

Population-Stratified Structural Variation Frequencies Using Genomic Data

Pranav M N
Roll No: ME23B186

ME4901-Undergraduate Research Project

Guided By: Dr.Raushan Singh and Manjri Bharti

Background: Genome vs. Genomic Data

What is a Genome?

- The **genome** is the complete set of genetic material (DNA) present in an organism.

What is Genomic Data?

- **Genomic data** refers to the digitized, high-throughput sequence information extracted from an organism's genome.
- A **VCF** file details genomic data by listing chromosomal positions where an individual's DNA varies from the hg37 reference genome.

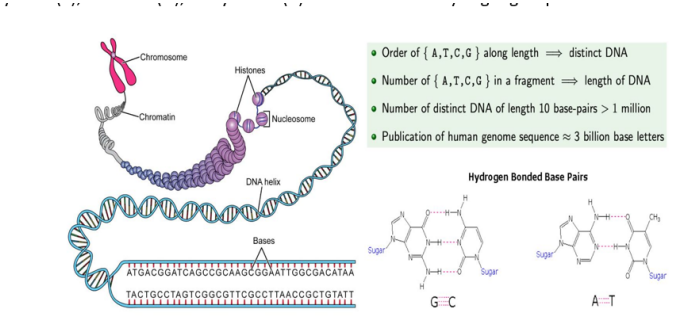


Figure 1: Biological representation of DNA

Figure: Biological representation of DNA and base pairing.

What are Structural Variants ?

Definition: Structural variants (SVs) are genomic alterations involving segments of DNA $\geq$50 bp in length.

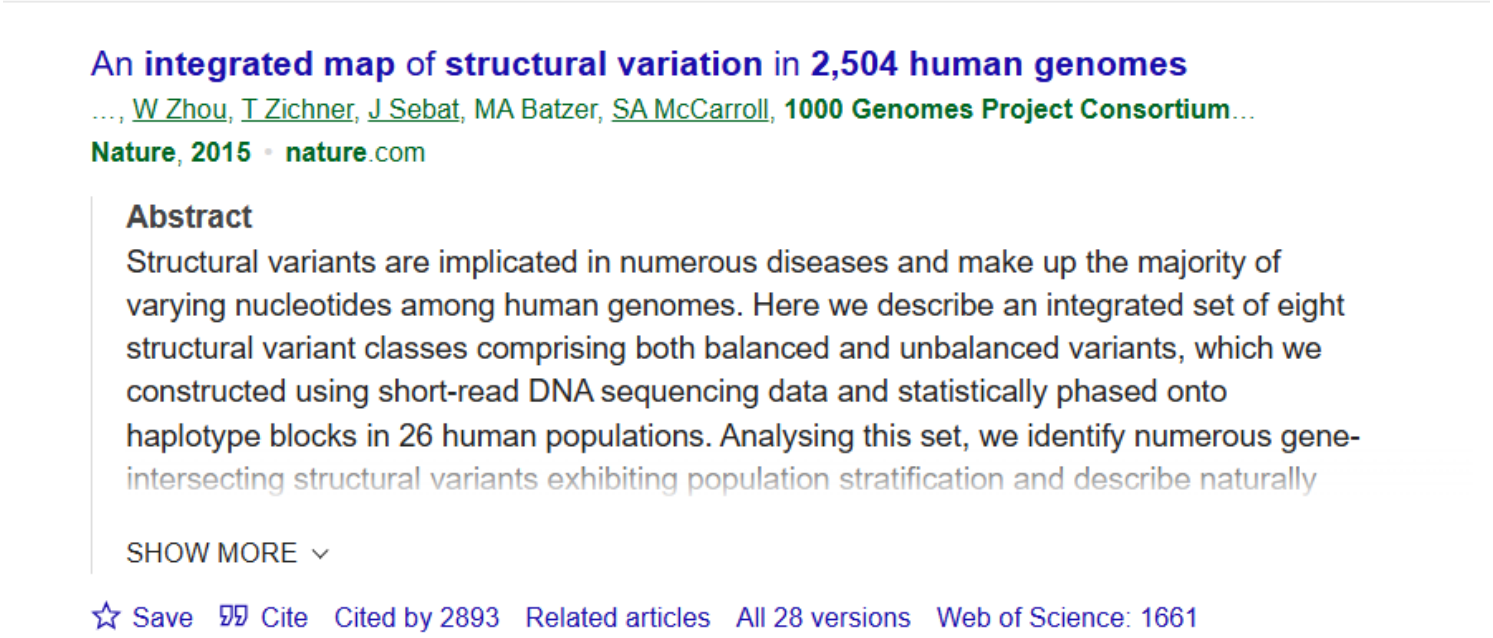
Classification of SVs:

