Genome Sequencing

III: Beyond 1000 Genomes

MMG 835, SPRING 2016
Eukaryotic Molecular Genetics

George I. Mias

Department of Biochemistry and Molecular Biology
gmias@msu.edu

Sequencing Populations

International HapMap Project

http://www.hapmap.org/

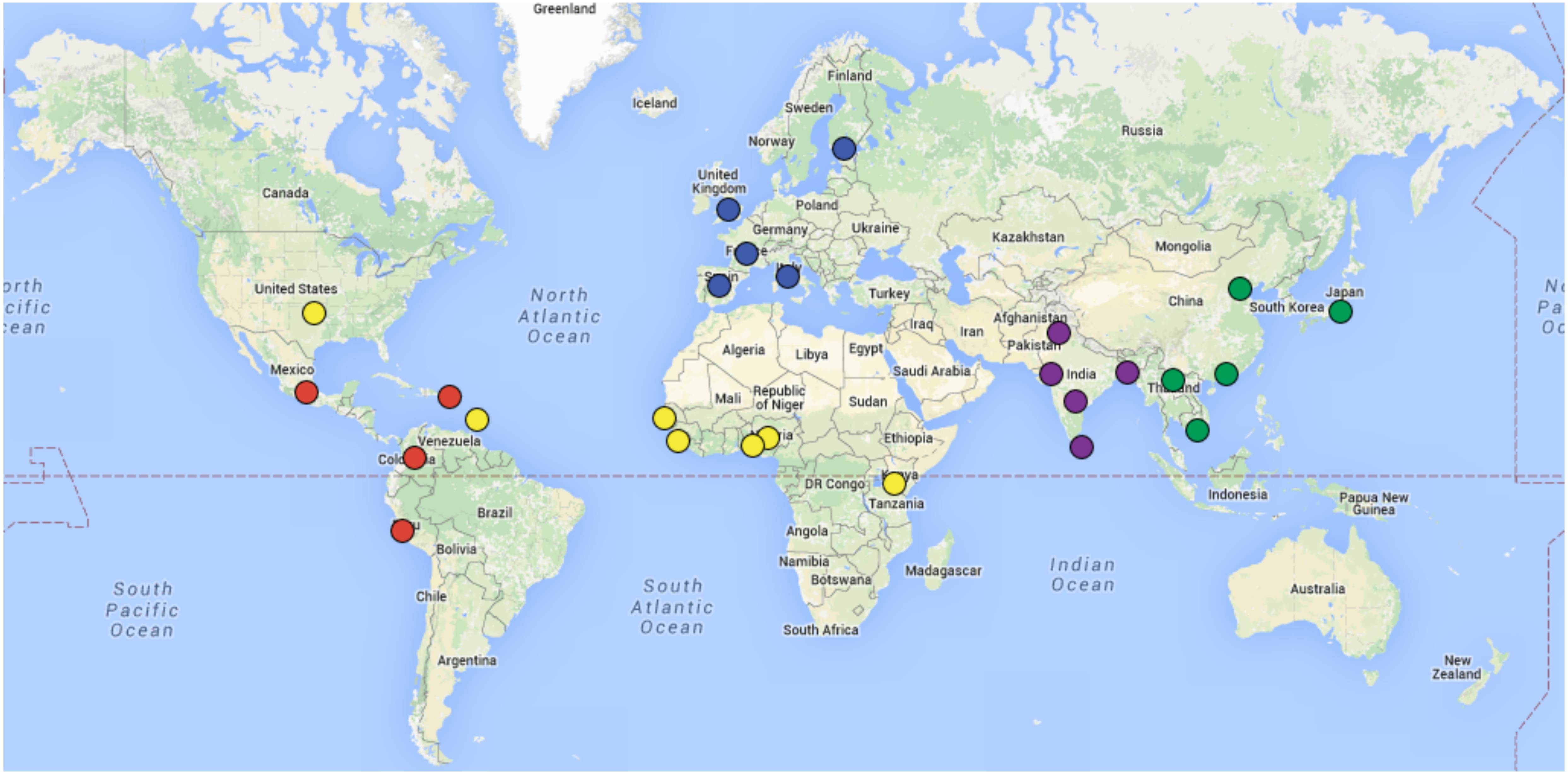
1000 Genomes Project

100k Genomes Project

Everyone Genomes Project

1000 Genomes Project

1000 Genomes Project



Populations: 🟡 - African; 🔴 - American; 🟢 - East Asian; 🔵 - European; 🟣 - South Asian;

1000genomes.org

1000 Genomes Project

- **Goal: find most genetic variants with frequencies of at least 1% in the populations studied.**
- Utilized new sequencing technology.
- Public Data.
- Project planned to sequence each sample to 4x genome coverage to allow the detection of most variants with frequencies as low as 1%.
- Multi-sample approach: Data from 2,504 samples combined.
- 26 populations
- 2008-2015

1000genomes.org

1000 Genomes Project

Pilot Project

Pilot	Purpose	Coverage	Strategy	Status
1 - low coverage	Assess strategy of sharing data across samples	2-4X	Whole-genome sequencing of 180 samples	Sequencing completed October 2008
2 - trios	Assess coverage and platforms and centres	20-60X	Whole-genome sequencing of 2 mother-father-adult child trios	Sequencing completed October 2008
3 - gene regions	Assess methods for gene-region-capture	50X	1000 gene regions in 900 samples	Sequencing completed June 2009

1000genomes.org

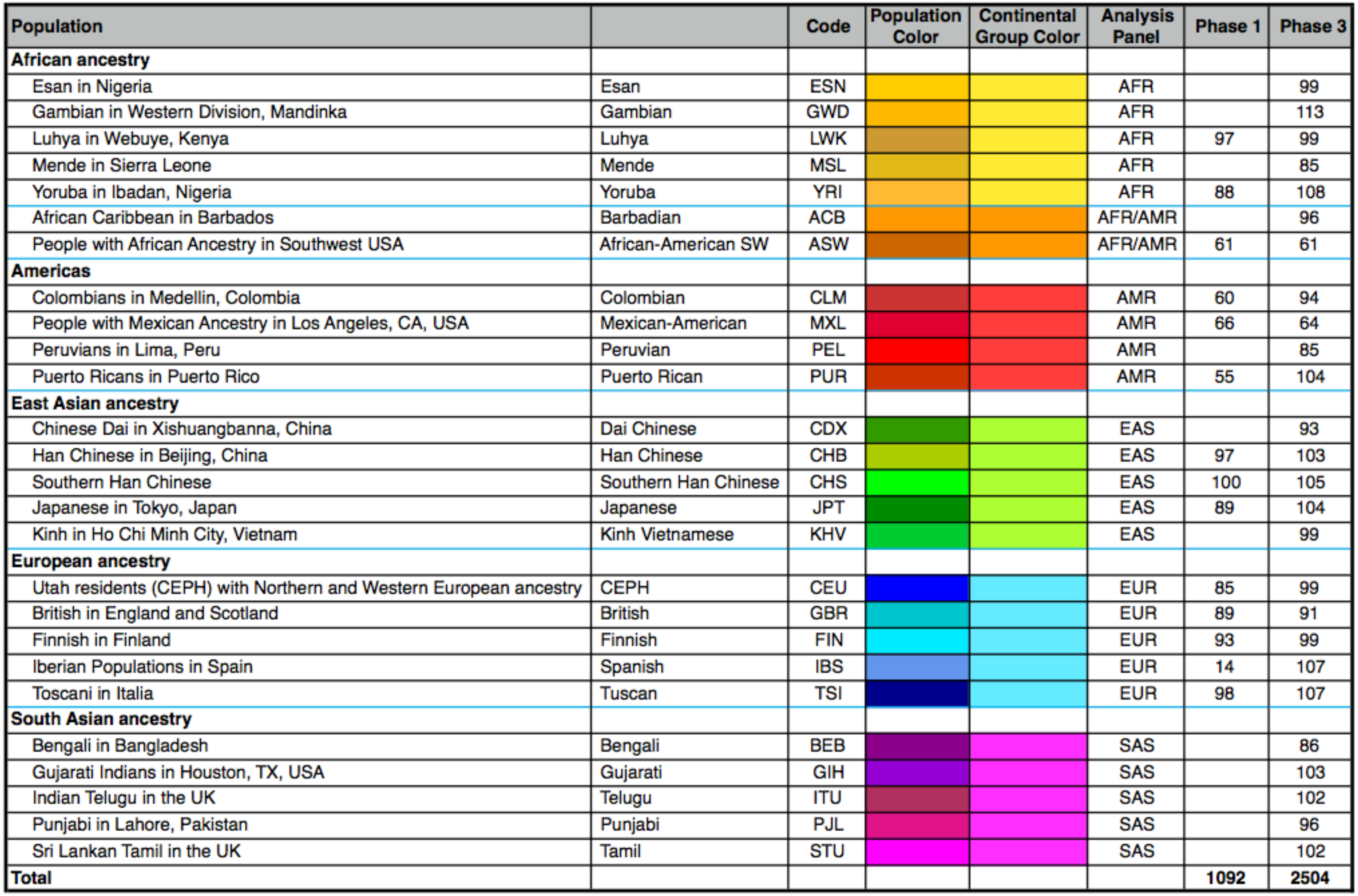
1000 Genomes Project

Main Project

- Data Freeze - 2nd May 2013.
- Multi-sample approach: Data from 2,504 samples combined.
- 26 populations.
- Low coverage and exome sequence data.
- 24 individuals sequenced to high coverage (validation).
- Results Published in 2015

1000genomes.org
The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).
Sudmant et al.Nature 526,75–81 (2015).

1000 Genomes Project



Population		Code	Population Color	Continental Group Color	Analysis Panel	Phase 1	Phase 3
African ancestry							
Esan in Nigeria	Esan	ESN	🟨	🟨	AFR		99
Gambian in Western Division, Mandinka	Gambian	GWD	🟧	🟨	AFR		113
Luhya in Webuye, Kenya	Luhya	LWK	🟫	🟨	AFR	97	99
Mende in Sierra Leone	Mende	MSL	🟨	🟨	AFR		85
Yoruba in Ibadan, Nigeria	Yoruba	YRI	🟧	🟨	AFR	88	108
African Caribbean in Barbados	Barbadian	ACB	🟧	🟧	AFR/AMR		96
People with African Ancestry in Southwest USA	African-American SW	ASW	🟫	🟧	AFR/AMR	61	61
Americas							
Colombians in Medellin, Colombia	Colombian	CLM	🟥	🟥	AMR	60	94
People with Mexican Ancestry in Los Angeles, CA, USA	Mexican-American	MXL	🟥	🟥	AMR	66	64
Peruvians in Lima, Peru	Peruvian	PEL	🟥	🟥	AMR		85
Puerto Ricans in Puerto Rico	Puerto Rican	PUR	🟥	🟥	AMR	55	104
East Asian ancestry							
Chinese Dai in Xishuangbanna, China	Dai Chinese	CDX	🟩	🟩	EAS		93
Han Chinese in Beijing, China	Han Chinese	CHB	🟩	🟩	EAS	97	103
Southern Han Chinese	Southern Han Chinese	CHS	🟩	🟩	EAS	100	105
Japanese in Tokyo, Japan	Japanese	JPT	🟩	🟩	EAS	89	104
Kinh in Ho Chi Minh City, Vietnam	Kinh Vietnamese	KHV	🟩	🟩	EAS		99
European ancestry							
Utah residents (CEPH) with Northern and Western European ancestry	CEPH	CEU	🟦	🟦	EUR	85	99
British in England and Scotland	British	GBR	🟦	🟦	EUR	89	91
Finnish in Finland	Finnish	FIN	🟦	🟦	EUR	93	99
Iberian Populations in Spain	Spanish	IBS	🟦	🟦	EUR	14	107
Toscani in Italia	Tuscan	TSI	🟦	🟦	EUR	98	107
South Asian ancestry							
Bengali in Bangladesh	Bengali	BEB	🟪	🟪	SAS		86
Gujarati Indians in Houston, TX, USA	Gujarati	GIH	🟪	🟪	SAS		103
Indian Telugu in the UK	Telugu	ITU	🟪	🟪	SAS		102
Punjabi in Lahore, Pakistan	Punjabi	PJL	🟪	🟪	SAS		96
Sri Lankan Tamil in the UK	Tamil	STU	🟪	🟪	SAS		102
Total						**1092**	**2504**

