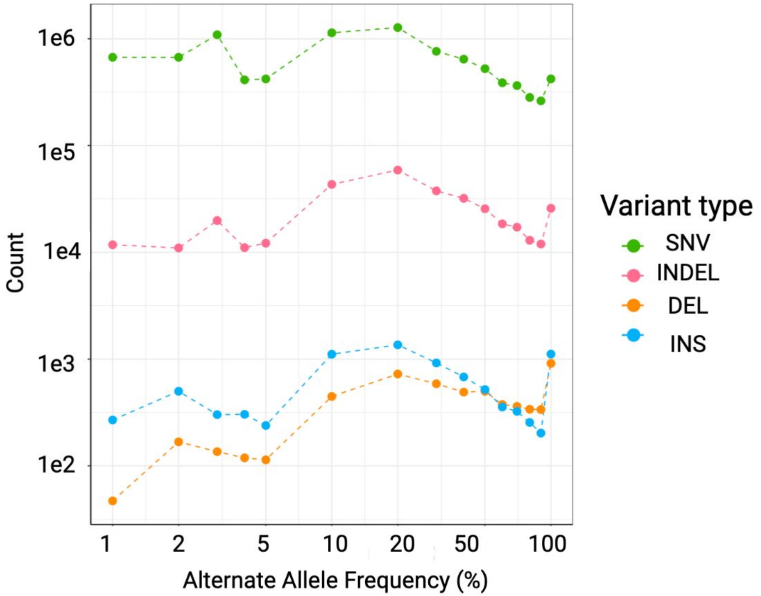
- Imputed 342,334 genotyped UK Biobank samples
- SVs observe comparable HWE distributions to SNVs/INDELS
  - 0.016% SVs, 0.014% INDELs, 0.008% SNVs HWE p<1e-10
- Rarer SNVs are imputed more accurately than SVs and INDELs
  - 35% SVs, 39% INDELS, 85% SNVs with AF < 1% were imputed

Variant Type	Minimac R^2	Count (% of total variant type)
SNV	0.3	20,018,920 (33.4)
INDEL	0.3	501,693 (73)
SV	0.3	16,032 (70.6) *DEL: 6,375* *INS: 9,657*

Common (AF >5%) SVs are Accurately Imputed in the UK Biobank (UKB)

- Empirical imputation accuracy evaluations were performed on 50 UKB samples.
- The SV truth set was generated by genotyping DELs and INSs from the HGSVC strict call set using PanGenie.
- SVs are imputed with comparable accuracy to SNVs at AF ≥ 5%:
    - DELs (mean R^2=0.75 +/- 0.12)
    - INSs (mean R^2=0.76 +/- 0.09)
    - SNVs (mean R^2=0.82+/- 0.03)

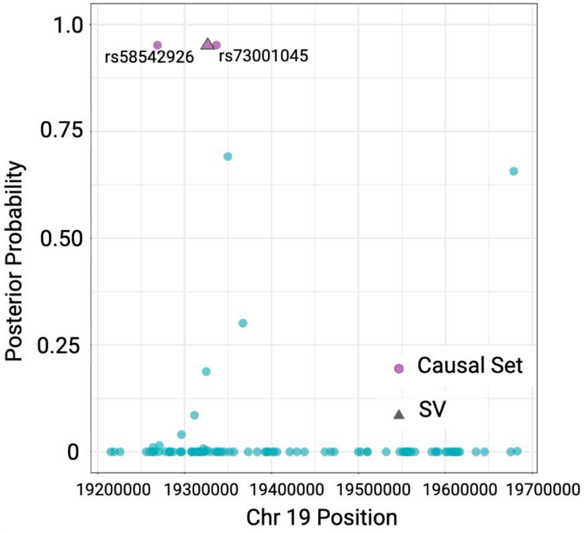


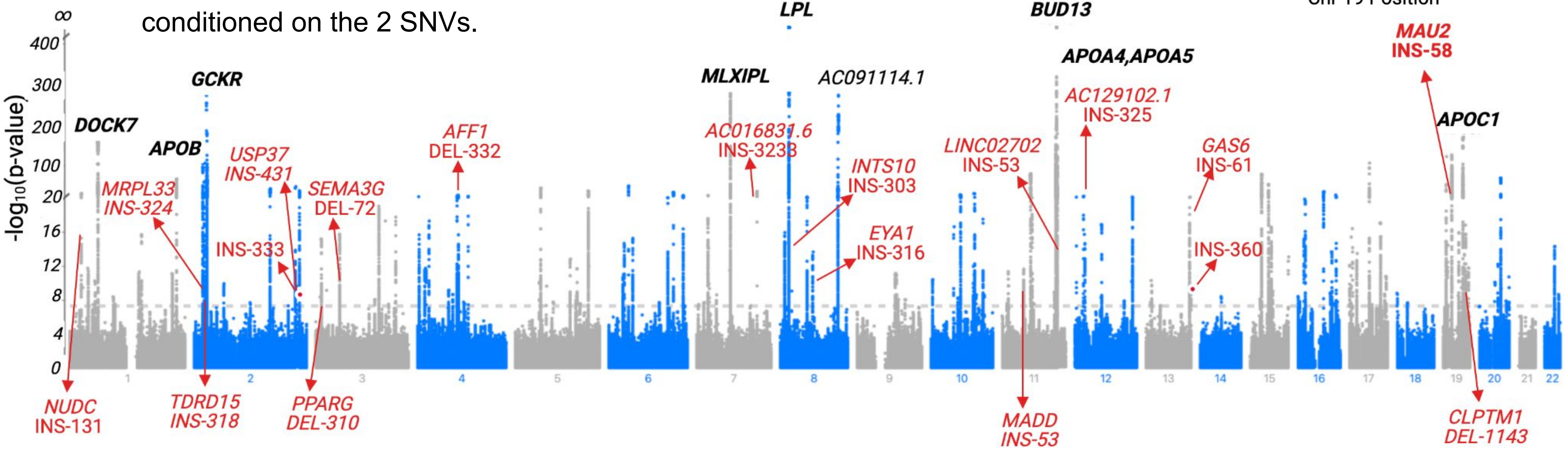


Lipid Trait GWAS identifies significant SVs

- 17 significant SVs with Bonferroni-corrected p-value <1.7e-9
- Top SV hit: *chr19:19326707-INS-58* in *MAU2*
    - P-value=5.5e-27, AF=64.7%, Beta=0.031

- FINEMAP identified this SV and 2 strongly correlated SNVs as potentially causal with ~98% posterior probability.
    - 99% posterior probability of ≥1 putative causal signal within 3 variant set.
    - 96% posterior probability of SV being likely causal when conditioned on the 2 SNV signals.
    - *chr19-19326707-INS-58* remained significant (P-value=4.9e-15) when conditioned on the 2 SNVs.


Posterior Probability
rs58542926
rs73001045
Causal Set
SV
Chr 19 Position


∞
-log10(p-value)
DOCK7
APOB
GCKR
MLXIPL
LPL
AC091114.1
BUD13
APOA4,APOA5
APOC1
MAU2
INS-58
AC129102.1
INS-325
AC016831.6
INS-3233
AFF1
DEL-332
USP37
INS-431
SEMA3G
DEL-72
MRPL33
INS-324
INS-333
INTS10
INS-303
EYA1
INS-316
LINC02702
INS-53
GAS6
INS-61
INS-360
NUDC
INS-131
TDRD15
INS-318
PPARG
DEL-310
MADD
INS-53
CLPTM1
DEL-1143