- **Deletion:** Loss of a genomic segment.
- **Duplication:** Gain of an extra copy of a genomic segment.
- **mCNV:** Multiallelic Copy Number Variants with highly variable copy counts across the population.
- **Inversion:** A DNA segment reversed in its orientation relative to the reference.
- **MEI:** Mobile Element Insertion (e.g., Alu, LINE-1,SVA) into new genomic locations.
- **NUMT:** Nuclear Mitochondrial DNA; the translocation of mitochondrial sequences into the nuclear genome.

Motivation

- **Primary Source of Genomic Variation:** Although structurally complex, structural variants (SVs $\geq$ 50 bp) account for the absolute majority of varying base pairs among individual human genomes.
- **Critical Clinical Relevance:** SV architectures are heavily linked to complex human phenotypes and disease predispositions, ranging from cognitive disabilities to obesity and cancer.
- **Limitations of Legacy Technologies:** Early SV surveys relied on low-resolution microarrays or clone-based approaches that were severely restricted by small sample sizes and lacked global population scale.
- **Persistent "Deletion Bias":** Prior phases of the 1000 Genomes Project encompassed fewer individuals and were largely or entirely restricted to tracking deletions, leaving other major classes unresolved.

Aim and Objectives



An **integrated map** of **structural variation** in **2,504 human genomes**

..., W Zhou, T Zichner, J Sebat, MA Batzer, SA McCarroll, **1000 Genomes Project Consortium**...

Nature, **2015** • **nature**.com

Abstract

Structural variants are implicated in numerous diseases and make up the majority of varying nucleotides among human genomes. Here we describe an integrated set of eight structural variant classes comprising both balanced and unbalanced variants, which we constructed using short-read DNA sequencing data and statistically phased onto haplotype blocks in 26 human populations. Analysing this set, we identify numerous gene-intersecting structural variants exhibiting population stratification and describe naturally

SHOW MORE

☆ Save Cite Cited by 2893 Related articles All 28 versions Web of Science: 1661

Objectives

- **Replicate Summary Statistics:** Plot variant class distributions.
- **Profile Population Variation:** Cluster continental population burdens.
- **Map Genomic Density:** Identify chromosomal mutation hotspots.
- **Analyze Functional Impacts:** Classify regional genomic overlaps.

Methodology: Data Collection

Dataset Origin & Scope

- **Source [2]**: Cohort data from the Phase-3 1000 Genomes Project.
- **Data Refinement Note:** While individual per-chromosome VCF files (chromosomes 1–22, X, Y) were evaluated, certain files presented corruptions. To guarantee analytical consistency, the fully integrated whole-genome structural variation file (`SV_combined`) was utilized.

Dataset Characteristics [2]

Property	**Value**
Individuals	2,504
Populations	26
Superpopulations	5 (AFR, EUR, EAS, SAS, AMR)
Structural Variants	$\sim$60,000
SV Classes	8 major types

Methodology: Cohort Sample Distribution

Superpop.	Count	% Genomes	Granular Sub-populations (Sample Size)
AFR	661	26.4%	GWD (113), YRI (108), ESN (99), LWK (99), ACB (96), MSL (85), ASW (61)
EAS	504	20.1%	CHS (105), JPT (104), CHB (103), KHV (99), CDX (93)
EUR	503	20.1%	IBS (107), TSI (107), CEU (99), FIN (99), GBR (91)
SAS	489	19.5%	GIH (103), ITU (102), STU (102), PJL (96), BEB (86)
AMR	347	13.9%	PUR (104), CLM (94), PEL (85), MXL (64)
Total	**2,504**	**100.0%**	*Balanced representation ensures high analytical power across ancestral lines.*

Results: Structural Variant Summary Statistics

- **Method:** Extracted the structural variant class from the `SVTYPE` field in the VCF `INFO` column using a custom **Python** script with the `pysam` library.
- **Total Counts Resolved:** 68,818 total SVs.
- **Conclusion:** In alignment with the paper's findings, deletions dominate the structural variation landscape.