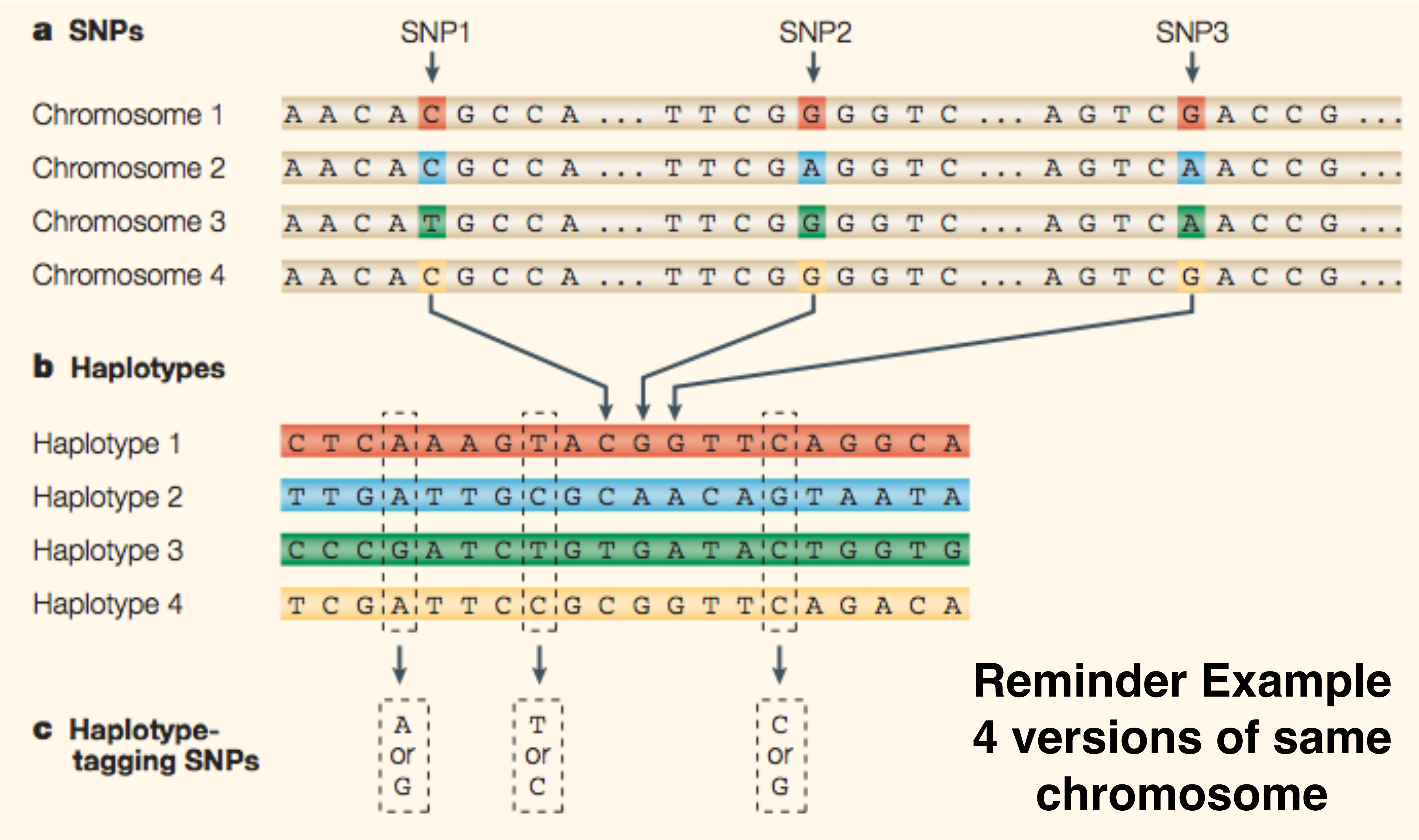
The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

1000 Genomes Project

- **All individuals:**
  - Whole-genome sequencing (mean depth 7.4X)
  - Targeted exome sequencing (mean depth 65.7X)
  - Individuals & available first-degree relatives (generally, adult offspring) genotyped with high-density SNP microarrays.

The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

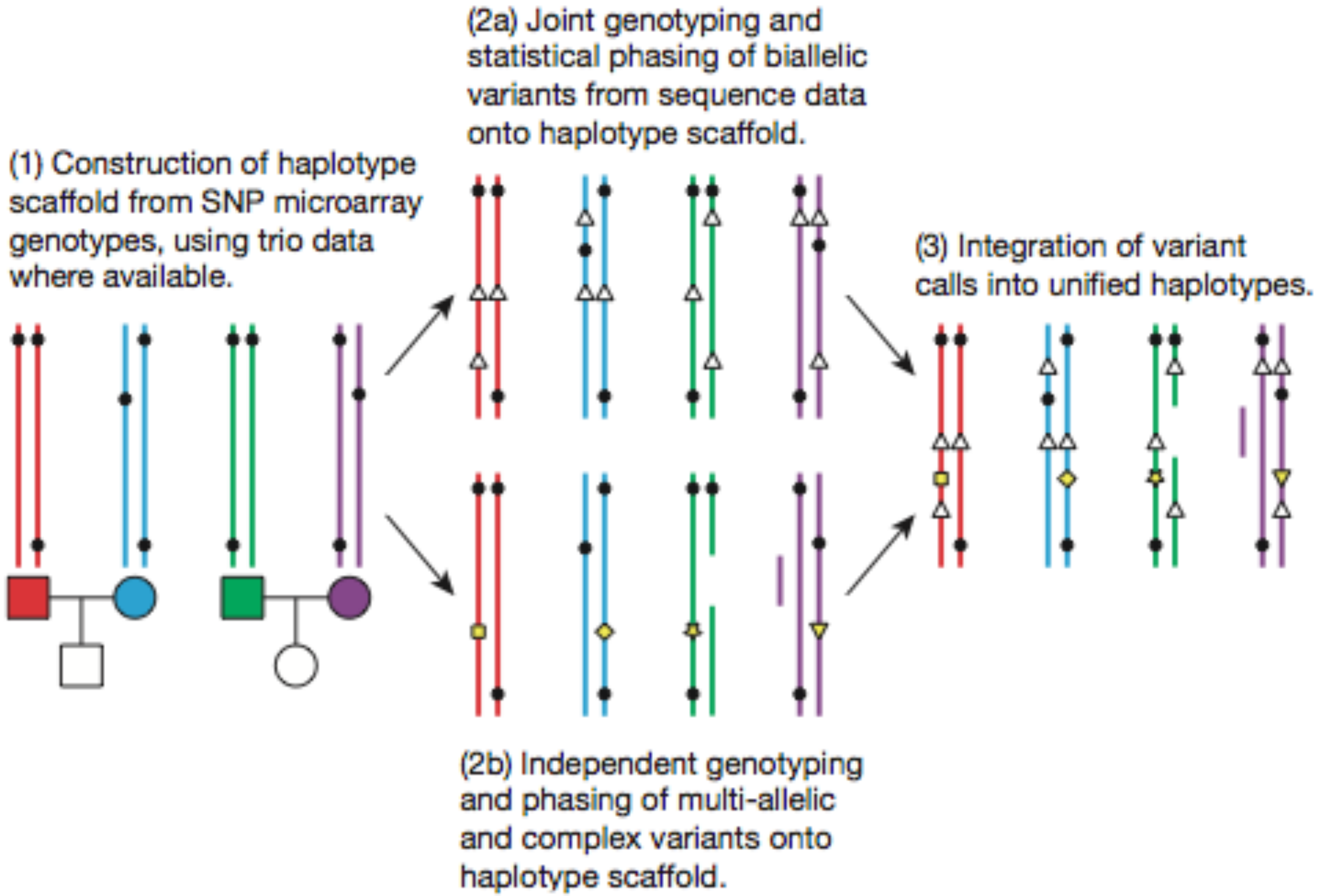
1000 Genomes Project


a SNPs
SNP1
SNP2
SNP3
Chromosome 1
A A C A C G C C A ... T T C G G G G T C ... A G T C G A C C G ...
Chromosome 2
A A C A C G C C A ... T T C G A G G T C ... A G T C A A C C G ...
Chromosome 3
A A C A T G C C A ... T T C G G G G T C ... A G T C A A C C G ...
Chromosome 4
A A C A C G C C A ... T T C G G G G T C ... A G T C G A C C G ...
b Haplotypes
Haplotype 1
C T C A A A G T A C G G T T C A G G C A
Haplotype 2
T T G A T T G C G C A A C A G T A A T A
Haplotype 3
C C C G A T C T G T G A T A C T G G T G
Haplotype 4
T C G A T T C C G C G G T T C A G A C A
c Haplotype-tagging SNPs
A or G
T or C
C or G

Reminder Example
4 versions of same chromosome

Hafler & De Jager, Nature Reviews Immunology 5, 83-91 (2005)

1000 Genomes Project



(2a) Joint genotyping and statistical phasing of biallelic variants from sequence data onto haplotype scaffold.

