Population sequencing and analysis projects at NYGC

Michael C. Zody

Scientific Director, Computational Biology, New York Genome Center

November 11, 2021

NEW YORK
GENOME CENTER℠

OVERVIEW

- Summary of large scale genome projects
  - TOPMed
  - CCDG
  - 1000 Genome high coverage
- Methods development for large scale projects
  - Absinthe insertion detector
  - Structural variant phasing and imputation

TOPMED

- Trans-Omics for Precision Medicine
- NHLBI project to create resources for deeply phenotyped cohorts
    - Whole genome sequencing for >130,000 samples of diverse ancestry
    - RNA-Seq, metabolomics, proteomics
- Flagship paper in Nature this year (Taliun et al., 2021)
    - Analyzed >53,000 genomes
    - >400M variants discovered (~50% singletons)
    - Imputation panel with >97,000 genomes
    - Discovery of >1000 non-reference sequences from AC = 1 to 100% AF

CCDG

- Centers for Common Disease Genomics
- NHGRI project to develop paradigms for understanding genetic architecture of common disease
  - Whole genome sequencing for >130,000 samples of diverse ancestry
  - Exome sequencing for an additional 198,000 samples
- Phenotypes include ASD, epilepsy, heart disease, stroke, IBD

CCDG ANALYSIS PLANS

- Whole genome sequencing now complete
- Final (“Freeze 3”) data set called
- Joint SNV/indel calling with GATK (Broad)
- Distributed SV call set (WashU, Baylor, NYGC)
  - Lumpy (deletions and inversions)
  - Absinthe (insertions)
  - Canvas + QuicKmer2 (depth of coverage/copy number)
  - Genotyping of long read derived variants with Paragraph
- SV calls will be genotyped on all samples and merged into a single set
- Imputation server based on the Michigan/TOPMed model
- Timeline for release in 2022

DATA AVAILABILITY FROM TOPMED AND CCDG

- Both projects intend to broadly share data
- Both projects consist of collections of older cohorts with a wide variety of patient consents ranging from general research to disease specific
- TOPMed data are currently available on a per cohort basis from dbGaP and BioData Catalyst
- CCDG data will be publicly available on AnVIL (access controlled through dbGaP)
- Imputation servers will be available as a service only (no downloadable panels) due to access restrictions

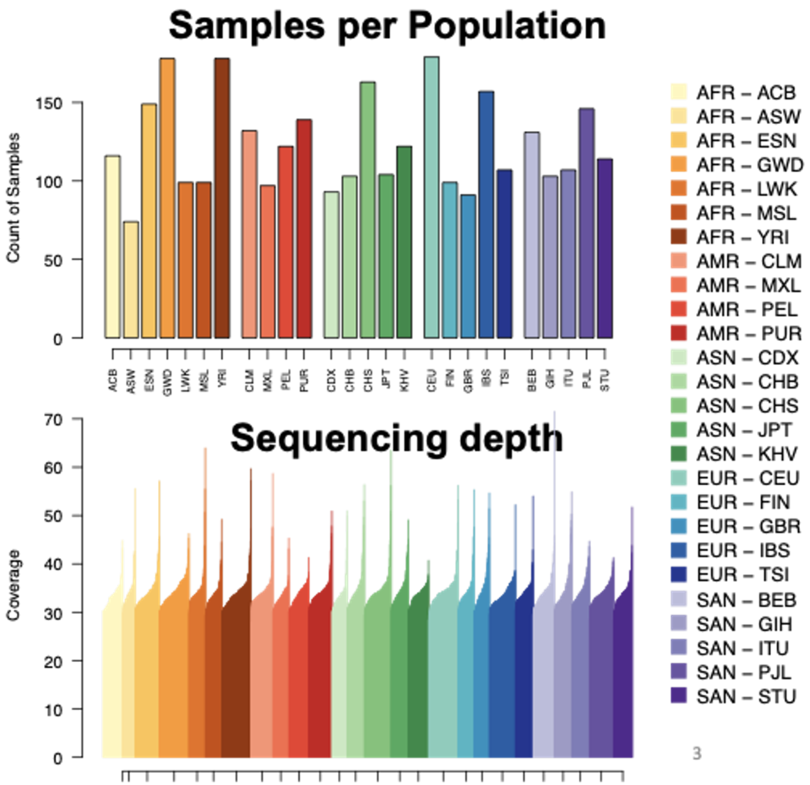
1000 GENOMES PROJECT SEQUENCING

- Supplement to CCDG
- 30x Illumina of all 2,504 phase 3 samples
- Additional 698 sample sequenced to complete 602 trios
- GATK joint calling for SNVs/indels
- Comprehensive combined SV calling from the HGSVC
- All data released through EBI/ISGR and NCBI:

  https://www.internationalgenome.org/data-portal/data-collection/30x-grch38
- Data are also available on AnVIL (Google cloud) and AWS
- Preprint up on biorxiv (Byrska-Bishop, Evani, Zhao, *et al.*, 2021)

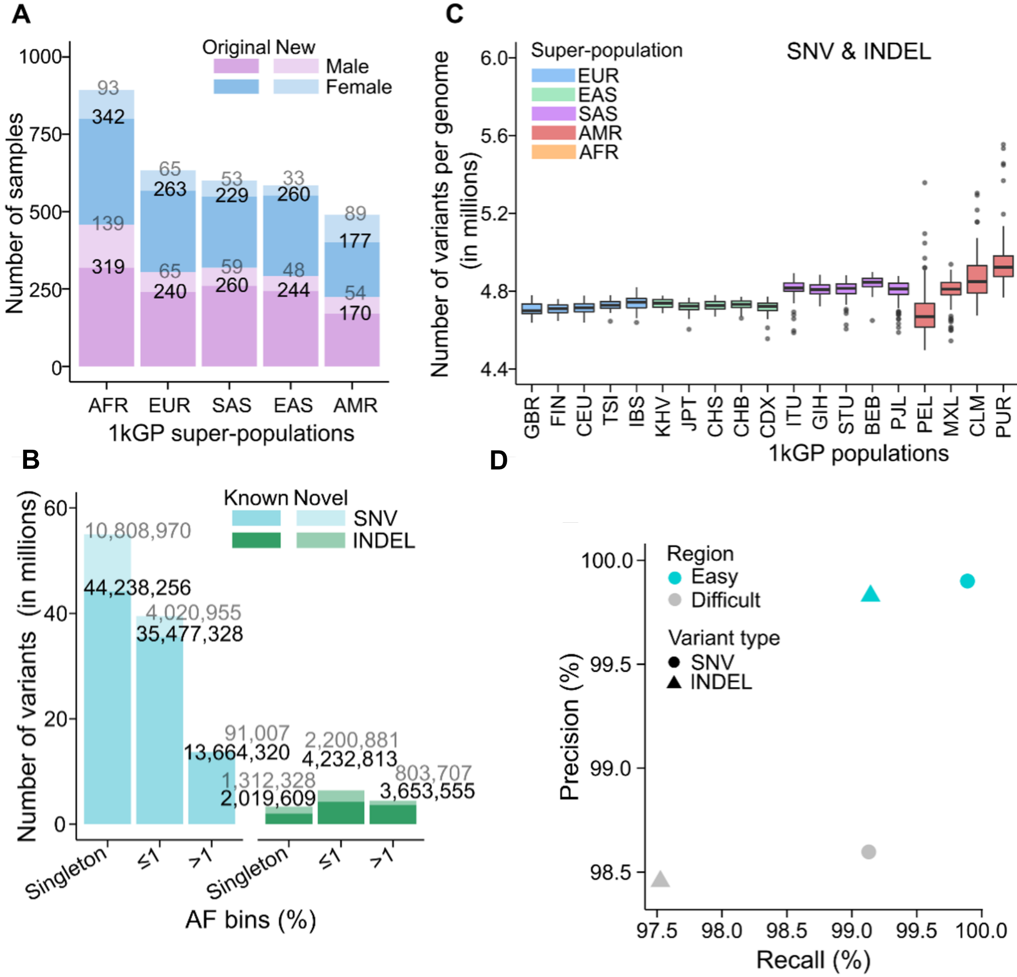
1000 GENOMES OVERVIEW

- ➢ 3,202 genomes (2,504 original + 698 new) collected from 26 populations, including:
  - 602 complete trios
  - 6 parent-child duos
- ➢ All samples were sequenced to a targeted depth of 30X by the NYGC.
- ➢ SNVs and INDELs were discovered using GATK's HaplotypeCaller; SVs were discovered with the GATK-SV pipeline[1], the svtools pipeline[2] and Absinthe[3].
- ➢ 2,504 unrelated samples were previously sequenced to ~7.4X (phase 3 callset)[4,5].



[1] Collins *et al.* 2020. Nature
[2] Abel *et al.* 2020. Nature
[3] Corvelo A. *in prep.*
[4] The 1000 Genomes Project Consortium. 2015. Nature
[5] Sudmant *et al.* 2015. Nature

SNV/INDEL DISCOVERY



Summary stats:

	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	
Novel	14,920,932	4,316,916		

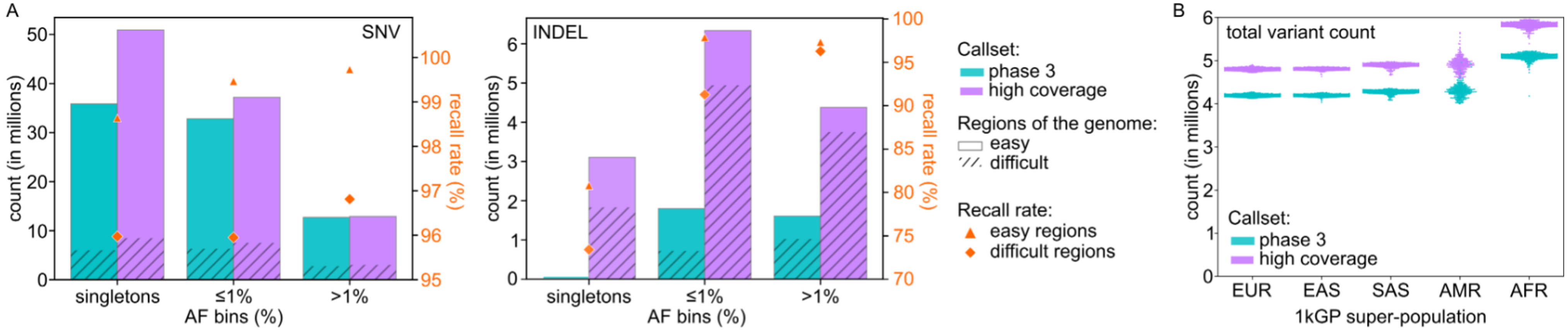
Comparison against the GIAB truth set:

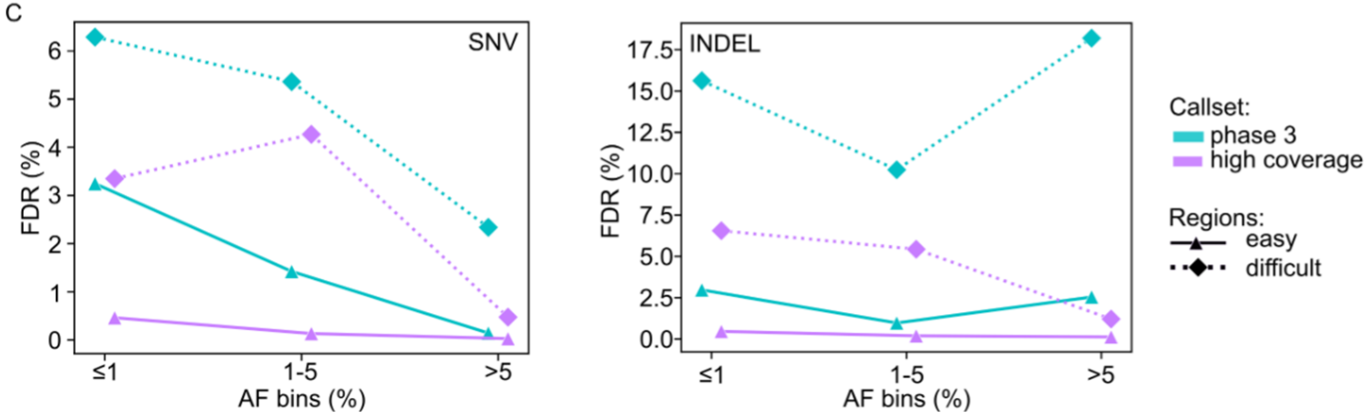
Variant type	FDR (%)
SNV	0.3
INDEL	1.15

Zook, J.M. et al. Nat. Biotechnol. *37*, 561–566 (2019)

COMPARISON TO 1KG PHASE 3

- Comparison restricted to the 2,504 samples shared between the two callsets.
- Used the GRCh38 lifted-over version of the phase 3 callset.