SV Type	Replicated Count	Original Paper Count [1]
Deletion	42,279	42,279
Mobile Element Insertion (MEI)	16,631	16,631
Duplication (DUP)	6,025	6,025
Copy Number Variant (CNV)	2,929	2,929
Inversion (INV)	786	786
Mitochondrial Insertion (NUMT)	168	168

Results: Per-Individual Structural Variant Burden Analysis

Mathematical Formulation:

$$\text{Average Burden Per Genome} = \frac{\sum_{i=1}^{N} \text{Non-Ref SV Counts}_i}{N}$$

Table: Replicated Variant Burden vs. Original Paper Median Length Metrics

SV Type	Total Non-Ref Counts	Average Burden Per Genome	Original Paper Median kbp* [1]
DEL (Deletion)	5,412,996	2,161.74 variants	5,615 *kbp*
INS / MEI	2,708,808	1,081.79 variants	518 *kbp*
mCNV	433,560	173.15 variants	11,346 *kbp*
INV (Inversion)	76,016	30.36 variants	78 *kbp*
DUP (Duplication)	34,932	13.95 variants	691 *kbp*
NUMT	3,802	1.52 variants	3 *kbp*
Total SVs	**8,666,312**	**3,461 variants**	–

- *Observation:* Calculated across 2,504 individual genomes, deletions and insertions visually dominate individual structural landscapes.
- **Note:* The original paper column highlights the median base pair length footprint per individual genome for context.

Results: Structural Variant Length Distribution

- **Comparative Scale Analysis:** Structural variant (SV) counts are plotted across a logarithmic size scale from 50 bp to >1 Mb to validate replication precision.

Replicated Size Distribution

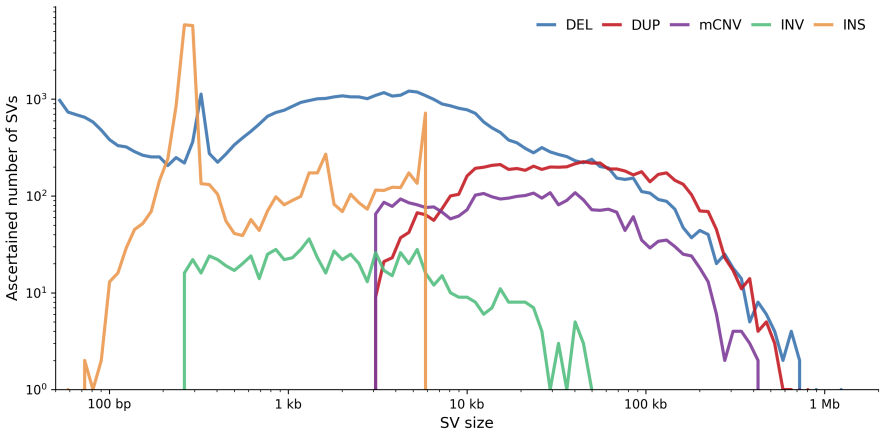


Figure: Computational profile generated via VCF parsing.

Original Paper Distribution [1] Sudmant et al. (2015).

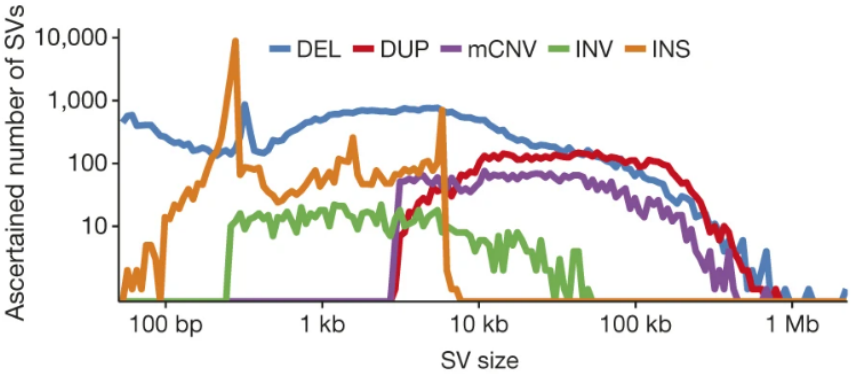


Figure: Reference framework curve from Phase 3 dataset.