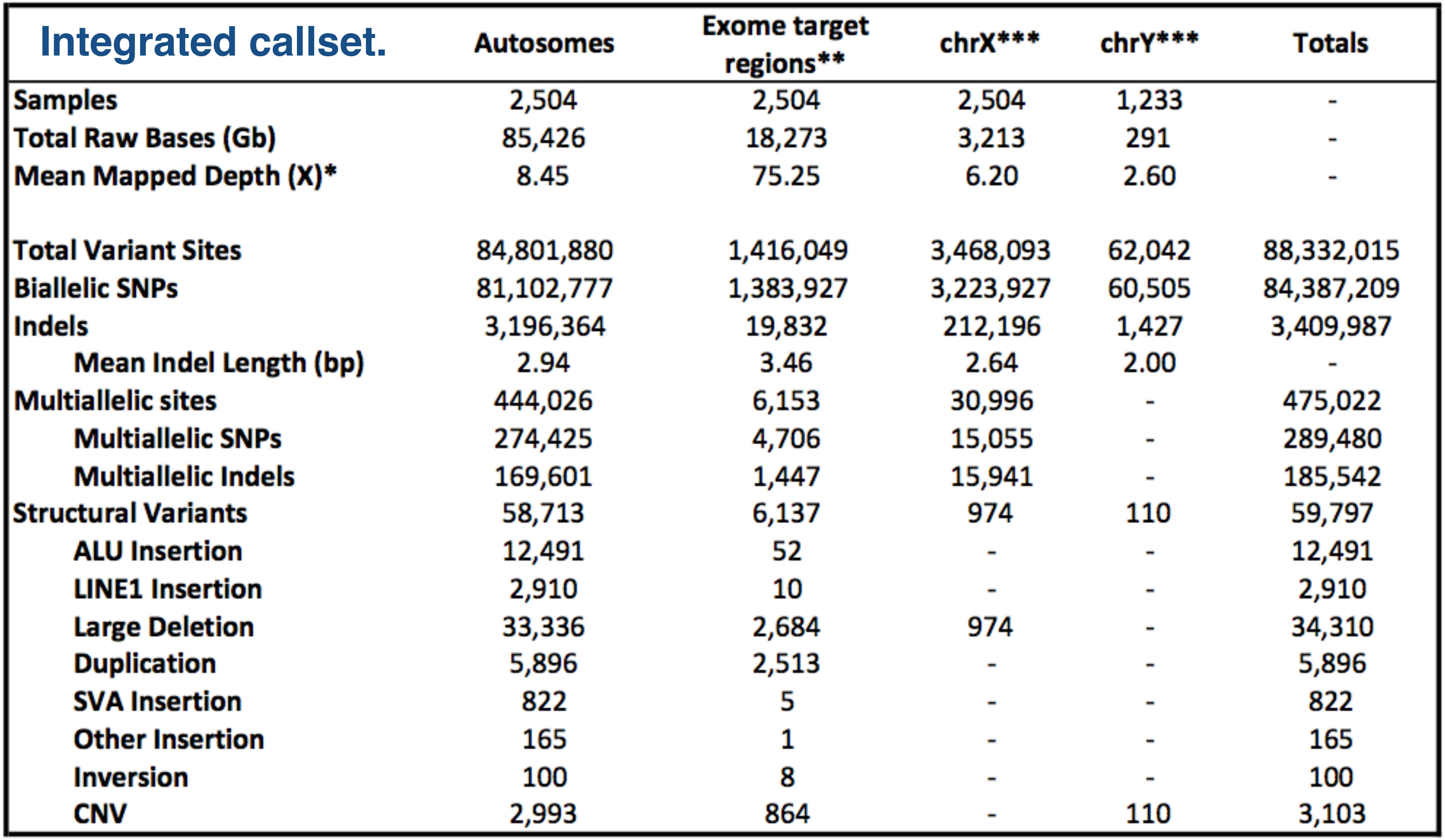
(1) Construction of haplotype scaffold from SNP microarray genotypes, using trio data where available.

(3) Integration of variant calls into unified haplotypes.

(2b) Independent genotyping and phasing of multi-allelic and complex variants onto haplotype scaffold.

The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

1000 Genomes Project

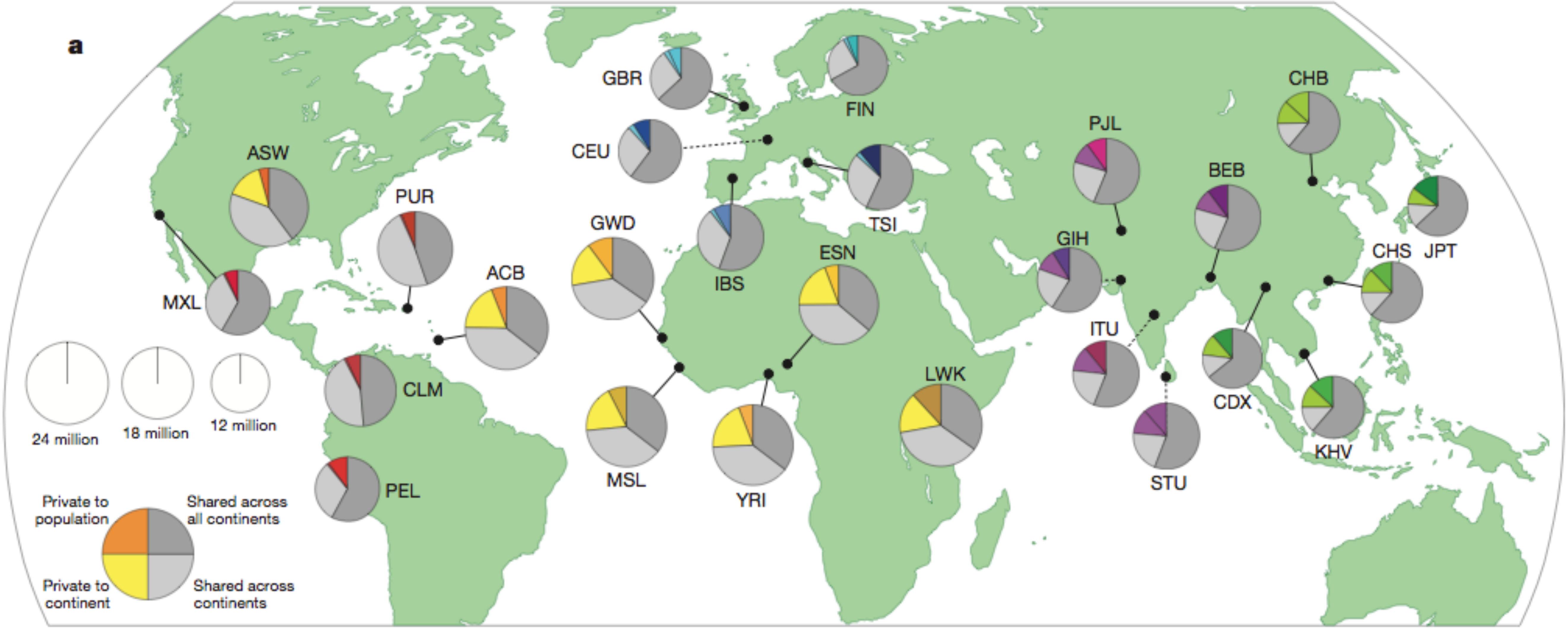


Integrated callset.	Autosomes	Exome target regions**	chrX***	chrY***	Totals
Samples	2,504	2,504	2,504	1,233	-
Total Raw Bases (Gb)	85,426	18,273	3,213	291	-
Mean Mapped Depth (X)*	8.45	75.25	6.20	2.60	-
Total Variant Sites	84,801,880	1,416,049	3,468,093	62,042	88,332,015
Biallelic SNPs	81,102,777	1,383,927	3,223,927	60,505	84,387,209
Indels	3,196,364	19,832	212,196	1,427	3,409,987
Mean Indel Length (bp)	2.94	3.46	2.64	2.00	-
Multiallelic sites	444,026	6,153	30,996	-	475,022
Multiallelic SNPs	274,425	4,706	15,055	-	289,480
Multiallelic Indels	169,601	1,447	15,941	-	185,542
Structural Variants	58,713	6,137	974	110	59,797
ALU Insertion	12,491	52	-	-	12,491
LINE1 Insertion	2,910	10	-	-	2,910
Large Deletion	33,336	2,684	974	-	34,310
Duplication	5,896	2,513	-	-	5,896
SVA Insertion	822	5	-	-	822
Other Insertion	165	1	-	-	165
Inversion	100	8	-	-	100
CNV	2,993	864	-	110	3,103

The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

1000 Genomes Project

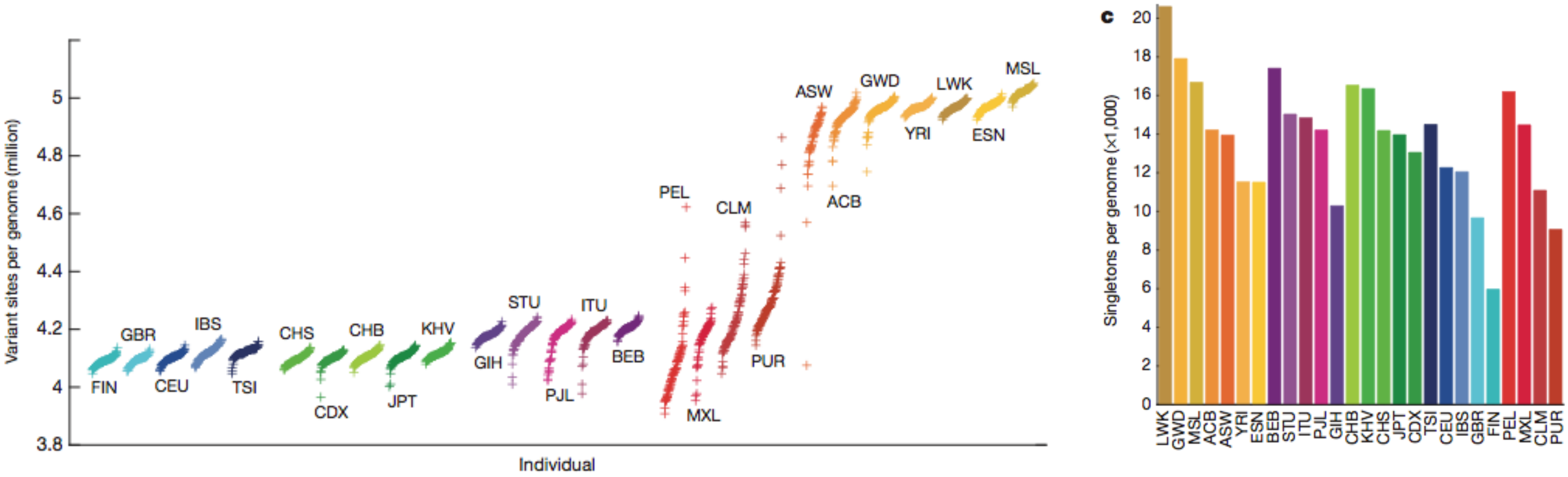
Polymorphic variants within populations.