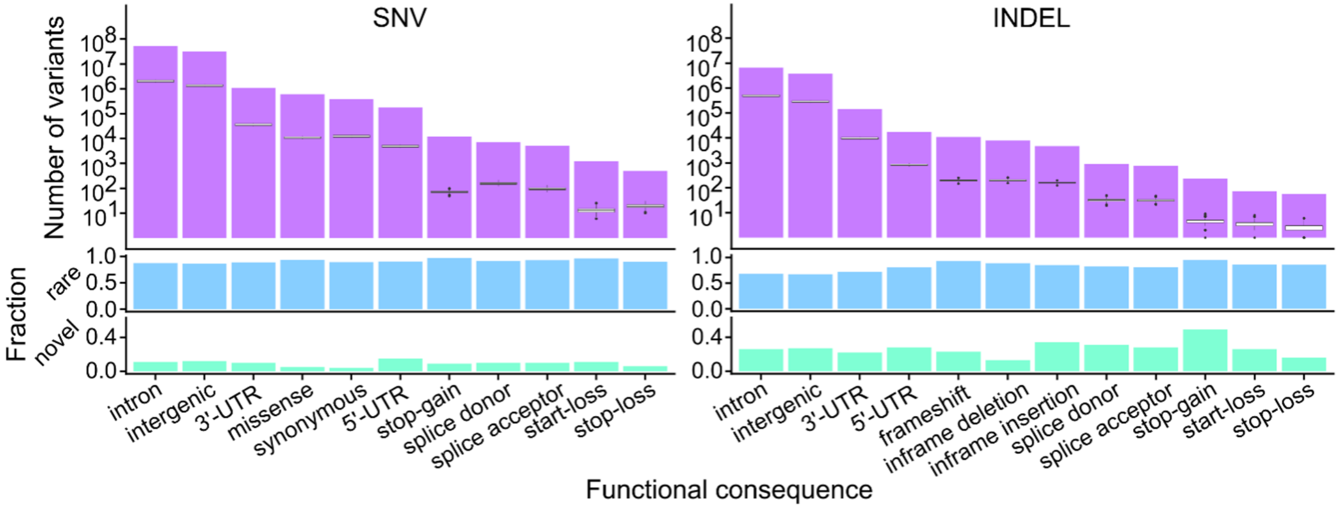


FDR (%):

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

Krusche *et al.* Nat Biotechnol 37, 555–560 (2019).

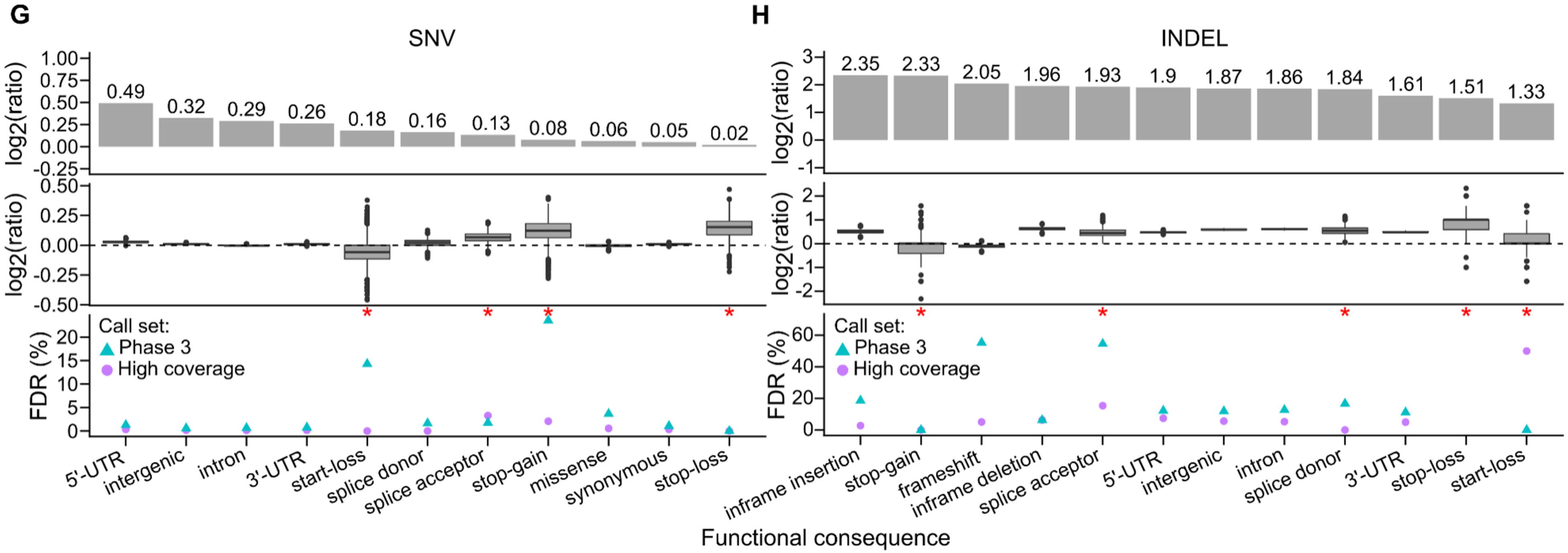
VARIANT FUNCTION PREDICTION



- **Cohort-level total**:
  - 605,896 missense mutations,
  - 384,451 synonymous mutations,
  - 36,520 predicted loss of function variants (pLOF), defined as stop gained (n=12,181), frameshift (n=10,850), and splice mutations (n=13,489).
- **Genome-level average (MAF < 1%)**:
  - 754 missense,
  - 569 synonymous,
  - 43 pLOFs (11 stop-gained, 14 frameshift, and 18 splice mutations).

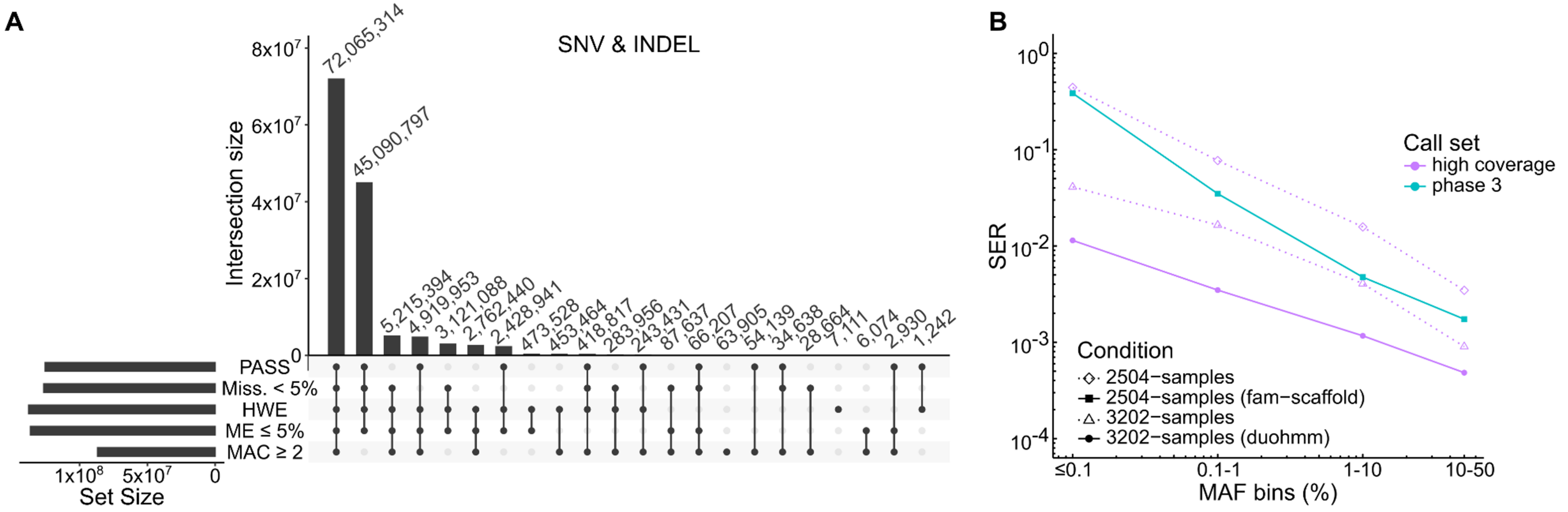
VEP v.104: McLaren W. *et al. Genome Biol* 17, 122 (2016).

COMPARISON TO 1KG PHASE 3



- **Cohort-level:**
  - SNVs: 1.01-1.41-fold increase in high coverage vs. phase 3.
  - INDELs: 2.52- and 13.48-fold increase in high coverage vs. phase 3.
- **Genome-level:**
  - SNVs: most categories show no significant difference, except for stop-gained (9% increase), stop-lost (11% increase), and start-lost (3% decrease).
  - INDELs: most categories show ~9-55% increase on average in the high coverage vs. phase 3, except for stop-gained and frameshift (3 and 7% decrease on average, respectively).

HAPLOTYPE PHASING OF SNV/INDEL



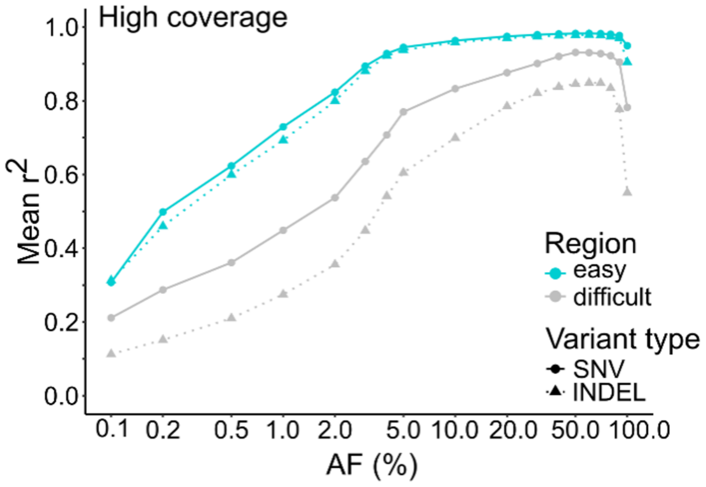
- Filtering criteria: VQSR PASS, missingness <5%, HWE PASS, ME <=5%, MAC>=2.
- Phasing performed using statistical phasing with pedigree-based correction (SHAPEIT2-duohmm) across autosomes (chrX was phased using Eagle2).