Results: Population-Level SV Burden

Task: Population SV Burden

- **Key Insight:** Populations with African Ancestry (**AFR**) consistently show the highest per-genome SV burden (∼3,600–3,800 SVs), primarily driven by elevated deletion and MEI rates.

Table: Average Per-Individual SV Burden Profiles

Pop	Super	N	DEL	MEI	Total Burden
GWD	AFR	113	2380.2	1272.6	**3888.3**
ESN	AFR	99	2396.2	1212.3	**3845.1**
MSL	AFR	85	2385.6	1193.3	**3820.0**
ACB	AFR	96	2365.0	1212.4	**3818.6**
YRI	AFR	108	2377.6	1170.0	**3781.4**
LWK	AFR	99	2377.3	1097.4	**3710.7**
ASW	AFR	61	2331.3	1027.2	**3591.5**
TSI	EUR	107	2080.8	1164.3	3457.2
CDX	EAS	93	2075.0	1158.5	3449.0
CEU	EUR	99	2064.0	1114.7	3393.6
PUR	AMR	104	2142.8	1012.7	3376.6
GIH	SAS	103	2101.0	1043.3	3361.0
FIN	EUR	99	2058.4	983.4	3253.5
MXL	AMR	64	2087.7	893.9	3183.4

Results: Population Allele Frequencies & Variant Sharing

Replicated Distribution

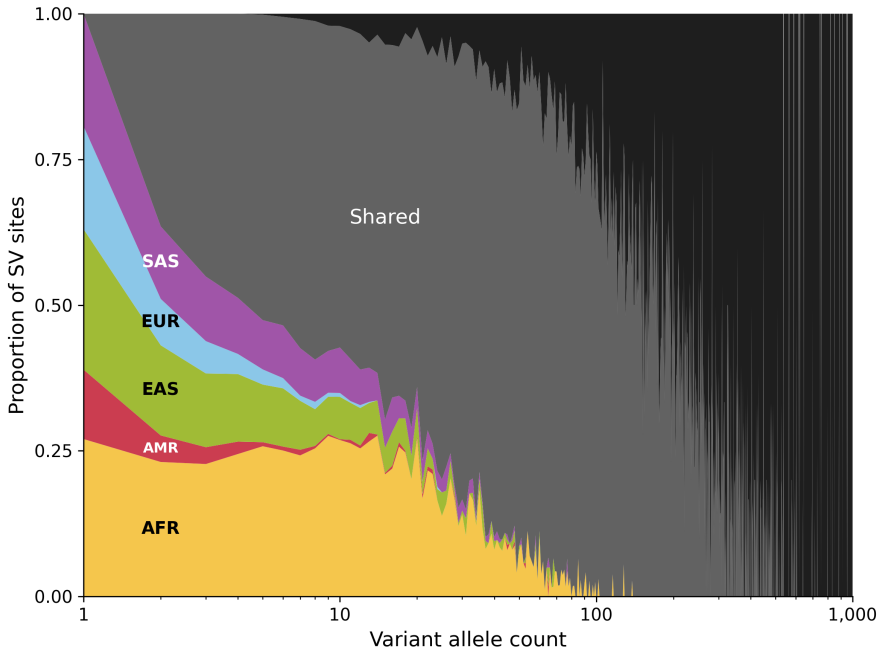


Figure: Proportion of SV sites binned by allele count (Reproduced).

Original Paper Distribution [1]