The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

1000 Genomes Project

Variants Per Genome



Europe East Asia South Asian Americas Africa Ancestry

The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

1000 Genomes Project

Variants Per Genome

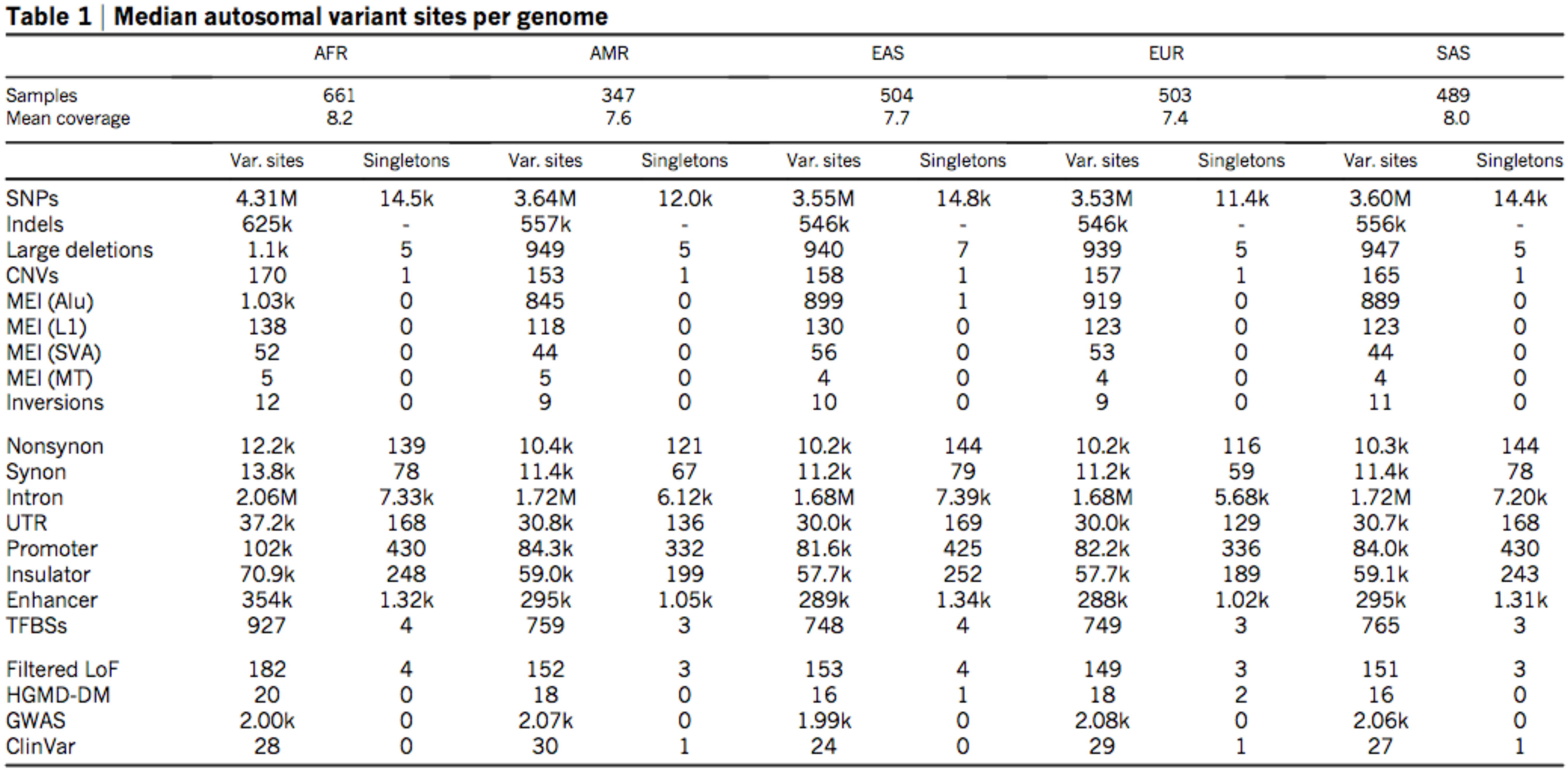


Table 1 | Median autosomal variant sites per genome

	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean coverage	8.2		7.6		7.7		7.4		8.0	
	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (L1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
Nonsynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBSs	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

1000 Genomes Project

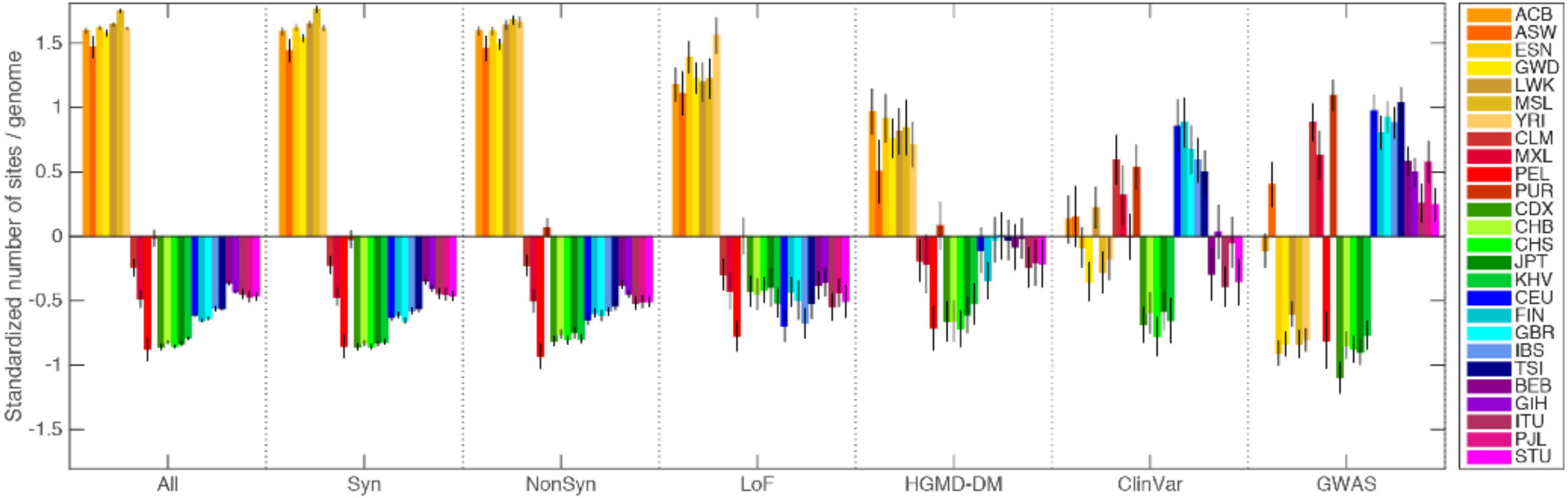
Variation Per Typical Genome

- **~ 4.1 million to 5.0 million sites**
- **~ 99.9% of variants are SNPs and short indels.**
- **Structural variants ~ 2,100 to 2,500 [20 Mb]**
    - ~1,000 large deletions, ,
    - ~160 copy-number variants,
    - ~ 915 Alu insertions,
    - ~ 128 L1 insertions,
    - ~ 51 SVA insertions,
    - ~ 4 NUMTs, and ,10 inversions)

The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

1000 Genomes Project

Variant sites per genome, partitioned by population and variant category



sites with a derived allele

synonymous sites with a derived allele

nonsynonymous sites with a derived allele

sites with a loss-of-function allele

sites with a HGMD disease mutation allele

sites with a ClinVar pathogenic variant

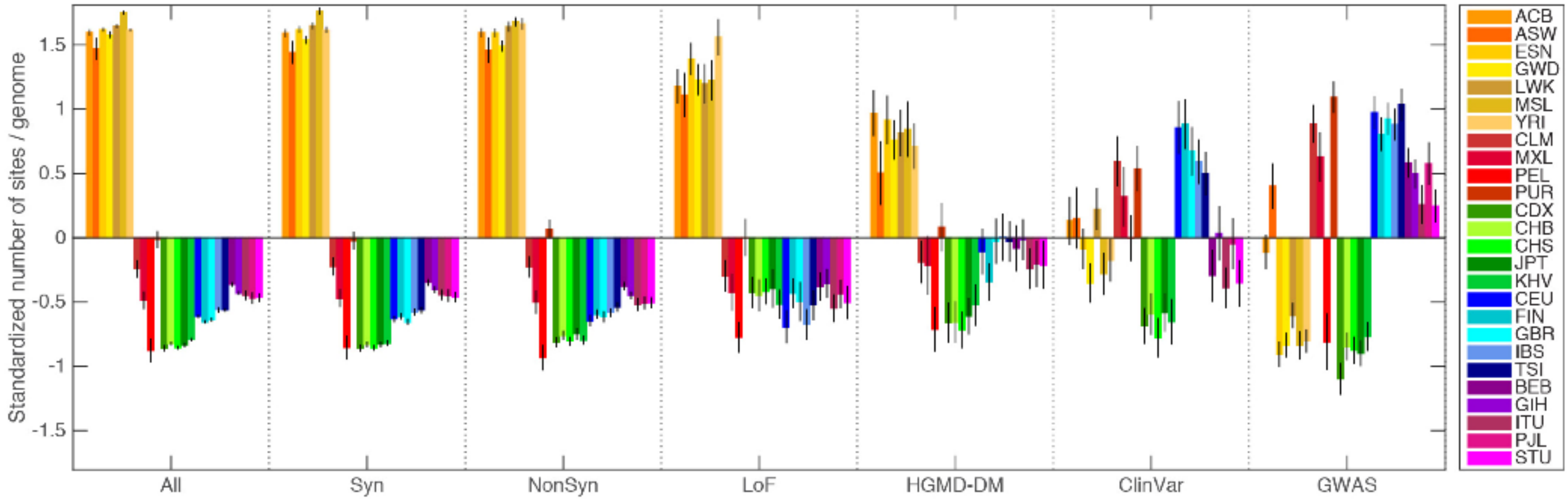
sites carrying a GWAS risk allele

HGMD: Human Gene Mutation Database

The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

1000 Genomes Project

Variant sites per genome, partitioned by population and variant category



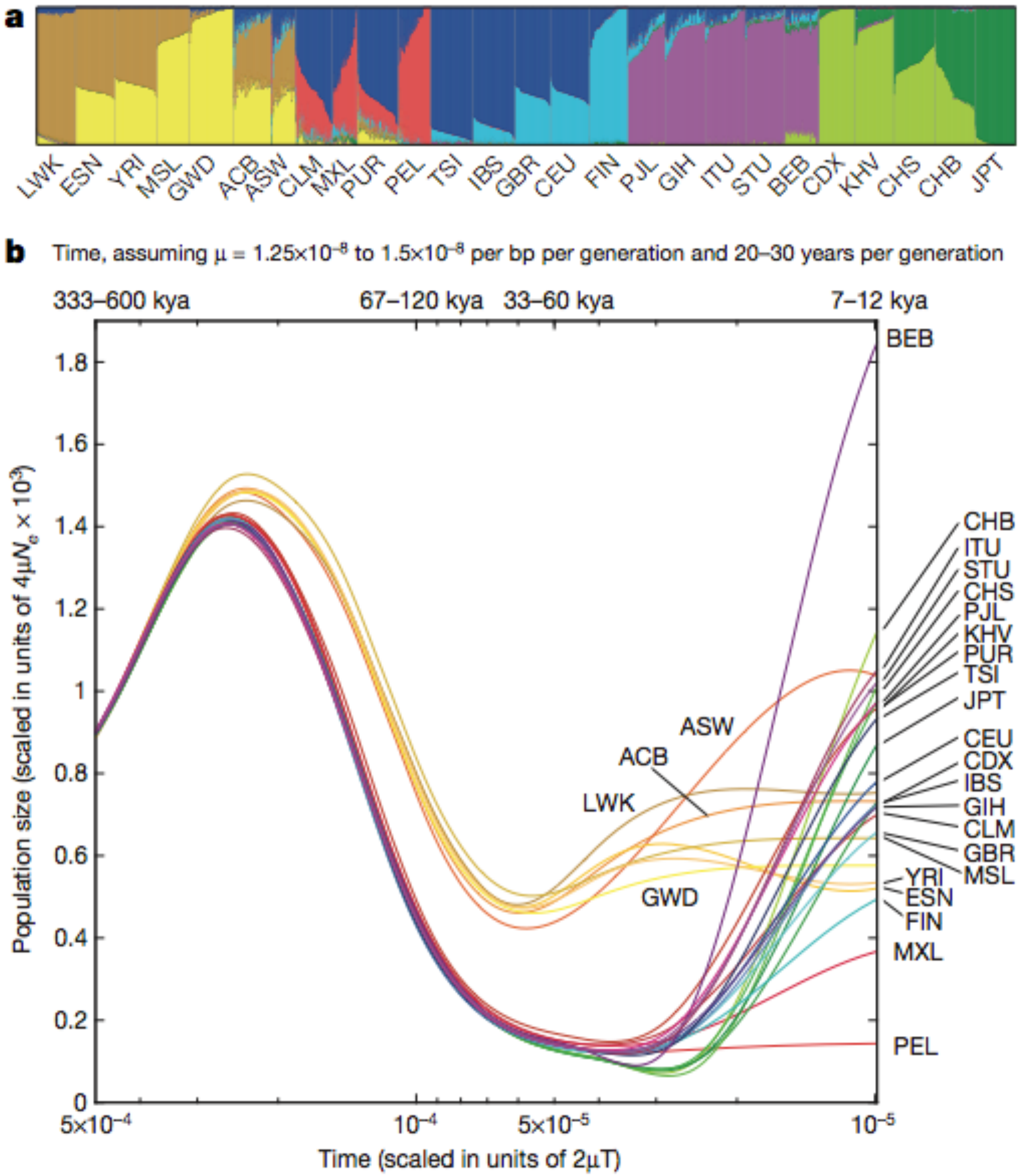
pattern of increased diversity in Africa