Delaneau, O. et al. Nat. Methods *9*, 179–181 (2011);
O’Connell, J. et al. PLoS Genet. *10*, e1004234 (2014);
Loh, P.-R., et al. Nat. Genet. *48*, 1443–1448 (2016).

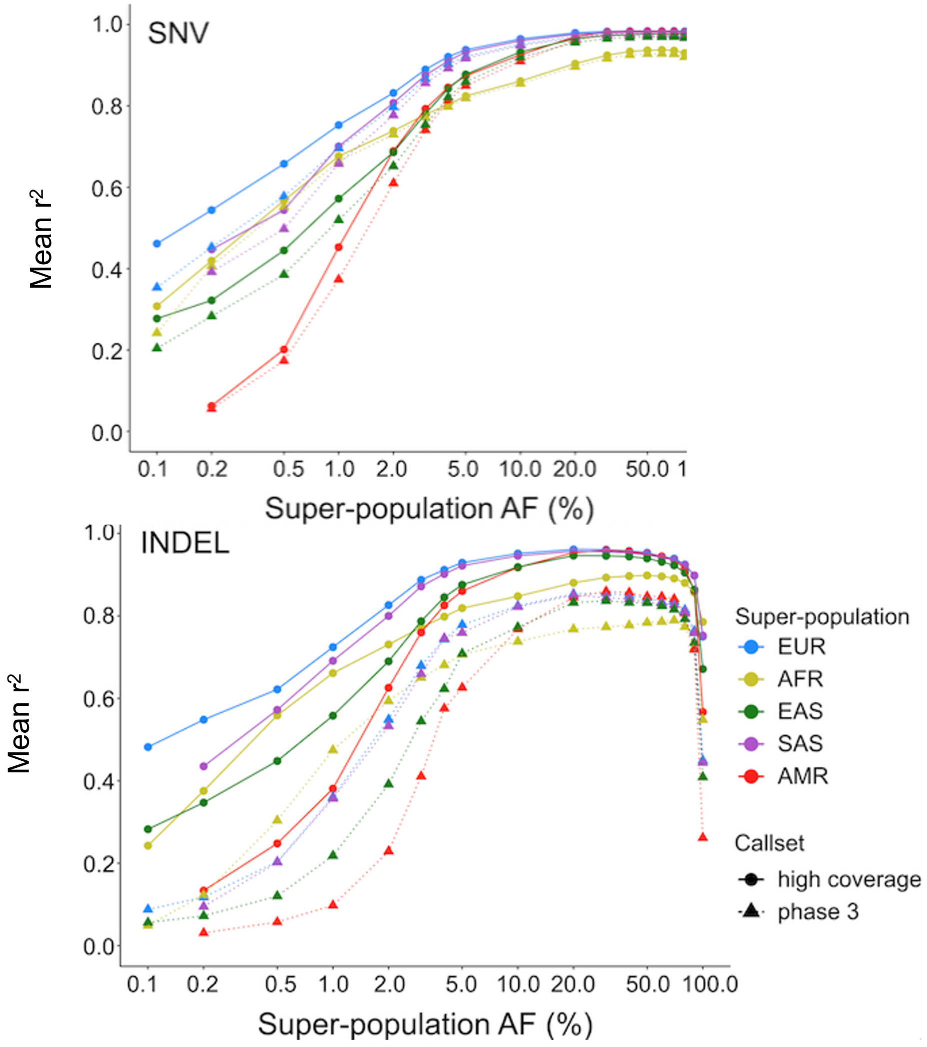
IMPUTATION PERFORMANCE

- Imputed a set of 279 diverse samples from the Simons Genome Diversity Project (SGDP) using IMPUTE2 software.
- Evaluated the accuracy of imputed genotypes by computing the squared correlation (r^2) between imputed allele dosages and dosages from WGS data across 110 samples, 22 from each of the five super-populations.

Performance of the high coverage panel stratified by variant type and genomic region:



Comparison against phase 3 (shared sites):

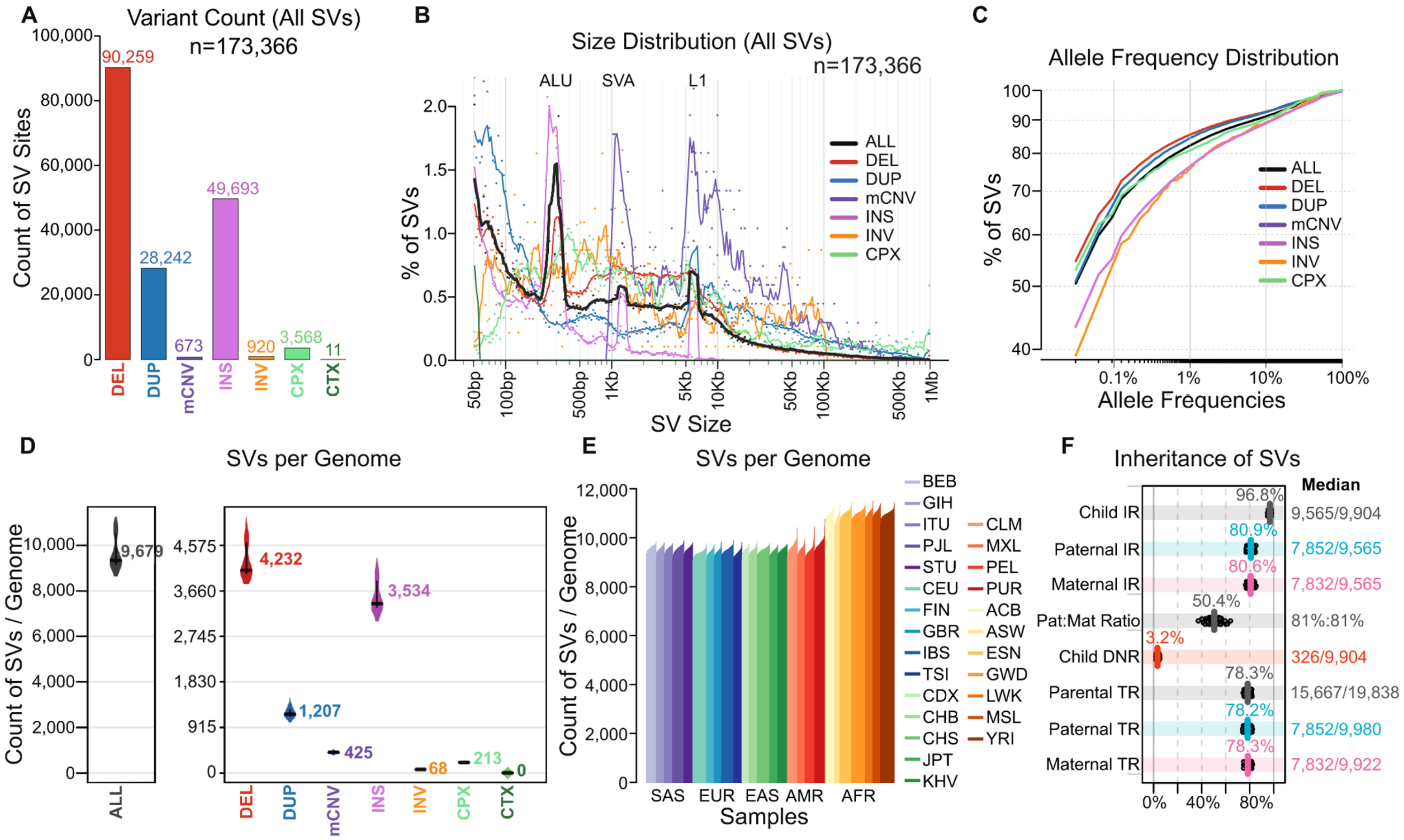


Howie, B.N. *et al.* PLoS Genet. *5*, e1000529 (2009).

INTEGRATED STRUCTURAL VARIANT CALLS

SV callset integrated from GATK-SV, SVTools and Absinthe:

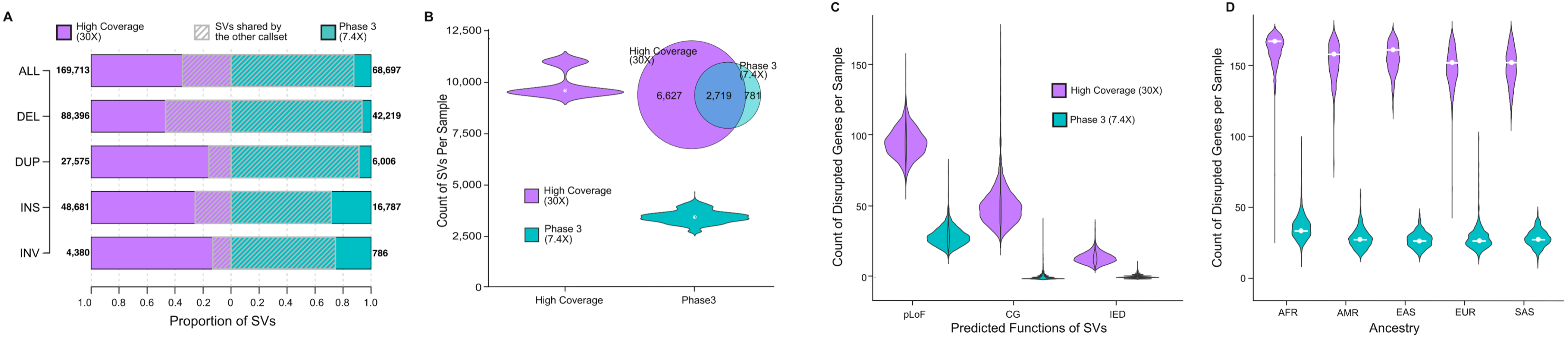
- ➢ A total of 173,366 SV sites across 3,202 samples in the high coverage callset.
- ➢ An average of 9,679 SVs per genome.
- ➢ More SVs are observed in African population.



INCREASED SV YIELD COMPARED TO PHASE 3

Increased sensitivity is observed in the SV callset from high-coverage (~35X) sequences than the 1KGP phase 3 callset (~7.4X):

- ➢ Over two times more SV sites are detected from the high-coverage sequences than 1kGP phase 3 (169,713 vs. 68,697).
- ➢ Increased sensitivity is also reflected in the SV count per sample.
- ➢ Most significant increase in sensitivity is reflected in small SVs < 250bp.
- ➢ More genes are altered by SVs in the new callset than 1kGP phase 3.
- ➢ More genes are altered in AFR population than others.