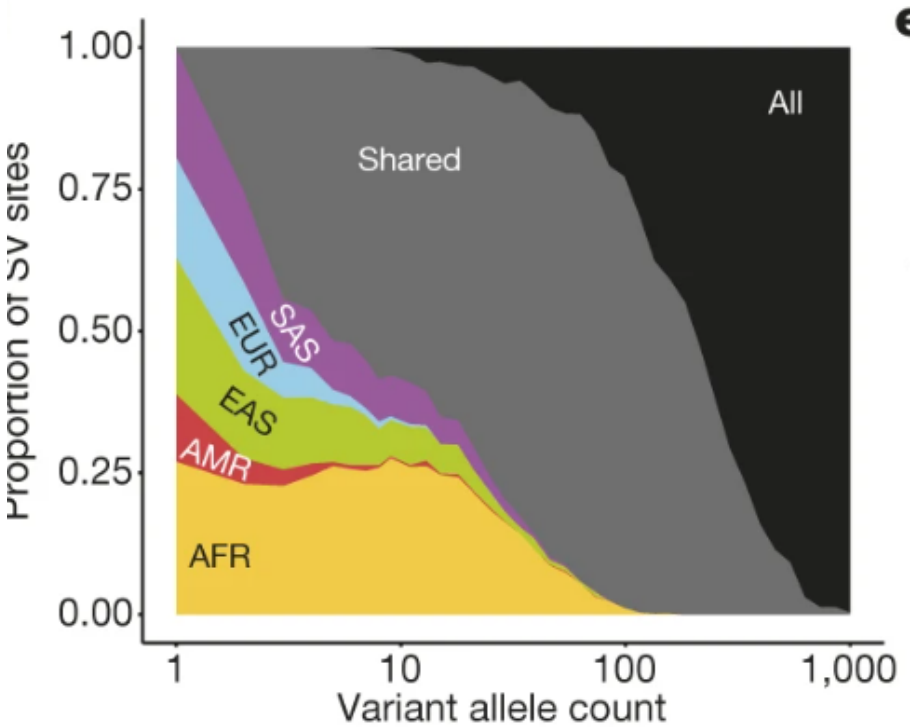


Figure: Reference framework from Sudmant et al. (2015).

Conclusion: Rare structural variants remain strictly population-specific, whereas common variants are universally shared globally.

Results: Population Structure via SV Genotypes

PCA Mathematical Foundation:

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (\mathbf{x}_i \cdot \mathbf{w})^2 \right\} = \arg\max \left\{ \frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}$$

Replicated Clustering

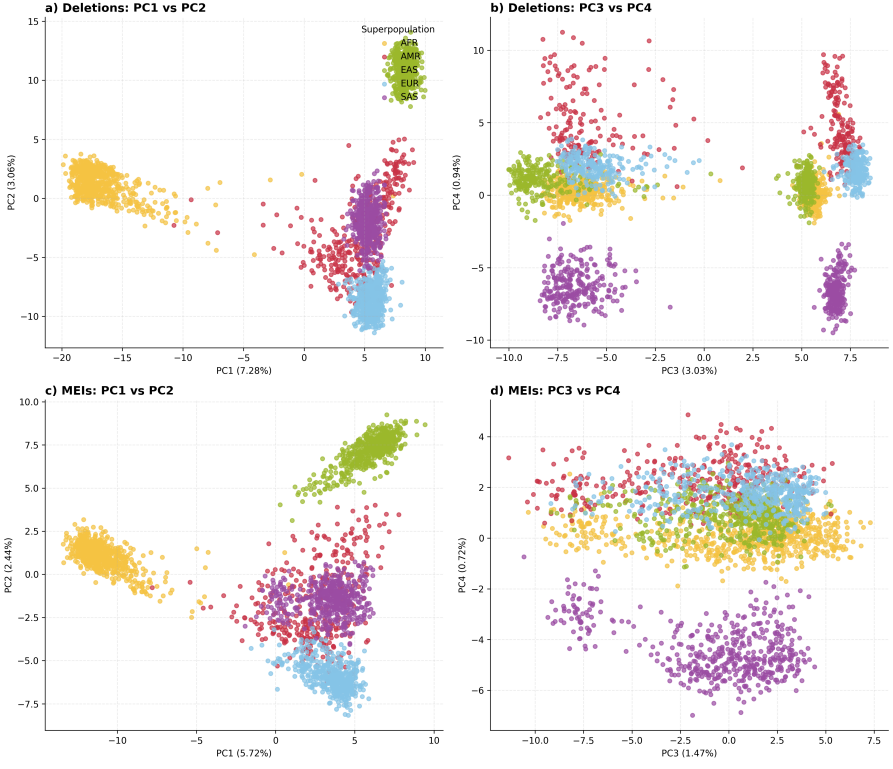


Figure: Replicated SV genotype PCA framework.

Original Paper Clustering [1] Sudmant et al. (2015).