ethnic bias of
current genetic
studies

The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

1000 Genomes Project

Population Structure and Demography



shared demographic history > 150,000-200,000 years ago

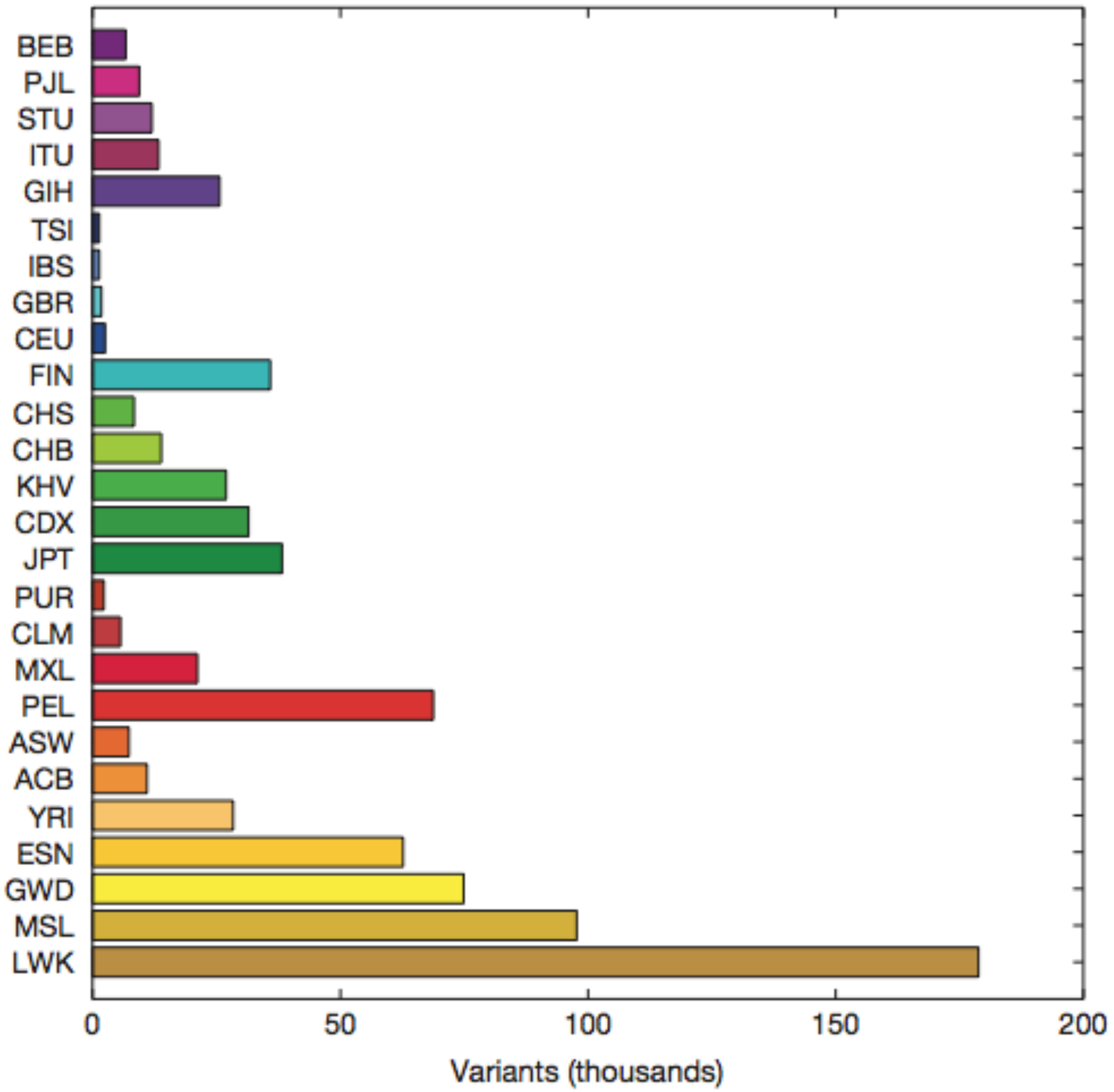
Note the bottlenecks

(large population reduction prior to recovery)

The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

1000 Genomes Project

Rare Variants



762,000 rare variants (frequency < 0.5%) within the global sample

but common (> 5%) within a population.

The1000 Genomes Project Consortium, Nature 526, 68–74 (2015).

1000 Genomes Project

Structural Variants

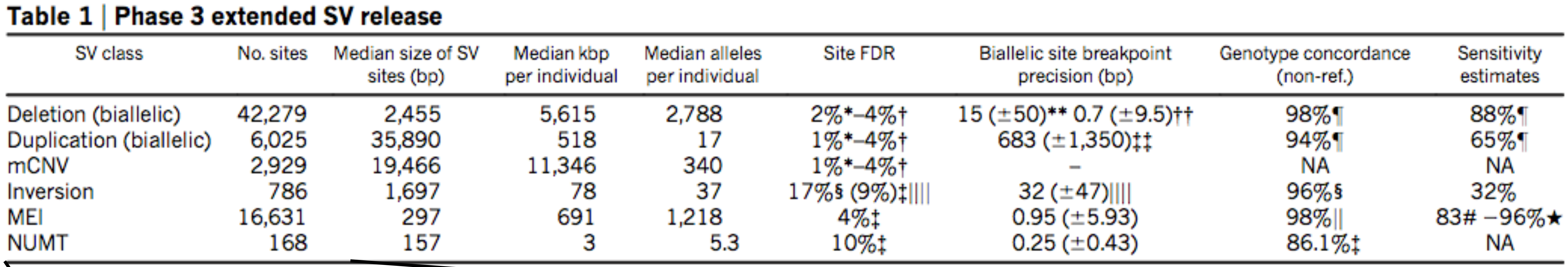


Table 1 | Phase 3 extended SV release

SV class	No. sites	Median size of SV sites (bp)	Median kbp per individual	Median alleles per individual	Site FDR	Biallelic site breakpoint precision (bp)	Genotype concordance (non-ref.)	Sensitivity estimates
Deletion (biallelic)	42,279	2,455	5,615	2,788	2%*–4%†	15 (±50)** 0.7 (±9.5)††	98%¶	88%¶
Duplication (biallelic)	6,025	35,890	518	17	1%*–4%†	683 (±1,350)‡‡	94%¶	65%¶
mCNV	2,929	19,466	11,346	340	1%*–4%†	–	NA	NA
Inversion	786	1,697	78	37	17%§ (9%)‡‖‖	32 (±47)‖‖	96%§	32%
MEI	16,631	297	691	1,218	4%‡	0.95 (±5.93)	98%‖	83# −96%★
NUMT	168	157	3	5.3	10%‡	0.25 (±0.43)	86.1%‡	NA

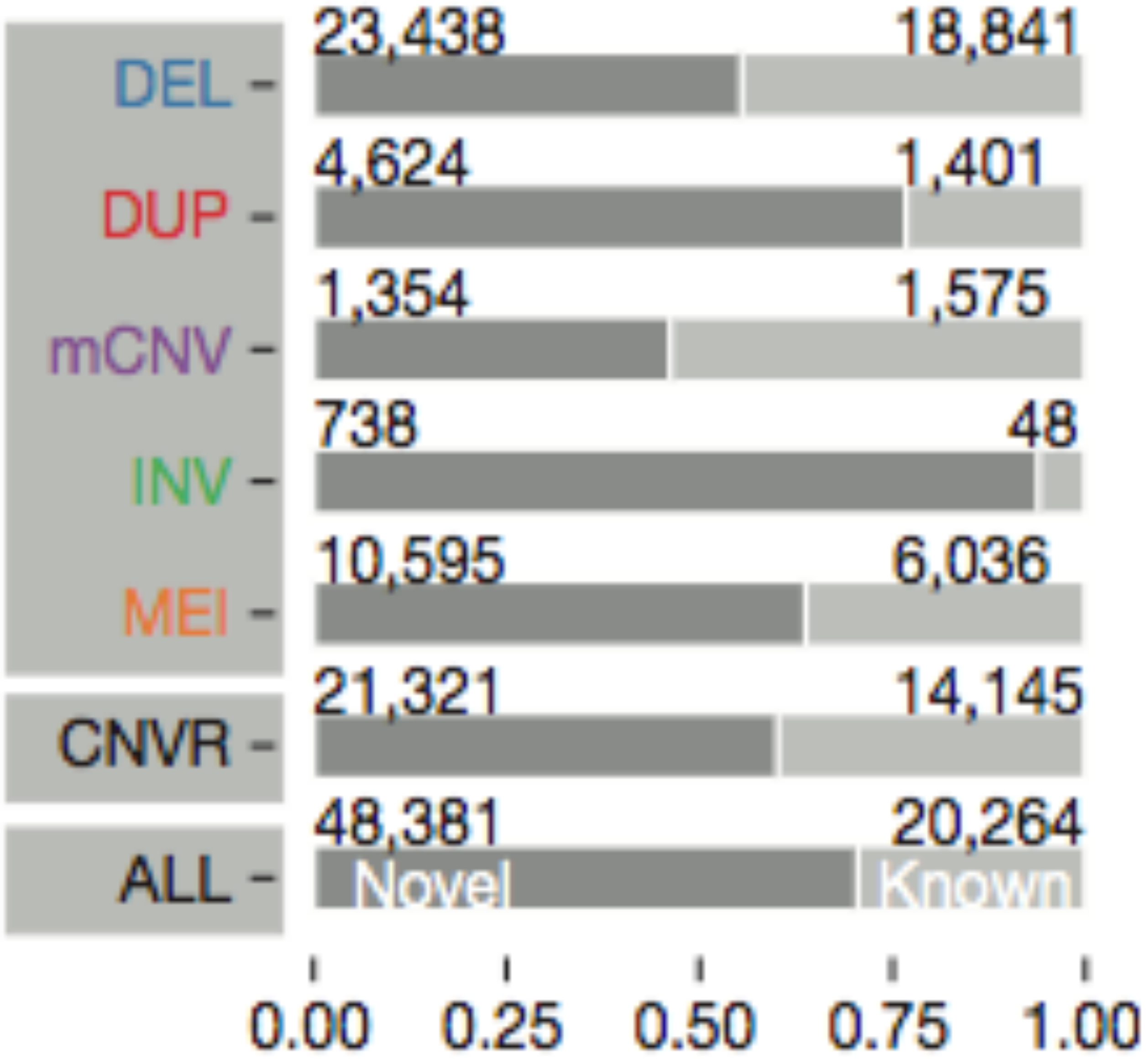
SV Class	No. sites
Deletion(biallelic)	42,279
Duplication (biallelic)	6,025
mCNVs (multi allelic copy-number variants)	2,929
Inversion	786
Mobile Element Insertion	16631
NUMT (nuclear mitochondrial insertions)	168

Sudmant et al.Nature 526,75–81 (2015).