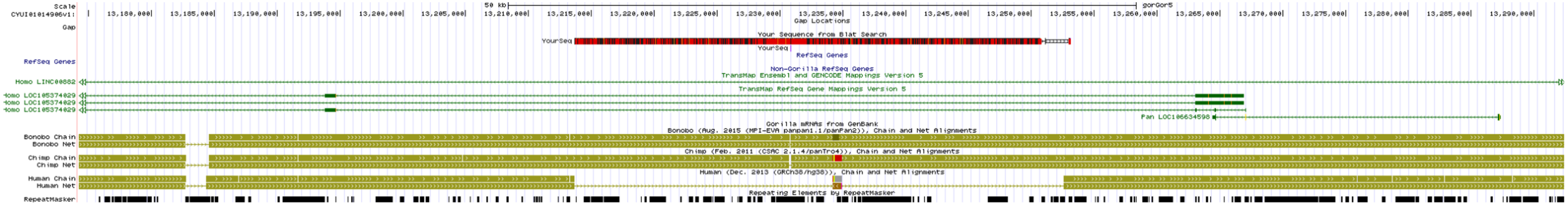
CONCLUSIONS

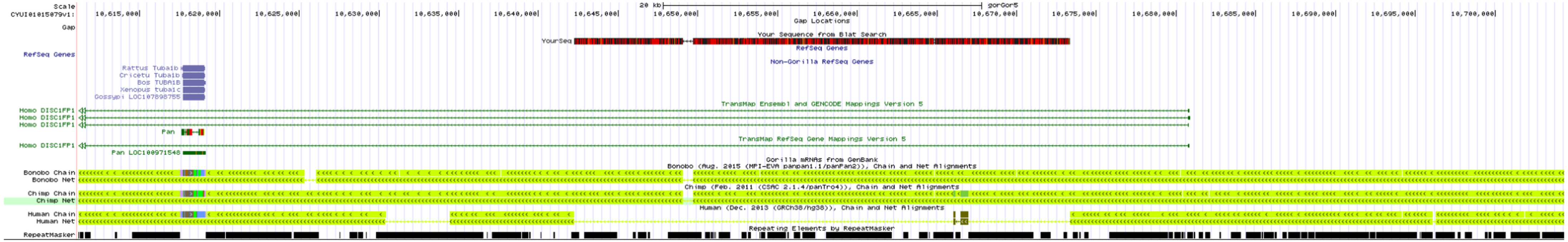
- We called **>111 million SNVs & >14 million INDELs** across the 3,202 samples with FDR of 0.3% and 1%, respectively.
- Relative to the phase 3 callset, we called **6% more SNVs and 48% more INDELs per genome**.
- The vast majority of the **new SNVs are in the rare MAF spectrum** (AC ≤ 2).
- We observed **gains in INDEL counts across the entire MAF spectrum**, with gains in the rare end of the spectrum being the most pronounced.
- The phased high coverage SNV/INDEL panel exhibits **an order of magnitude higher phasing accuracy** as compared to the phase 3 dataset across the entire MAF spectrum.
- Improvements in small variant calling, coupled with higher phasing accuracy of the high coverage panel, translated into **significantly better imputation accuracy**, especially for INDELs, across all of the 1kGP super-populations.
- We called **173,366 SV** sites across 3,202 samples with FDR ≤ 3.2%
- More genes are altered by SVs in the high coverage call set as compared to phase 3

ABSINTHE INSERTION CALLING

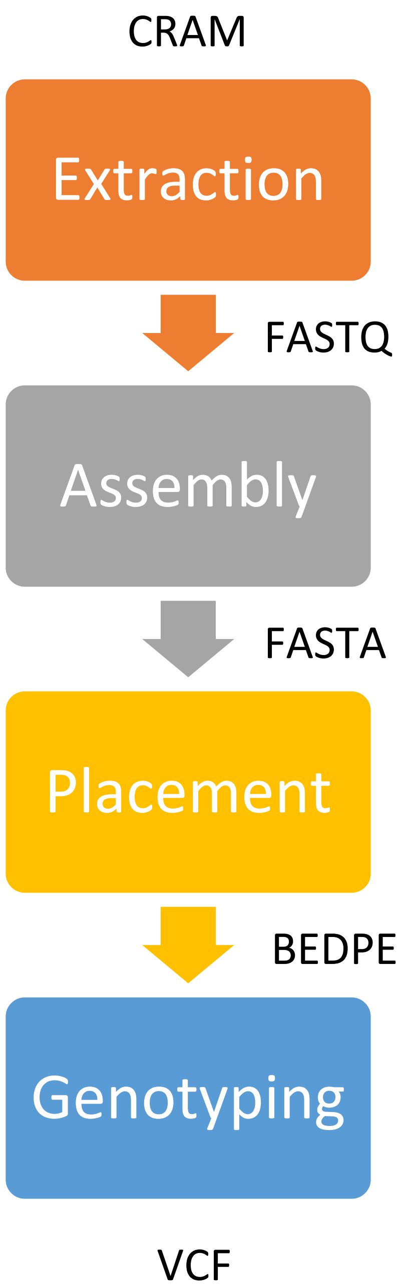
- Calling “insertions” from short reads has traditionally been difficult
- Absinthe identifies reads that don’t map or mismap and assembles them
- The resulting contigs can then be placed back on the reference

EXAMPLES OF ASSEMBLED INSERTIONS





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

RESULTS FROM TOPMED

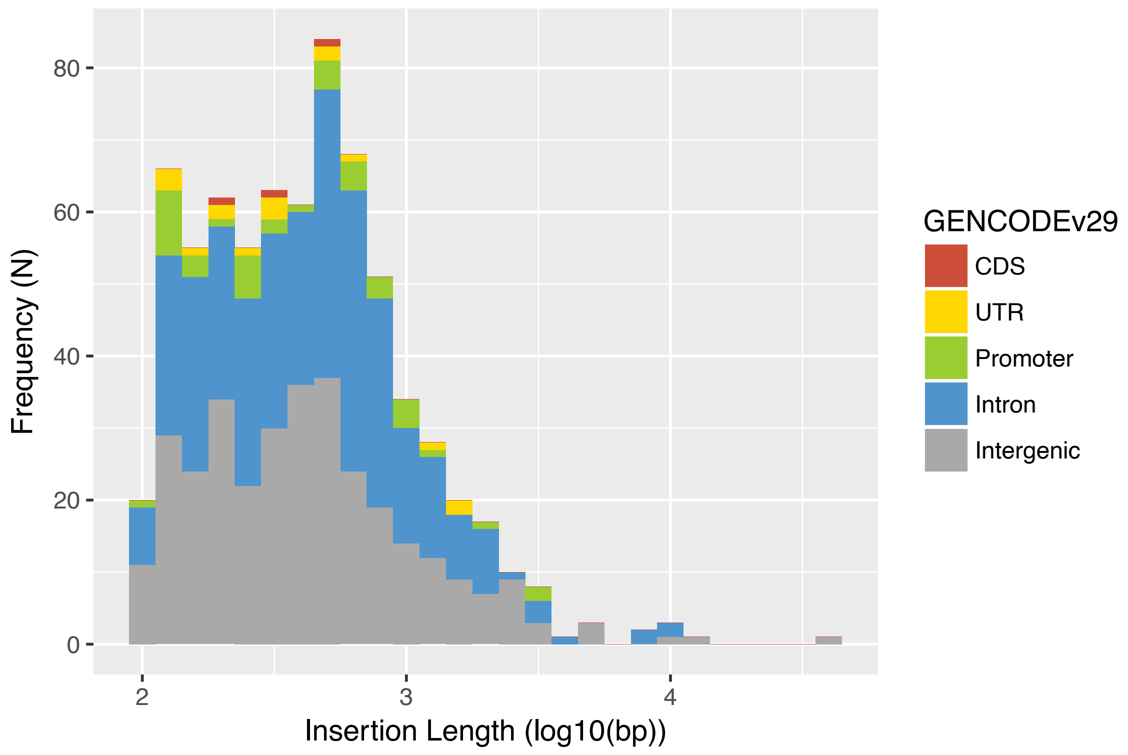
- 53,831 genomes (reads aligned to GRCh38)
- Genotype using Paragraph, rather than simply determining presence/absence (insertions only)

		53k GRCh38
Insertions	N	713
	(bp)	514,642
Breakends	(N)	304
	(bp)	186,343

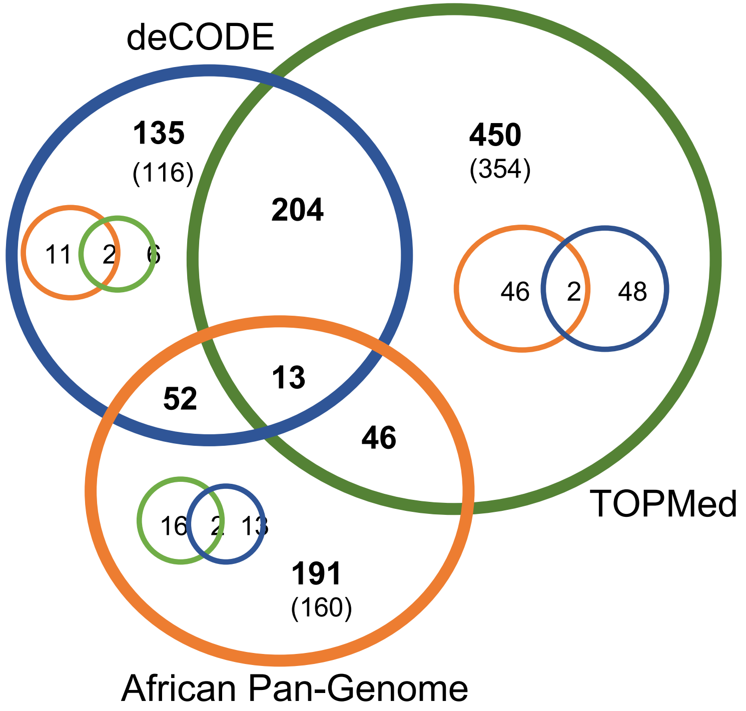
Chen *et al*. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biology 2019