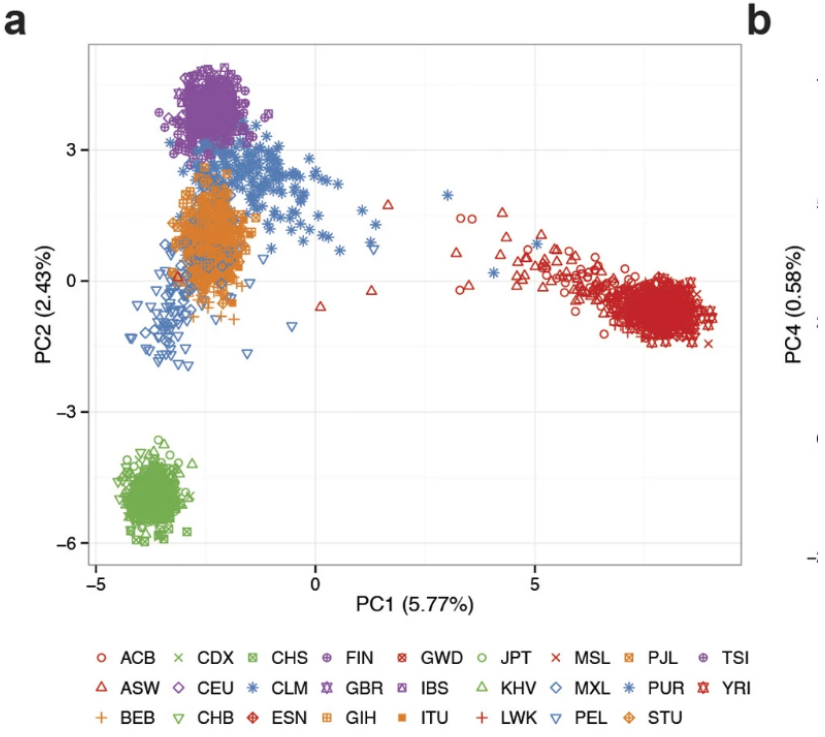
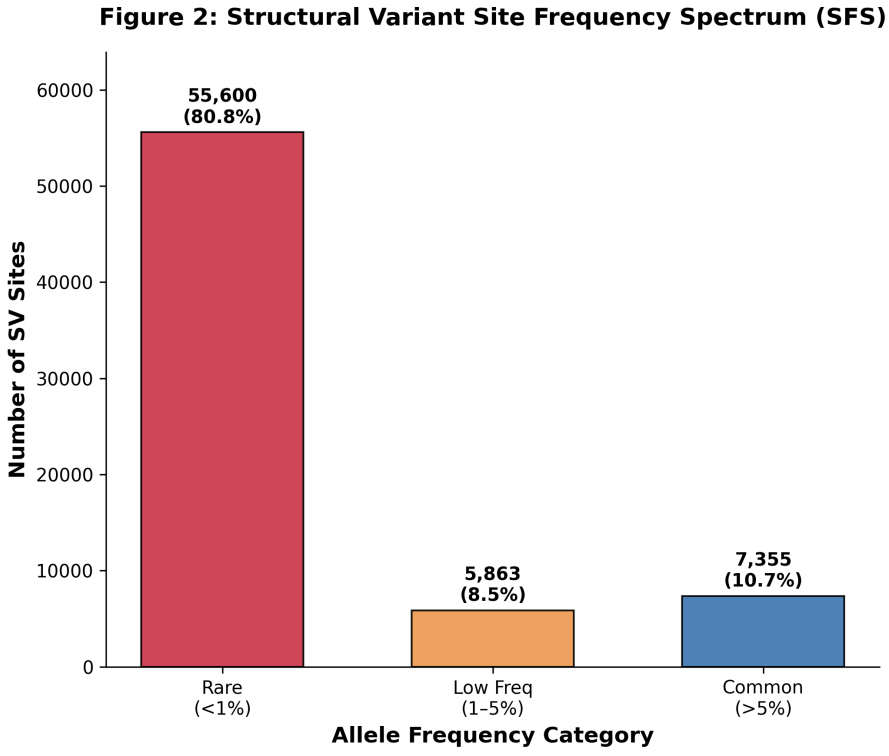


Figure: Original reference continental clusters (PC1 vs PC2).