1000 Genomes Project

Structural Variants

Novelty

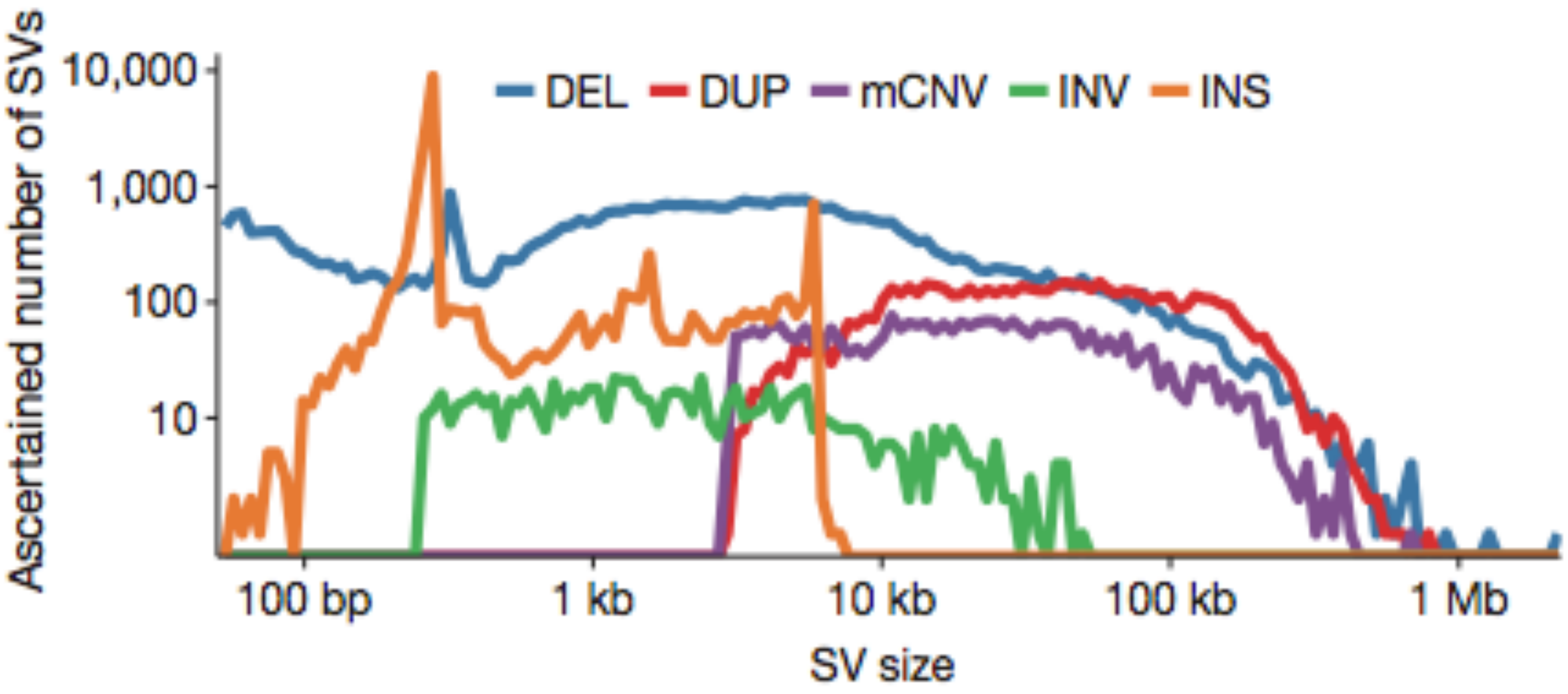


Sudmant et al.Nature 526,75–81 (2015).

1000 Genomes Project

Structural Variants

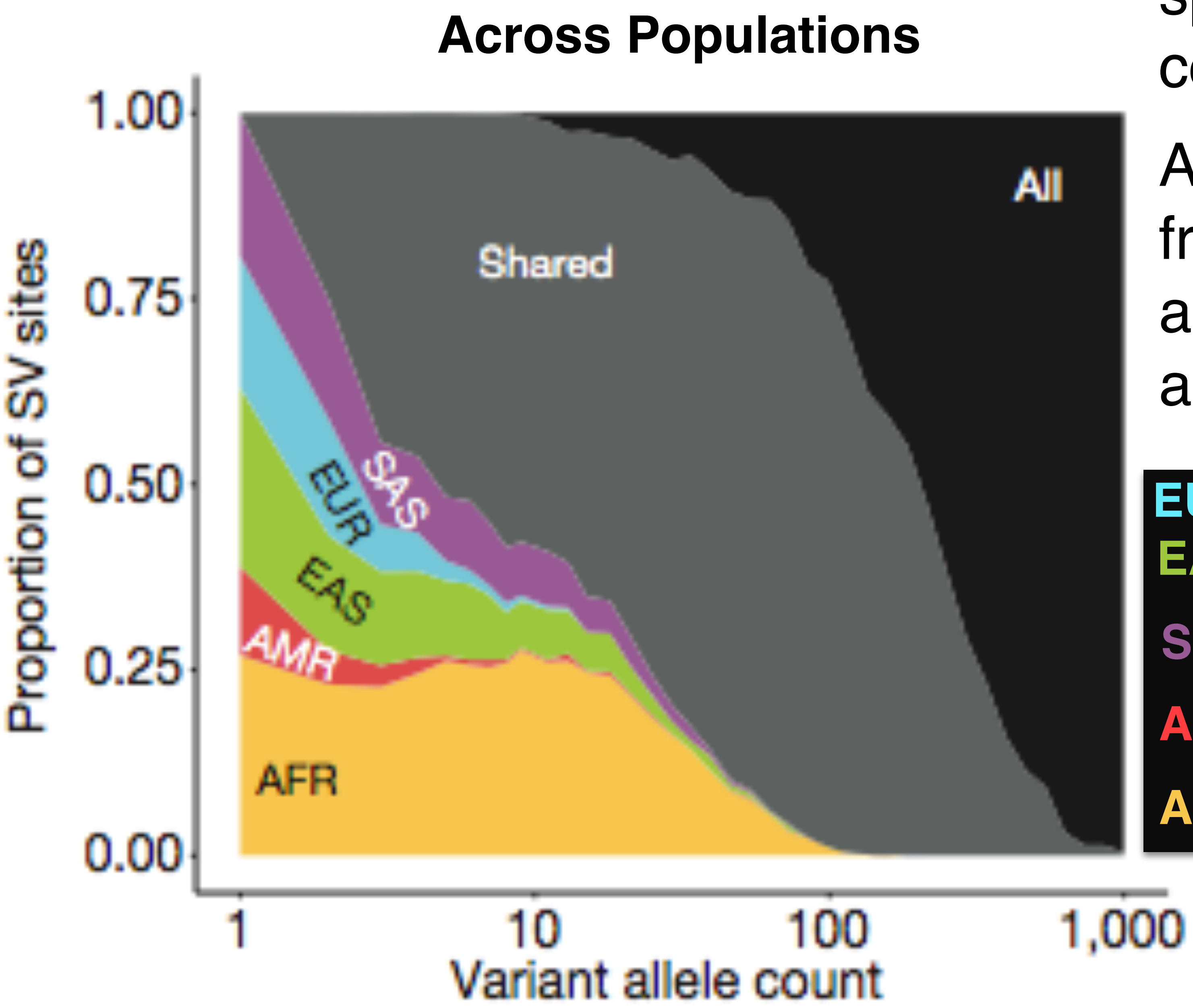
Size Distribution



Sudmant et al.Nature 526,75–81 (2015).

1000 Genomes Project

Structural Variants



Rare SVs typically specific to individual continental groups.

At variant allele frequency > 2% nearly all SVs are shared across continents.

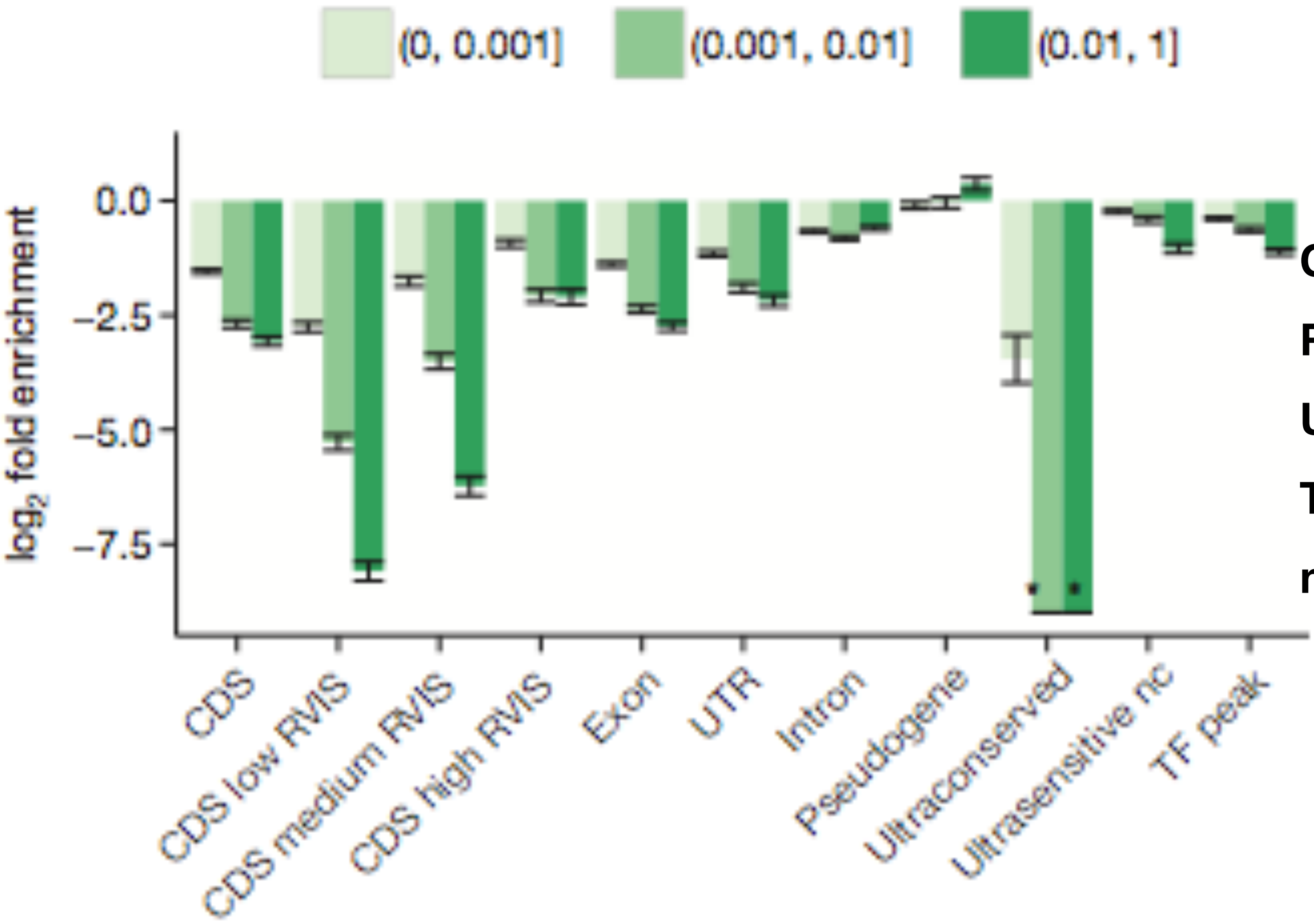
EUR, Europe
EAS, East Asia
SAS, South Asian
AMR, Americas
AFR, Africa

Sudmant et al.Nature 526,75–81 (2015).