RESULTS FROM TOPMED

Length distribution

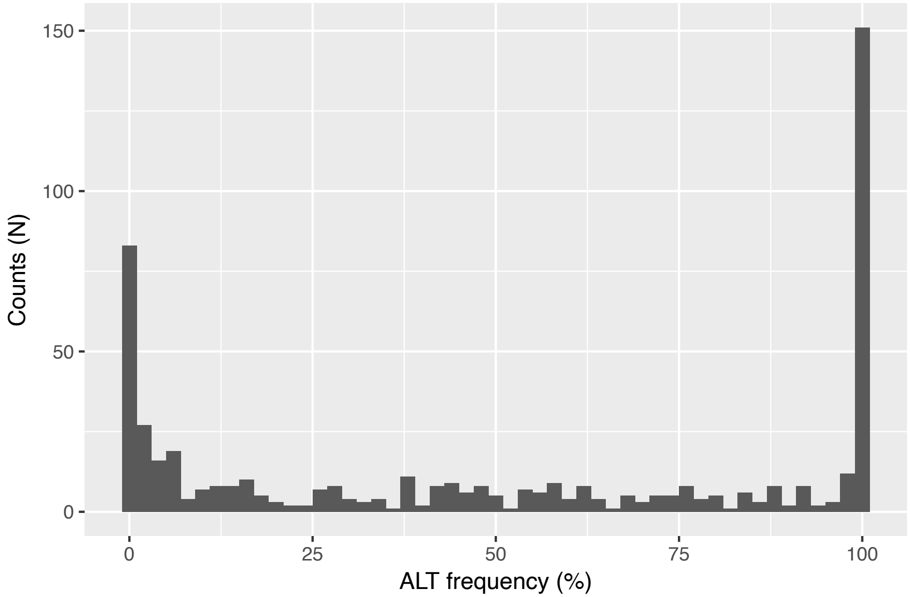


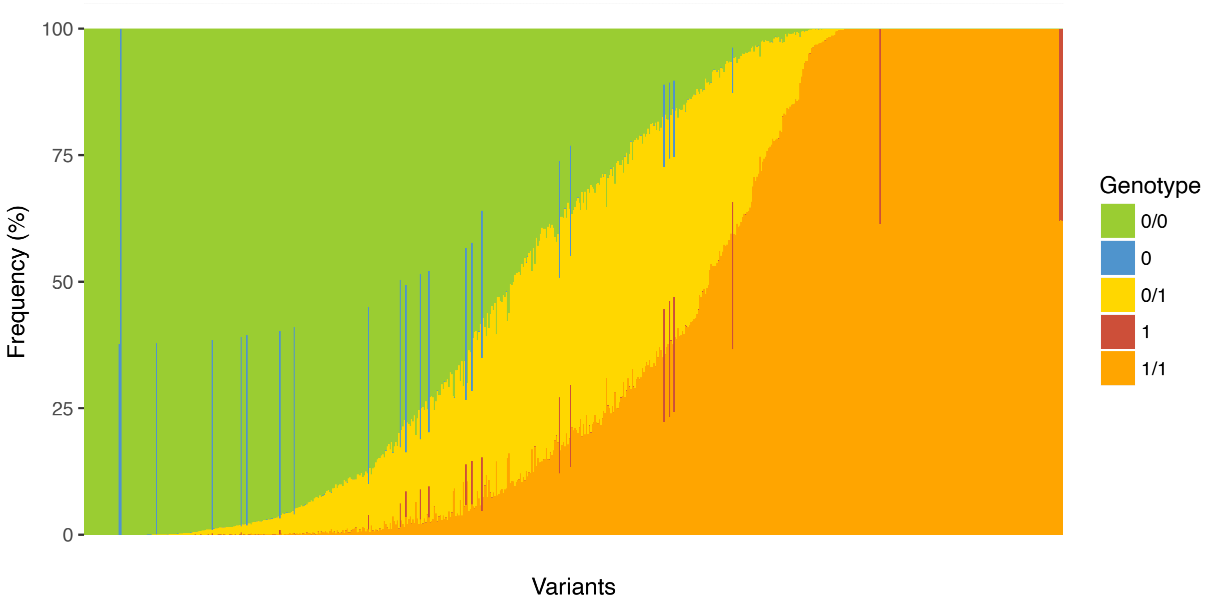
Positional concordance with insertions identified using short-read data from two previous studies



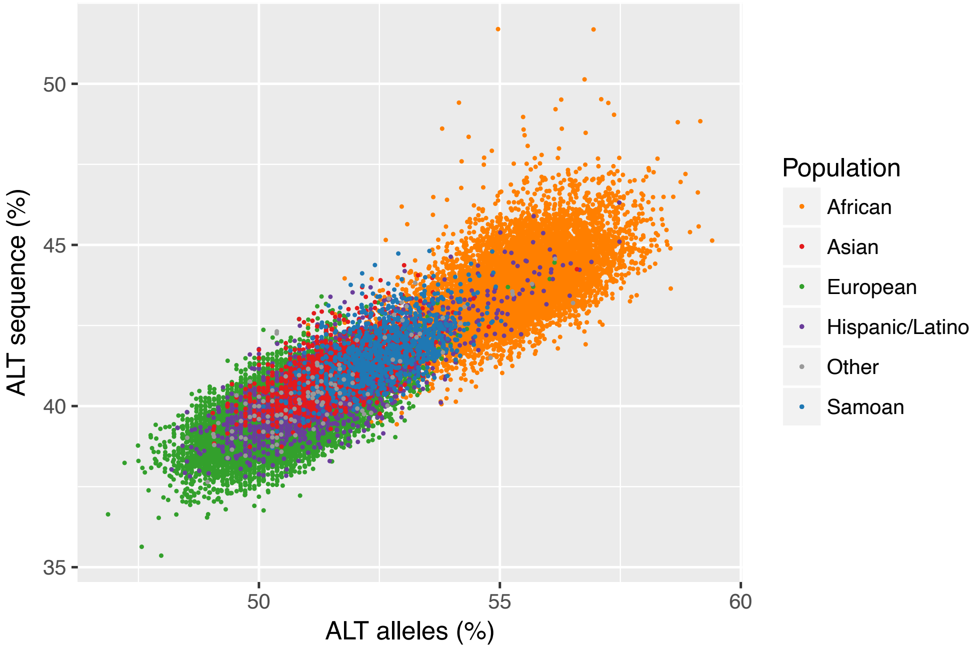
Taliun et al., *Nature*, 2021

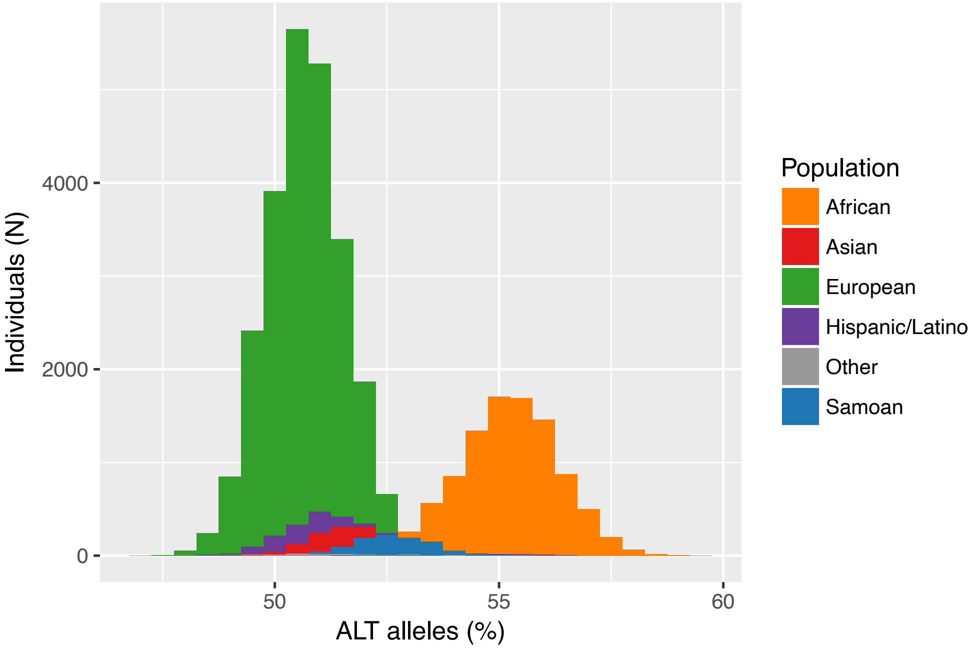
ALLELE AND GENOTYPE FREQUENCY



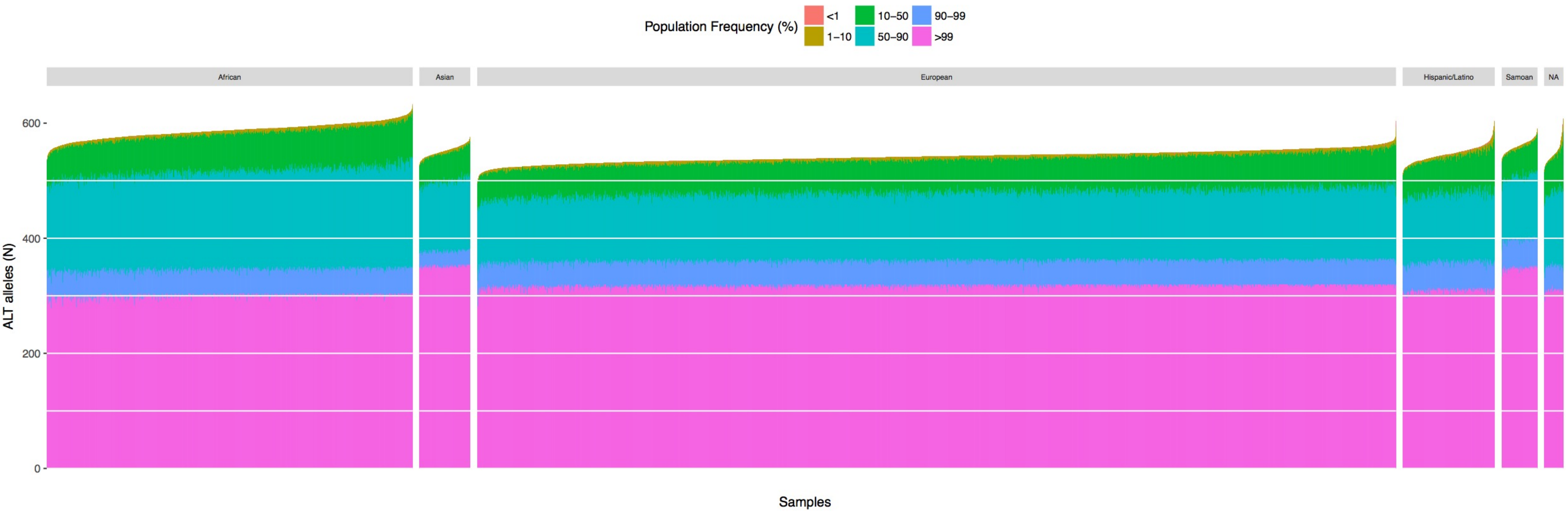


ALT ALLELE DISTRIBUTION BY ANCESTRY





ALT FREQUENCY WITH POPULATIONS



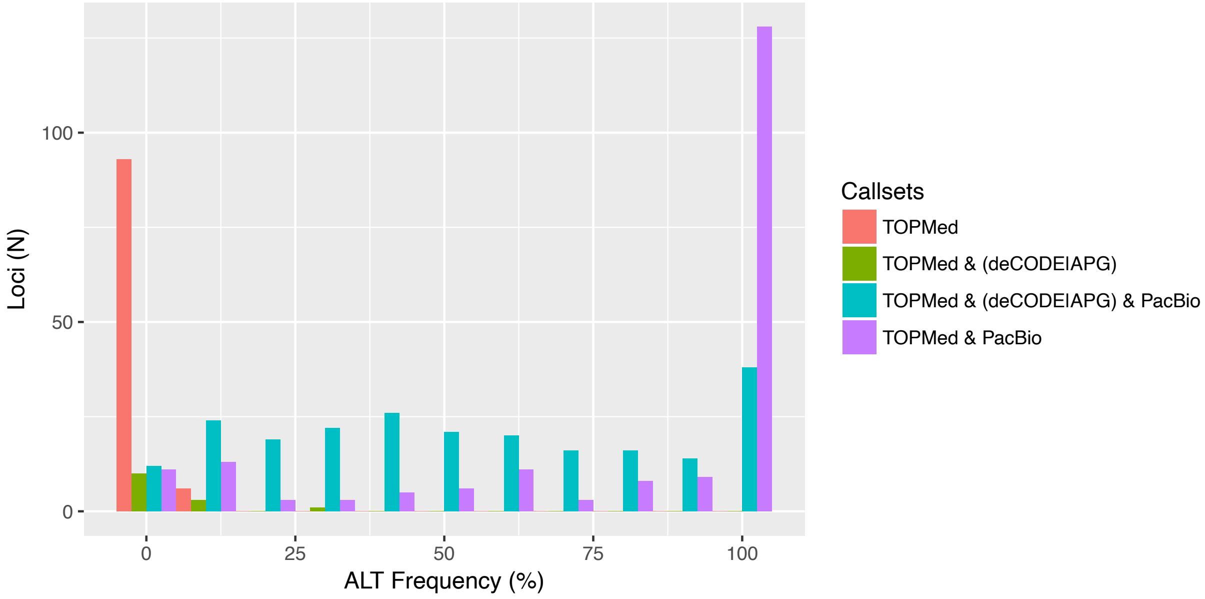
Higher fraction of >99% alleles in Asians and Samoans