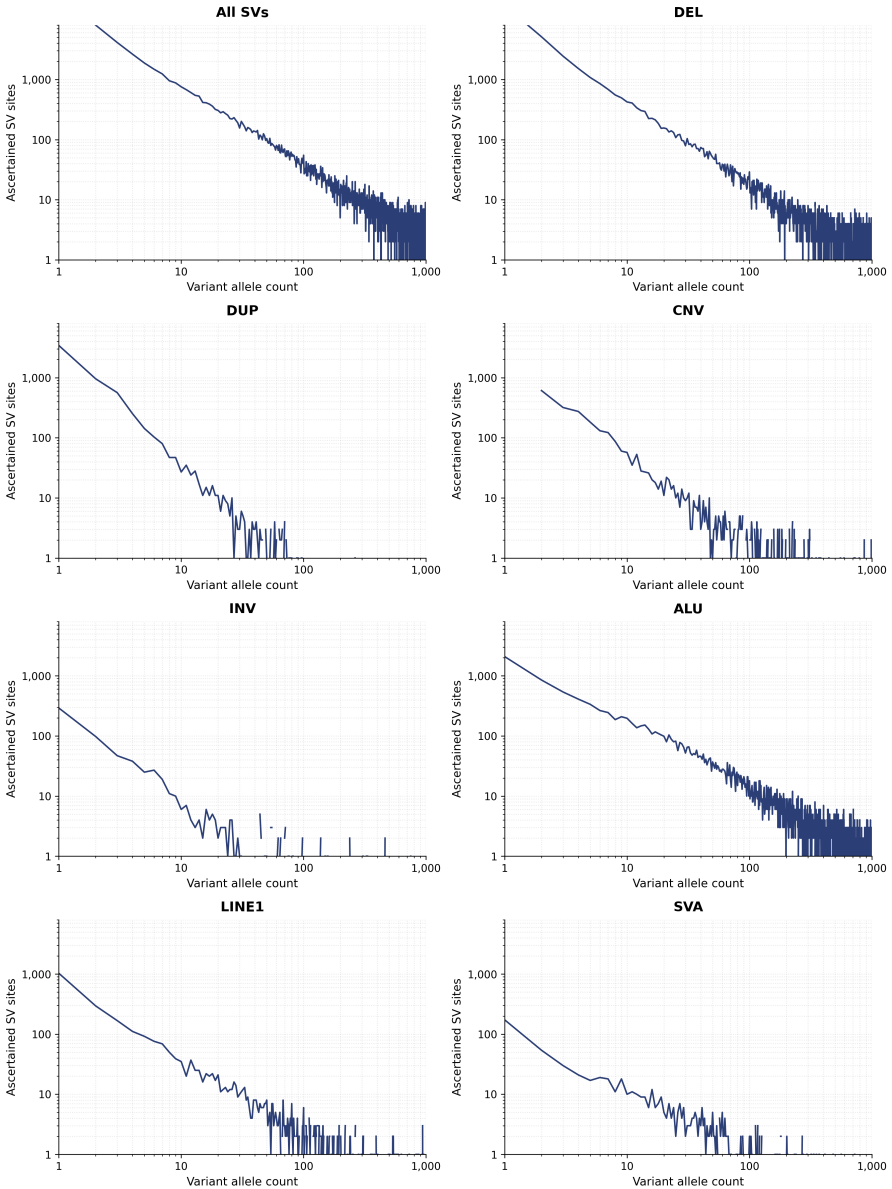
Results: Global Site Frequency Spectrum (SFS)



Conclusion: Human structural variation is heavily dominated by rare alleles ($< 1\%$ frequency), reflecting recent population expansions and purifying selection.

Results: SFS Distribution Across Variant Classes

Replicated Distributions



Original Paper Framework [1]

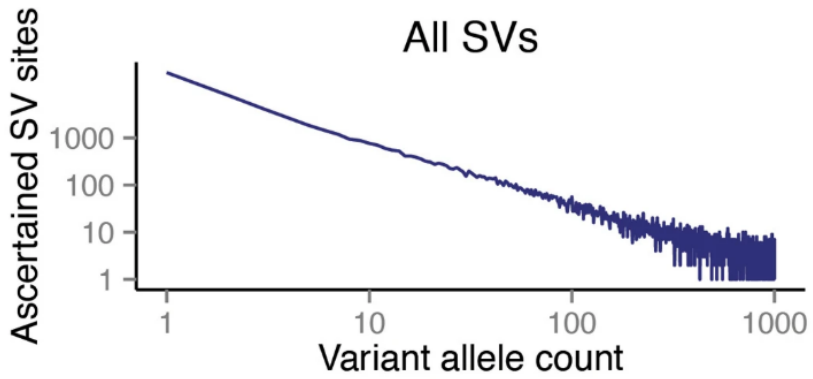


Figure: Reference variant frequency landscape from Sudmant et al. (2015).