1000 Genomes Project

Structural Variants

Functional impact enrichment by variant allele frequency



CDS, coding sequence

RVIS, residual variation intolerance score

UTR, untranslated regions

TF, transcription factor binding site

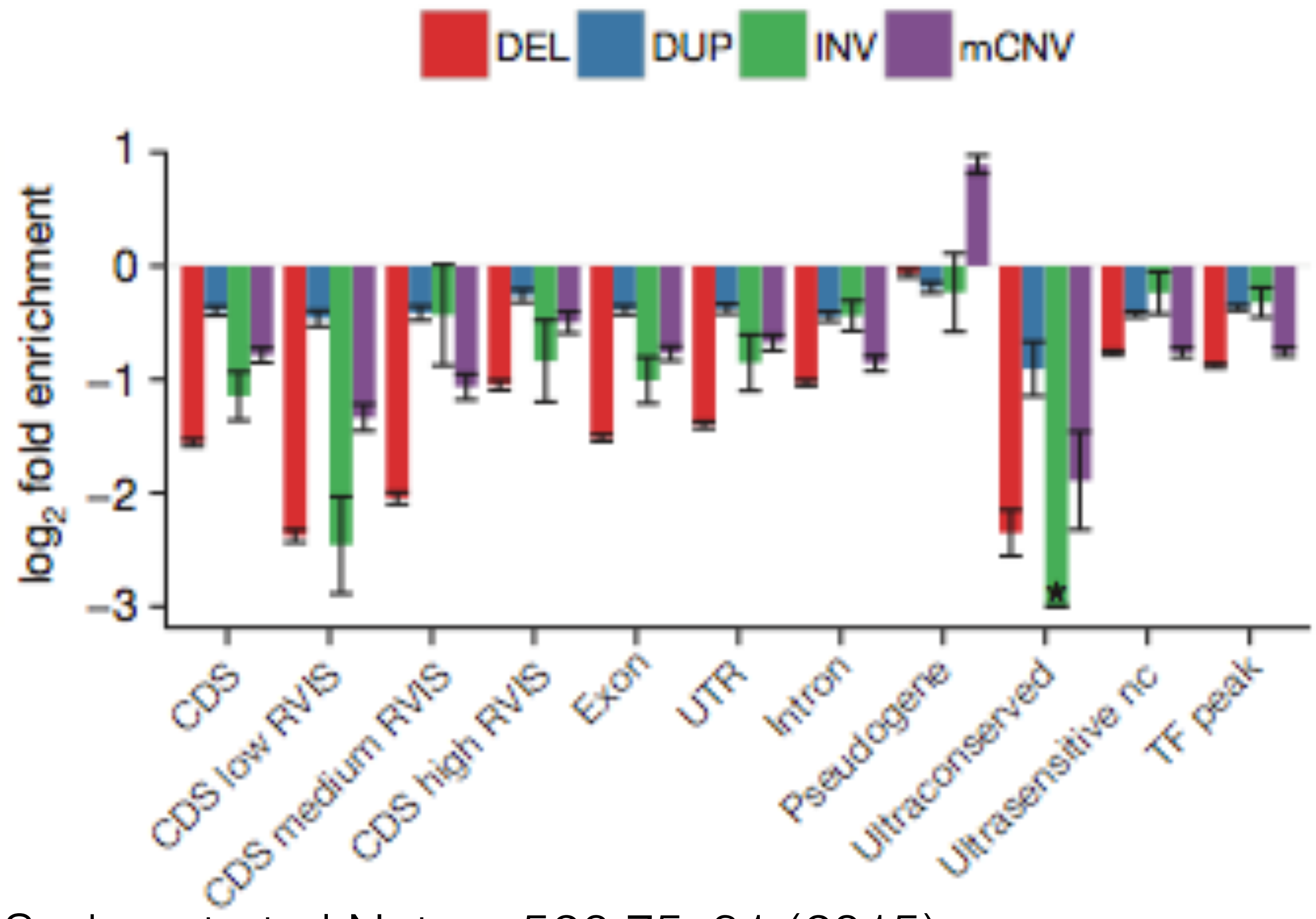
nc, non coding

Sudmant et al.Nature 526,75–81 (2015).

1000 Genomes Project

Structural Variants

Functional impact enrichment by type



CDS, coding sequence

RVIS, residual variation intolerance score

UTR, untranslated regions

TF, transcription factor binding site

nc, non coding

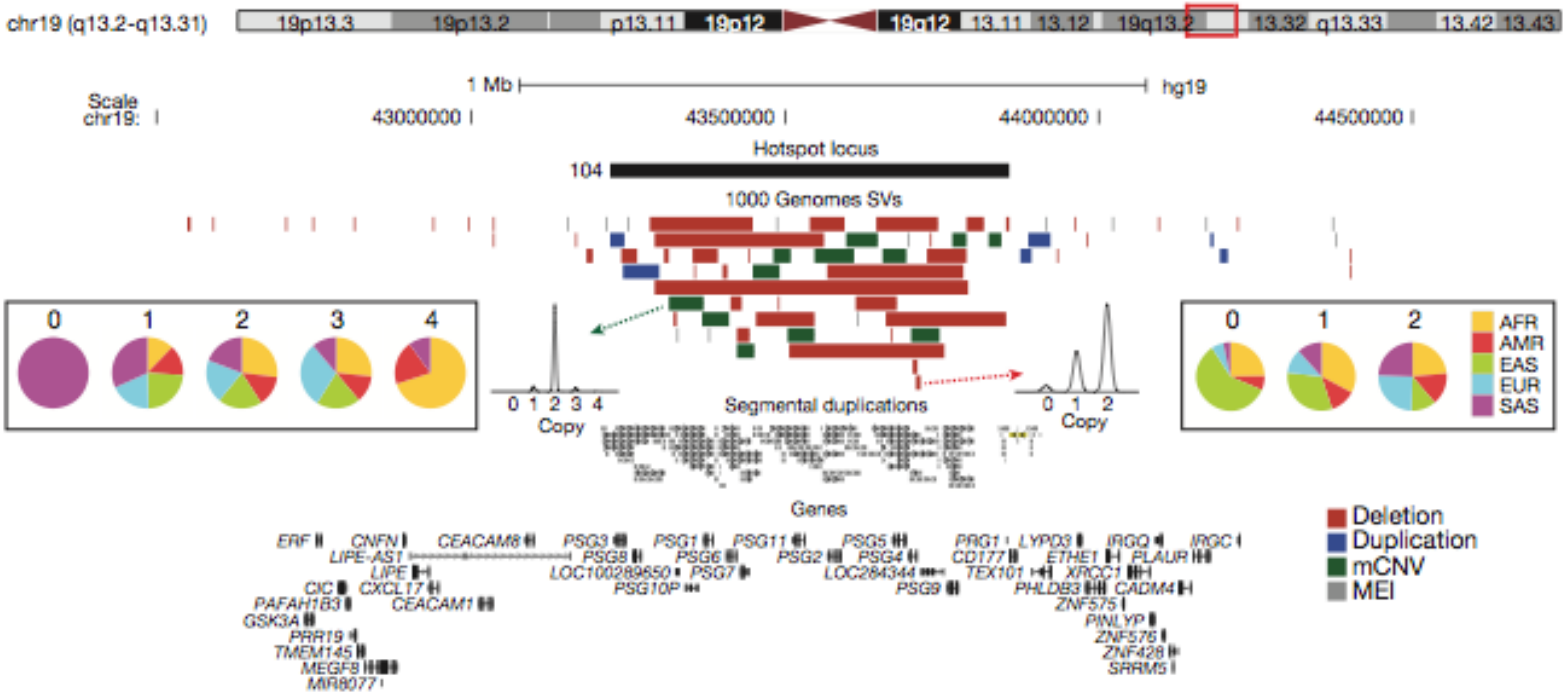
Sudmant et al.Nature 526,75–81 (2015).

1000 Genomes Project

Structural Variants

3,163 total regions where SVs cluster
(>2 SVs mapping within 500 bp)

Example: SV Clustering (47 SVs): pregnancy-specific glycoprotein (PSG) family

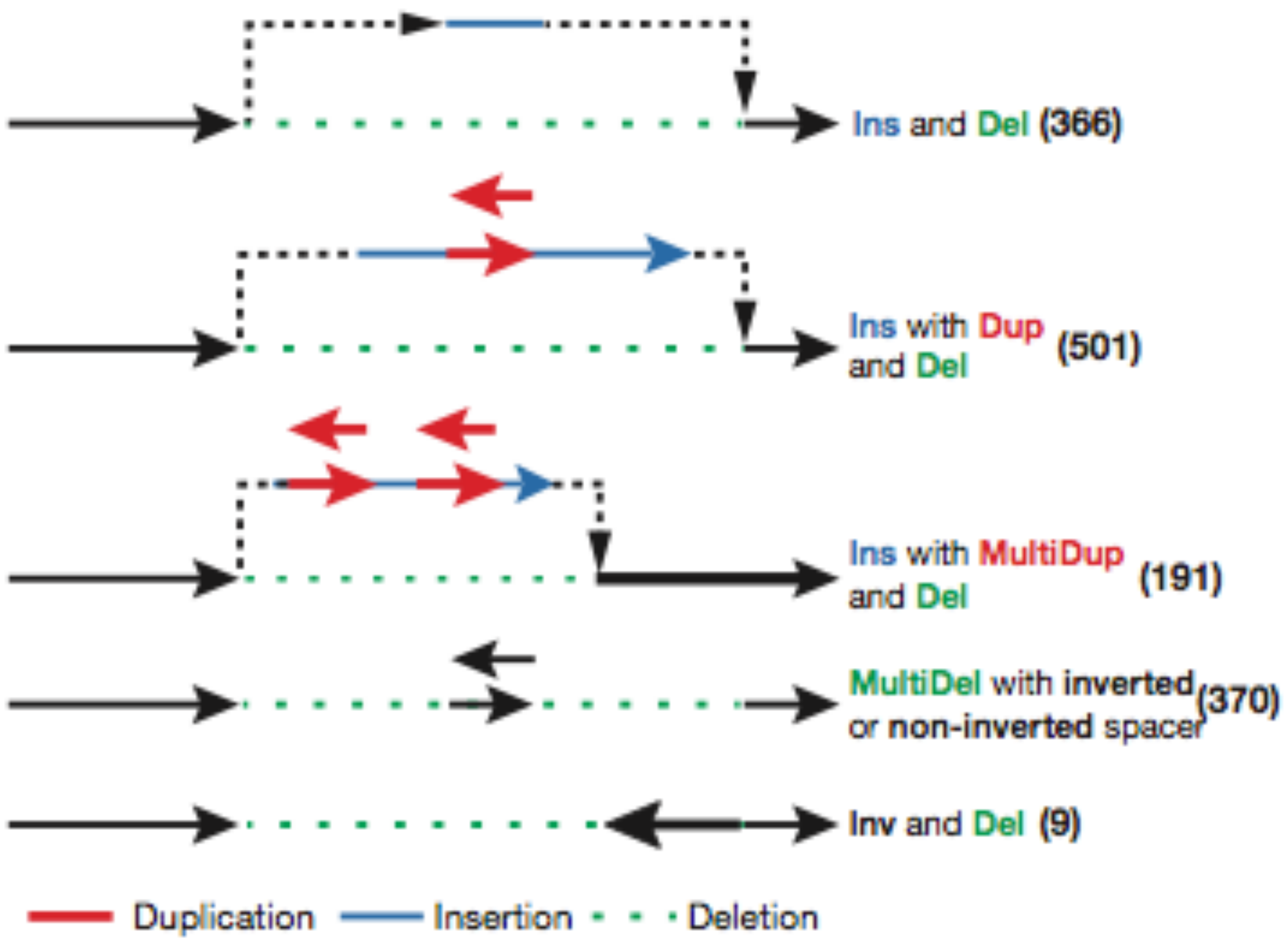


Sudmant et al.Nature 526,75–81 (2015).