Excess ALT alleles observed in individuals of African ancestry fall in the frequency range of 10-90%

VALIDATION WITH LONG READS

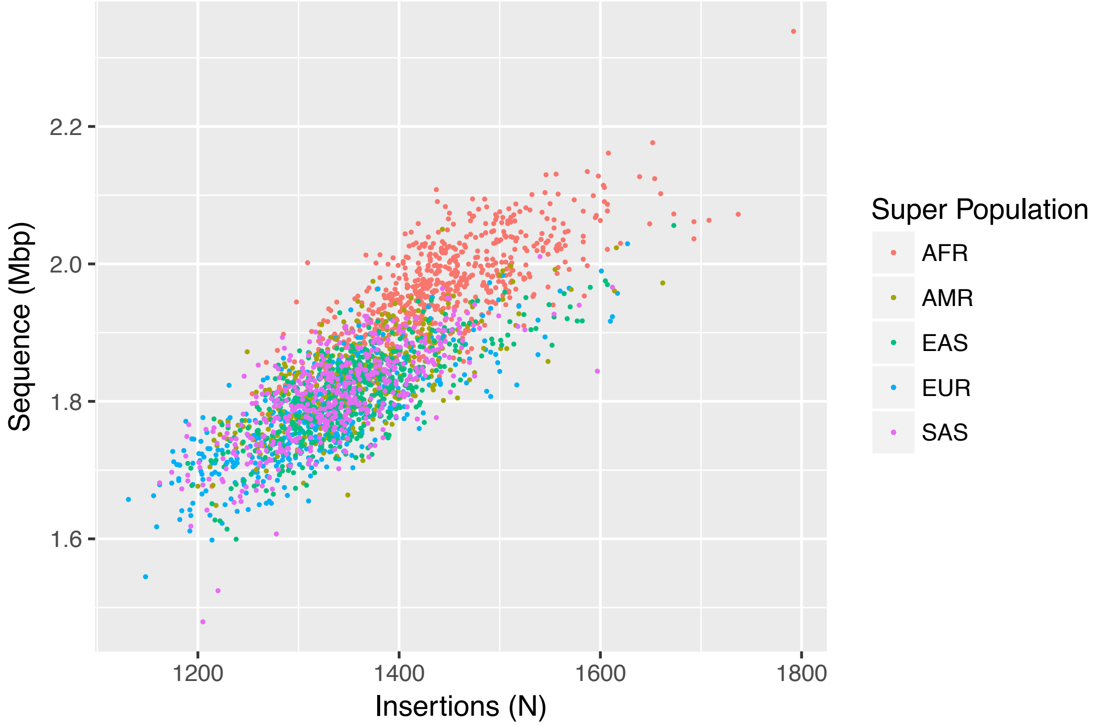
ALT allele frequency by overlap with deCODE, APG and PacBio*

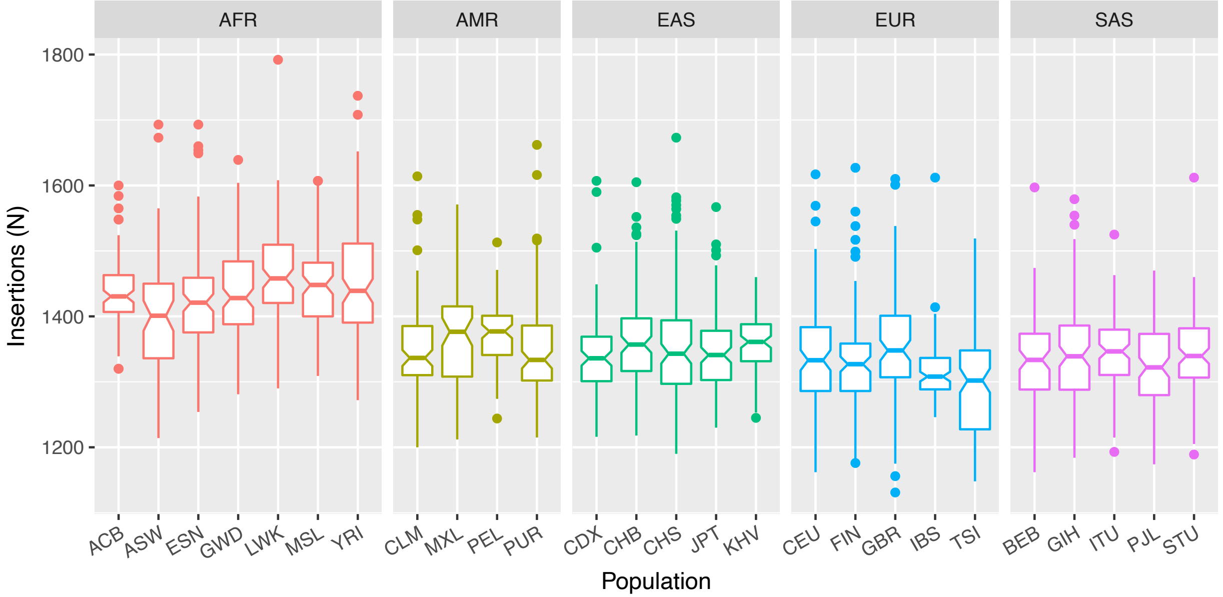
79% overlap PacBio insertions



*Audano *et al*. Characterizing the Major Structural Variant Alleles of the Human Genome. Cell. 2019 Jan 24;176(3):663-675.e19

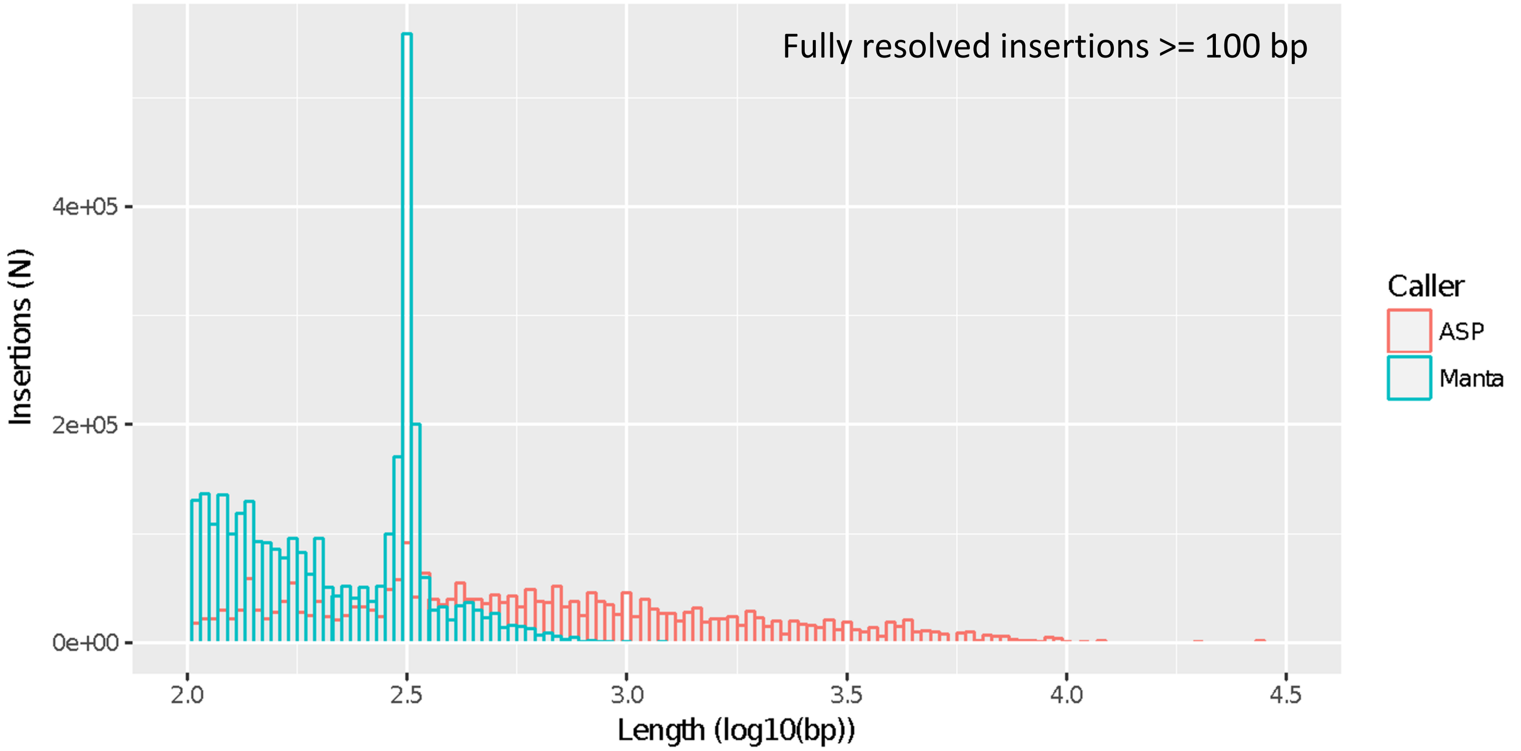
CALLING IN 1000 GENOMES



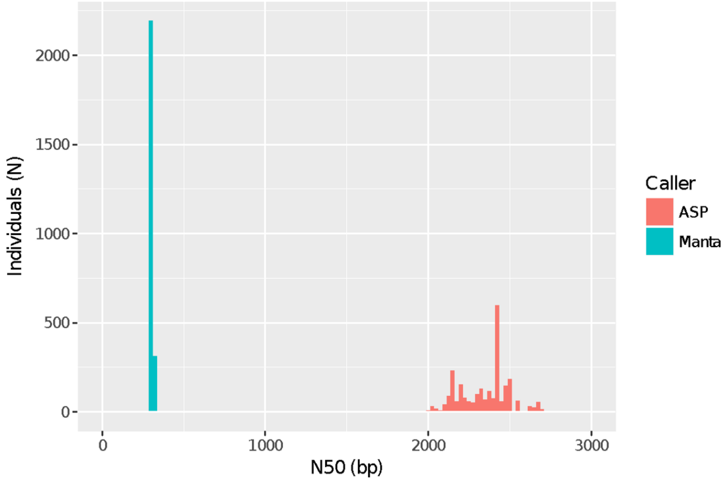


- 1,300-1,500 insertions per individual (1.6 – 2.2 Mbp)
- Larger number of insertions in individuals from African populations