Results: Mobile Element Insertion (MEI) Profiling

Core MEI Lineages:

- **Alu Elements:** Short Interspersed Nuclear Elements (~300 bp) that depend on external cellular machinery to copy and paste themselves.
- **LINE-1 Elements:** Long Interspersed Nuclear Elements (~6 kb) that encode autonomous replication machinery driving retrotransposition.
- **SVA Elements:** Composite retrotransposons (SINE-VNTR-Alu) that represent evolutionary young, hominid-specific insertion sequences.

Table: Observed MEI Class Counts

MEI Type	**Observed Count**
Alu	12,748
LINE-1	3,048
SVA	835
Total MEIs	**16,631**

Conclusion and Future Scope

Conclusion

- **Successful Replication:** Validated ∼68,818 high-confidence SVs, averaging ∼3,461 variants per individual genome.
- **Population Specificity:** Confirmed >80% of SVs are rare and population-specific, peaking in African lineages.
- **Ancestral Signal:** Demonstrated SV genotypes cleanly resolve global continental clusters, mirroring SNP-based resolution.

Future Scope: Focus on Indian Genomics

- **Localized SV Database:** Establish a high-resolution structural variant map tailored to unique Indian population architectures.
- **Targeted Pharmacogenomics:** Leverage regional SV signatures to optimize drug efficacy and mitigate adverse clinical reactions locally.

References



The 1000 Genomes Project Consortium, *An integrated map of structural variation in 2,504 human genomes*, Nature, 526(7571), 75-81, 2015.



The 1000 Genomes Project Consortium, *1000 Genomes Project Phase 3 Variant Calls (Data Set PRJEB6930)*, European Variant Archive (EVA), EMBL-EBI, `https://ftp.ebi.ac.uk/pub/databases/eva/PRJEB6930/`.



BCFtools Documentation, *Utilities for variant calling and manipulating VCFs and BCFs*, `https://samtools.github.io/bcftools/bcftools.html`.



Pysam Developers, *Pysam: A Python module for reading, manipulating and writing genomics data sets*, `https://github.com/pysam-developers/pysam`.

Thank You

Supplementary Slide

Populations Sampled

#	Population Name	Code
1	African Ancestry in Southwest US	ASW
2	African Caribbean in Barbados	ACB
3	Bengali in Bangladesh	BEB
4	British in England and Scotland	GBR
5	Chinese Dai in Xishuangbanna, China	CDX
6	Colombian in Medellin, Colombia	CLM
7	Esan in Nigeria	ESN
8	Finnish in Finland	FIN
9	Gambian in Western Division, The Gambia	GWD
10	Gujarati Indian in Houston, TX	GIH
11	Han Chinese in Beijing, China	CHB
12	Iberian populations in Spain	IBS
13	Indian Telugu in the UK	ITU
14	Japanese in Tokyo, Japan	JPT
15	Kinh in Ho Chi Minh City, Vietnam	KHV
16	Luhya in Webuye, Kenya	LWK
17	Mende in Sierra Leone	MSL
18	Mexican Ancestry in Los Angeles, California	MXL
19	Peruvian in Lima, Peru	PEL
20	Puerto Rican in Puerto Rico	PUR
21	Punjabi in Lahore, Pakistan	PJL
22	Southern Han Chinese, China	CHS
23	Sri Lankan Tamil in the UK	STU
24	Toscani in Italy	TSI
25	Utah residents with Northern and Western European ancestry	CEU
26	Yoruba in Ibadan, Nigeria	YRI

Continental groups*

Group	Code
South Asian	SAS
European	EUR
East Asian	EAS
American	AMR
African	AFR

**population memberships are resolved in Sample_IDs sheet*