INSERTION LENGTH DISTRIBUTION



- Consistent across individuals

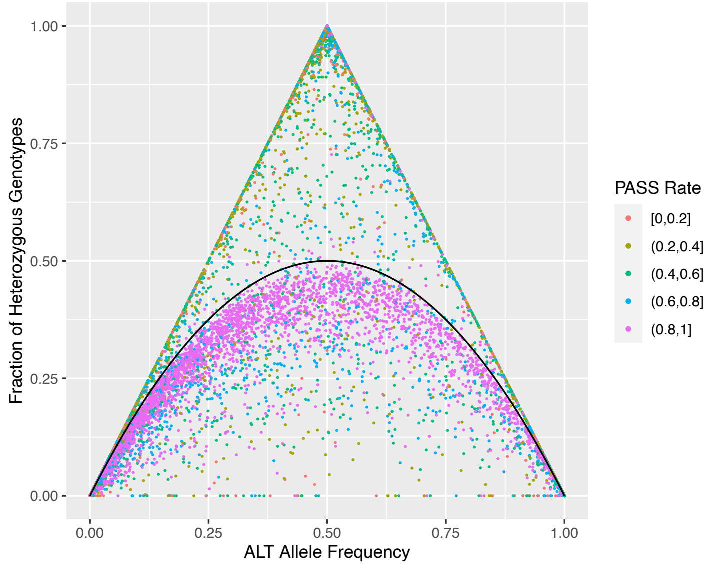


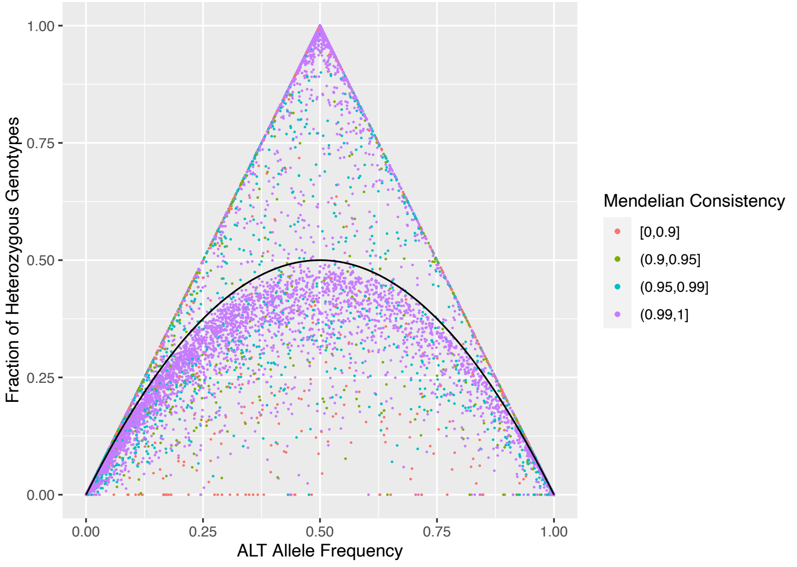
- Absinthe calls are a good complement to Manta’s as they extend well into the range of 1Kb – 10Kbp
- Several fully resolved insertions are longer than 10Kbp

1000 GENOMES MERGED CALLSET

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



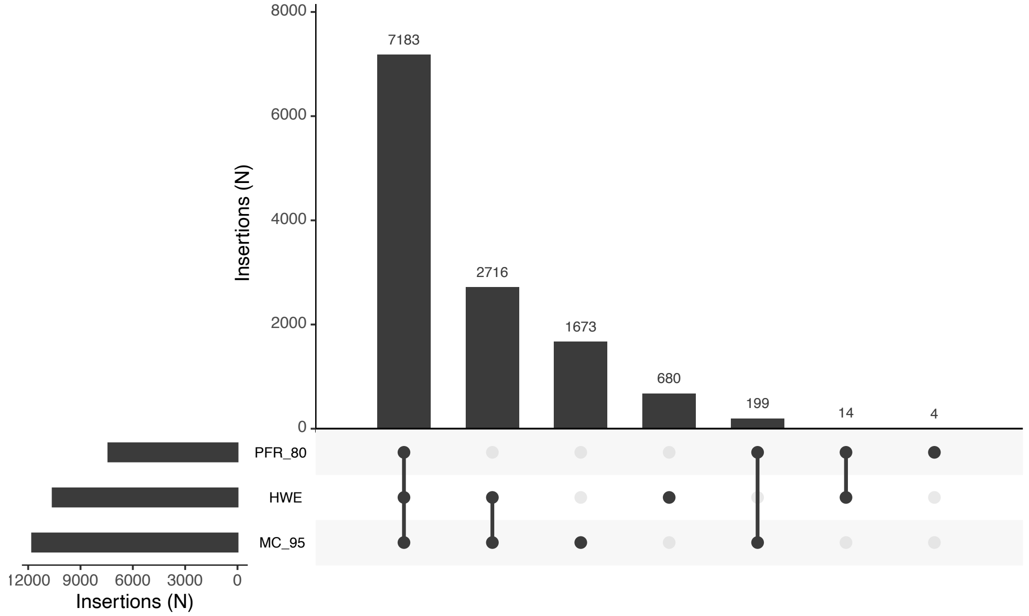


Genotyping:

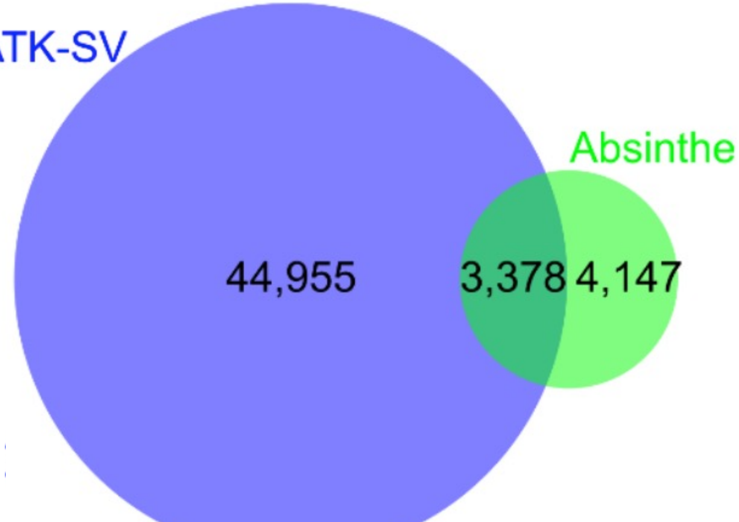
- Paragraph (Chen et al, 2019)

Filters:

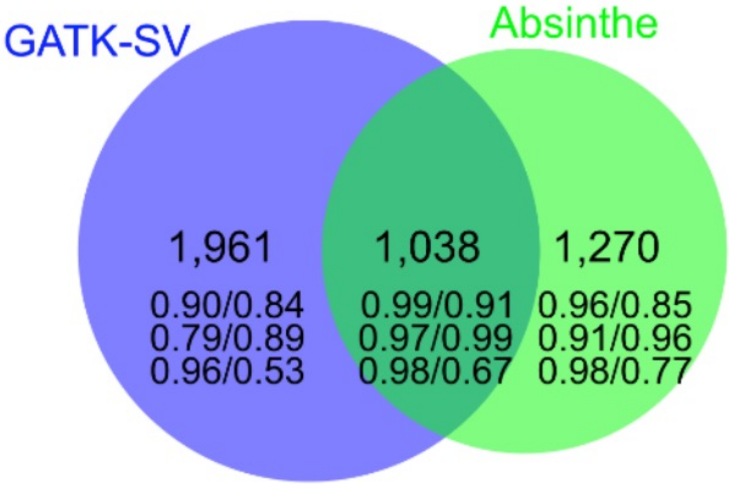
- Super population PASS-filter rate [all >= 0.8]
- Super population HWE [any > 10^{-6}]
- Mendelian Consistency based on 602 trios [>= 0.95]
- Output:
  - 7,183 HQ genotyped insertions



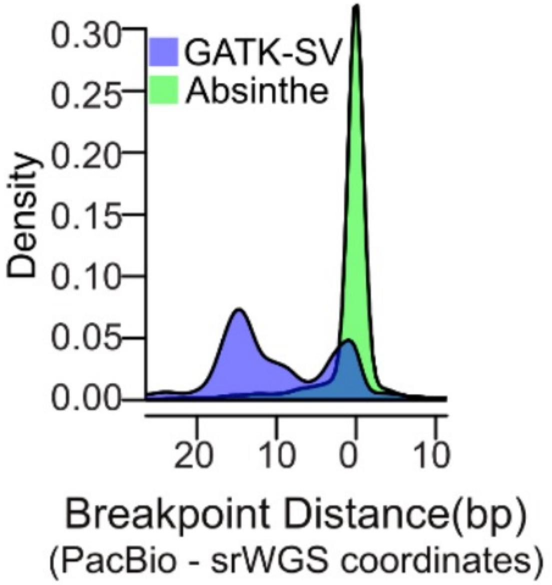
COMPARISON TO GATK-SV CALLS

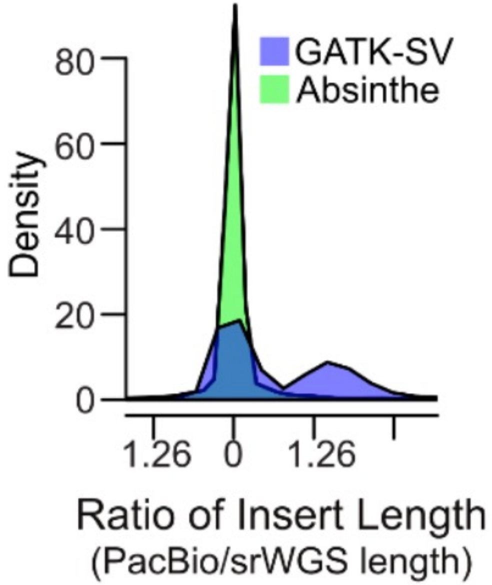


Per sample averages



N insertions
validated by VaPoR / assessable by VaPoR
in PacBio SVs (Ebert et al. 2021) / in PacBio SVs (Chaisson et al. 2019)
transmission rate / rate of bi-parentally inherited

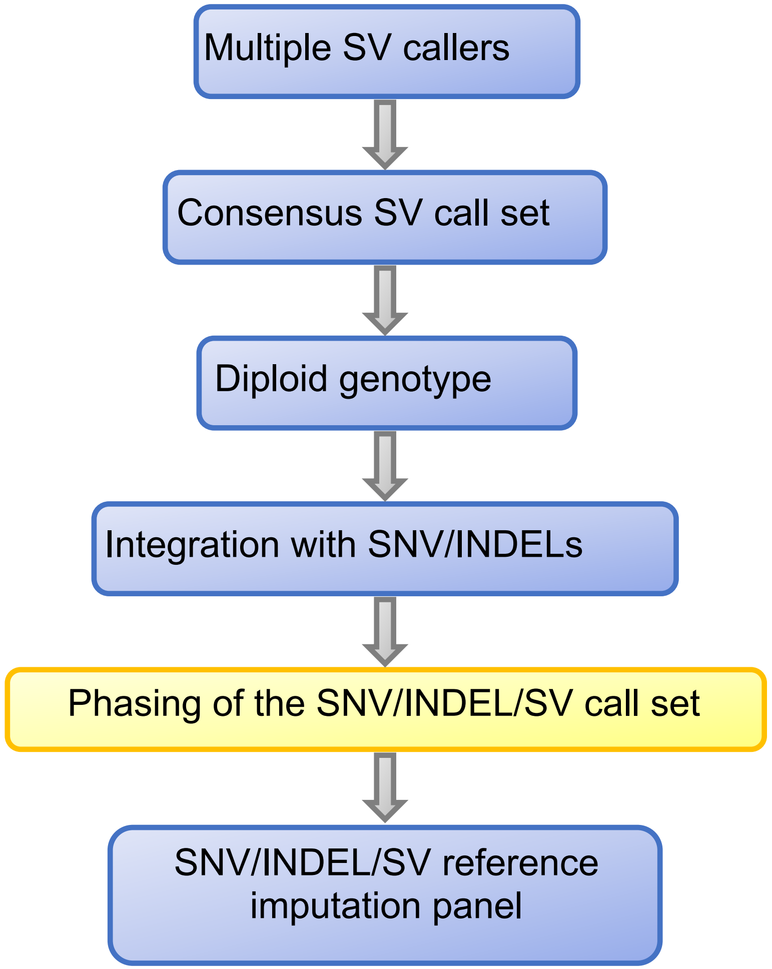




STRUCTURAL VARIANT IMPUTATION

- Imputation panels for SNVs and small indels have greatly improved our power to run associations for traits
- SVs are harder to call from sparse data than SNVs
- SVs have typically not been included on imputation panels
- Association of SVs to phenotype has typically been done case-by-case leveraging associations discovered from linked SNVs
- We would like to be able to directly associated SVs with phenotype

PHASING ACCURACY OF SVS



Multiple SV callers

Consensus SV call set

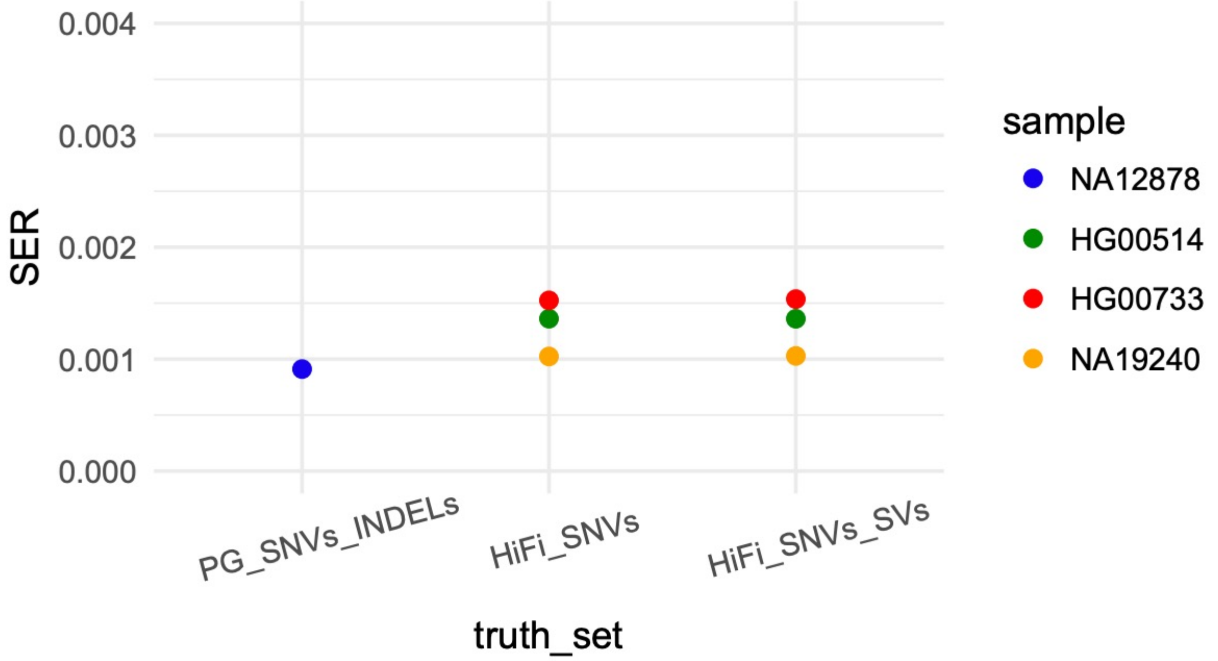
Diploid genotype

Integration with SNV/INDELs

Phasing of the SNV/INDEL/SV call set

SNV/INDEL/SV reference imputation panel

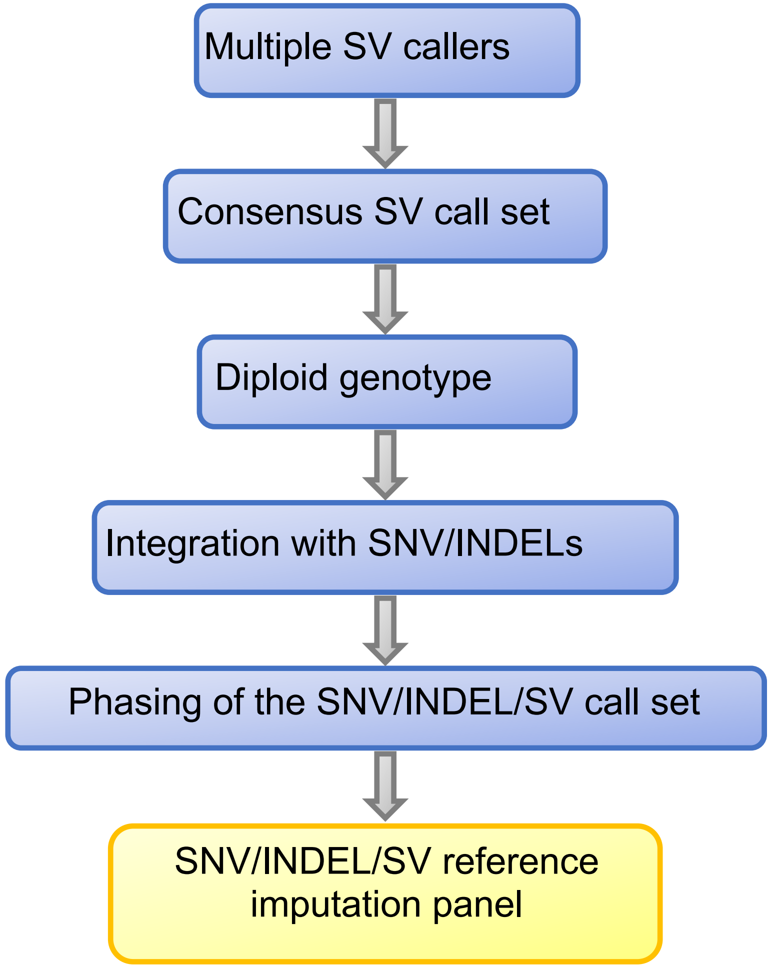
Using PacBio-HiFi haplotype-resolved SNV/SV call sets* to assess accuracy of SV phasing.



△ SER for DELs = 0.012

* HGSVC, pre-publication.

EVALUATION OF IMPUTATION PERFORMANCE



Multiple SV callers

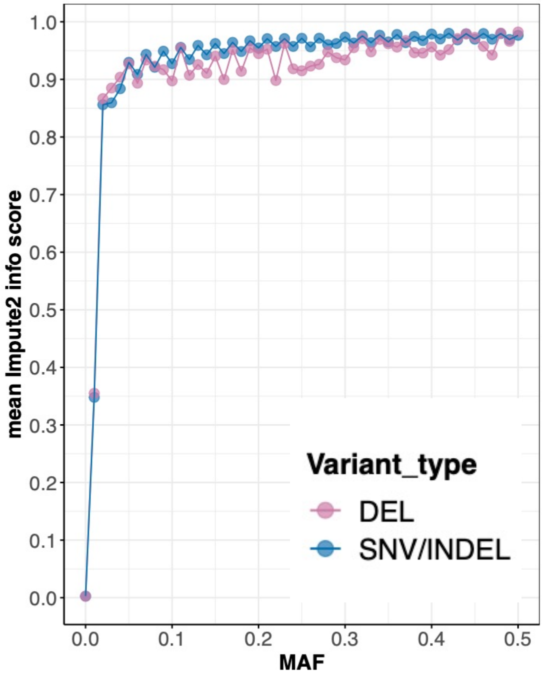
Consensus SV call set

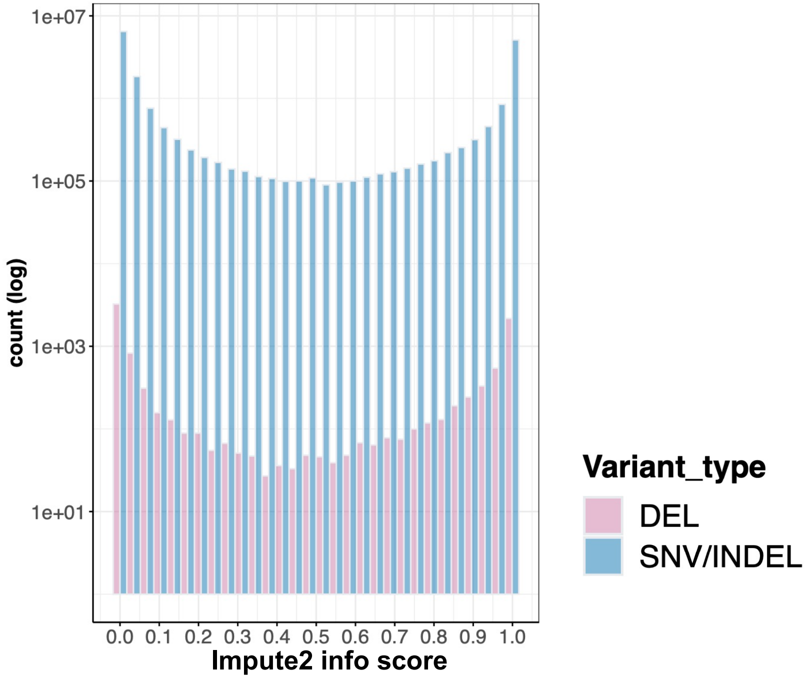
Diploid genotype

Integration with SNV/INDELs

Phasing of the SNV/INDEL/SV call set

SNV/INDEL/SV reference imputation panel





SV GT concordance evaluation against the GIAB truth set*:

Imputed sample	Info score threshold	Sensitivity	Precision
HG002	>=0.5	98.12%	95.55%

* Zook JM *et al*. *Sci data*, 3:160025 (2016)

STRUCTURAL VARIATION IN ALZHEIMER'S

- Create a harmonized, publicly available SV call set from a 972 familial and 39,000 unrelated LOAD case-control ADSP dataset of multi-ethnic ancestry.

- Augment ADSP SV call-set in by using SVs derived from long-read sequencing data from 200 AD patients.

- Increase sample size by imputing SVs in individuals without WGS data from the AD Genetics Consortium (ADGC).

- Identify common and rare SVs associated with LOAD and related endophenotypes.

ACKNOWLEDGEMENTS

NYGC:
Marta Byrska-Bishop
André Corvelo
Uday Evani
Anna Basile
Wayne Clarke
Rajeeva Musunuri
Giuseppe Narzisi
Kshithija Nagulapalli
Alexi Runnels
Lara Winterkorn
Soren Germer

HGSVC:
Michael Talkowski
Xuefang Zhao
Harrison Brand
Ira Hall
Haley Abel
Allison Regier
Evan Eichler
Peter Audano
Susan Fairley
Ernesto Lowy-Gallego
Paul Flicek

AD SV Grant:
Badri Vardarajan

TOPMed Consortium:
Daniel Taliun
Gonçalo Abecasis

Funding: NHGRI, NIA