1000 Genomes Project

Structural Variants

Complexity of Deletions: 5 Classes

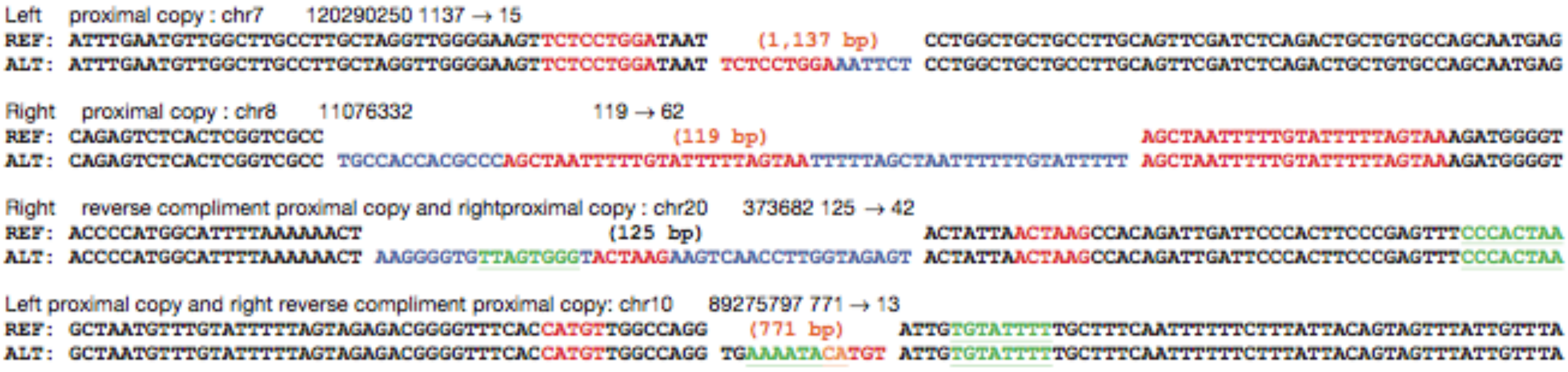

Ins and Del (366)
Ins with Dup and Del (501)
Ins with MultiDup and Del (191)
MultiDel with inverted or non-inverted spacer (370)
Inv and Del (9)
Duplication
Insertion
Deletion

Sudmant et al.Nature 526,75–81 (2015).

- Out of 29,954 deletions with resolved breakpoints 6% (1,822) intersect another deletion with distinct breakpoints.
- 16% (4,813) showed the presence of additional inserted sequence at deletion breakpoints.
- 1,651 deletions with mean size of 3.1 kbp and at least 10 bp of additional DNA sequence between the original SV site boundaries grouped into 5 classes (214 do not fit)

1000 Genomes Project

Structural Variants



```
Left    proximal copy : chr7      120290250 1137 → 15
REF:  ATTTGAATGTTGGCTTGCCTTGCTAGGTTGGGGAAGTTCTCCTGGATAAT    (1,137 bp)    CCTGGCTGCTGCCTTGCAGTTCGATCTCAGACTGCTGTGCCAGCAATGAG
ALT:  ATTTGAATGTTGGCTTGCCTTGCTAGGTTGGGGAAGTTCTCCTGGATAAT TCTCCTGGAAATTCT CCTGGCTGCTGCCTTGCAGTTCGATCTCAGACTGCTGTGCCAGCAATGAG

Right   proximal copy : chr8      11076332                     119 → 62
REF:  CAGAGTCTCACTCGGTCGCC                          (119 bp)                               AGCTAATTTTTGTATTTTTAGTAAAGATGGGGT
ALT:  CAGAGTCTCACTCGGTCGCC TGCCACCACGCCCAGCTAATTTTTGTATTTTTAGTAATTTTTAGCTAATTTTTTGTATTTTT AGCTAATTTTTGTATTTTTAGTAAAGATGGGGT

Right   reverse compliment proximal copy and rightproximal copy : chr20     373682 125 → 42
REF:  ACCCCATGGCATTTTAAAAAACT                     (125 bp)                   ACTATTAACTAAGCCACAGATTGATTCCCACTTCCCGAGTTTCCCACTAA
ALT:  ACCCCATGGCATTTTAAAAAACT AAGGGGTGTTAGTGGGTACTAAGAAGTCAACCTTGGTAGAGT ACTATTAACTAAGCCACAGATTGATTCCCACTTCCCGAGTTTCCCACTAA

Left proximal copy and right reverse compliment proximal copy: chr10     89275797 771 → 13
REF:  GCTAATGTTTGTATTTTTAGTAGAGACGGGGTTTCACCATGTTGGCCAGG    (771 bp)    ATTGTGTATTTTTGCTTTCAATTTTTTCTTTATTACAGTAGTTTATTGTTTA
ALT:  GCTAATGTTTGTATTTTTAGTAGAGACGGGGTTTCACCATGTTGGCCAGG TGAAAATACATGT ATTGTGTATTTTTGCTTTCAATTTTTTCTTTATTACAGTAGTTTATTGTTTA
```

REF, reference allele

ALT, alternative allele

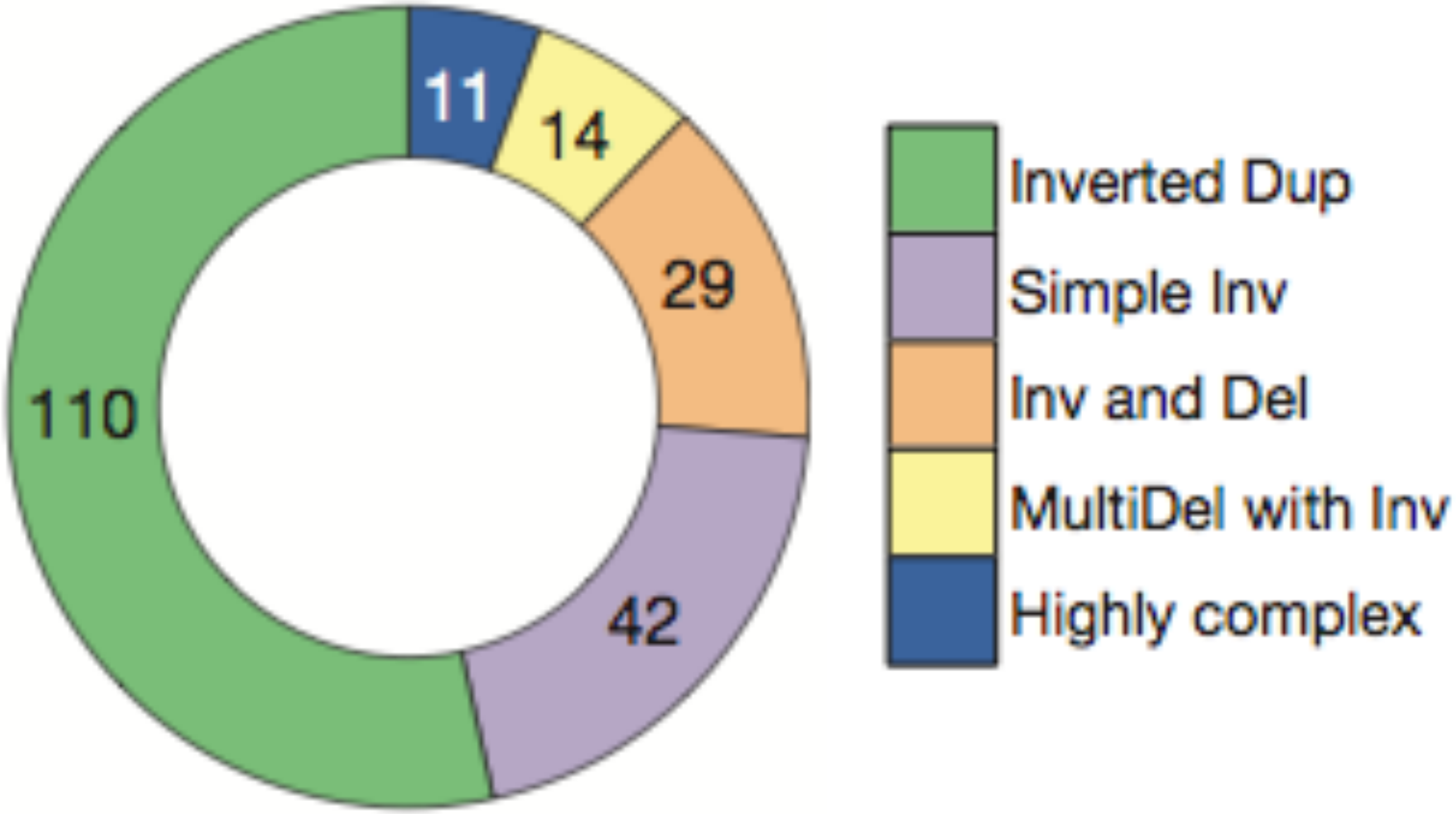
- Split-read smaller-scale complex deletions (7,804 examined):
- 664 small deletions exhibit complexity (median size 67bp)
- 64 (of the 664) contained insertions >3bp that may be derived from a nearby template.

Sudmant et al.Nature 526,75–81 (2015).

1000 Genomes Project

Structural Variants

Summary of Inversion Complexity



Sudmant et al.Nature 526,75–81 (2015).

Sudmant et al.Nature 526,75–81 (2015).

The UK10K Project

The UK10K Project

Aims

Genome-wide sequencing of deeply phenotyped cohorts,
Exome (protein-coding regions) analysis of selected extreme phenotypes to:

1. **Elucidate singleton variants by maximising variation detected.**
   - Pre-existing cohorts of related phenotypes.
   - Genome-wide sequencing of 4,000 samples from the **TwinsUK** and **ALSPAC** cohorts to 6x sequencing depth. (**ALSPAC,** Avon Longitudinal Study of Parents and Children)
2. **Directly associate genetic variations to phenotypic traits**
   TwinsUK and ALSPAC cohorts have been deeply phenotyped
   Analysis of shared genetic variation within twin pairs - link to disease.
3. **Uncover rare variants contributing to disease**
   - 6,000 exomes of extreme phenotypes of specific conditions
   - identified obesity and neurodevelopmental disorder cohorts
   - 8 other areas
4. **Assign uncovered variations into genotyped cohort and case/control collections**
5. **Provide a sequence variation resource for future studies**

uk10k.org

The UK10K Project

Information about the UK10K Study Samples:

- Whole genome cohorts (4000)
- Neurodevelopment Sample Sets (up to 3000 whole exomes)
- Obesity Sample Sets (2000 whole exomes)
- Rare Diseases Sample Sets (1000 whole exomes)

uk10k.org

The UK10K Project

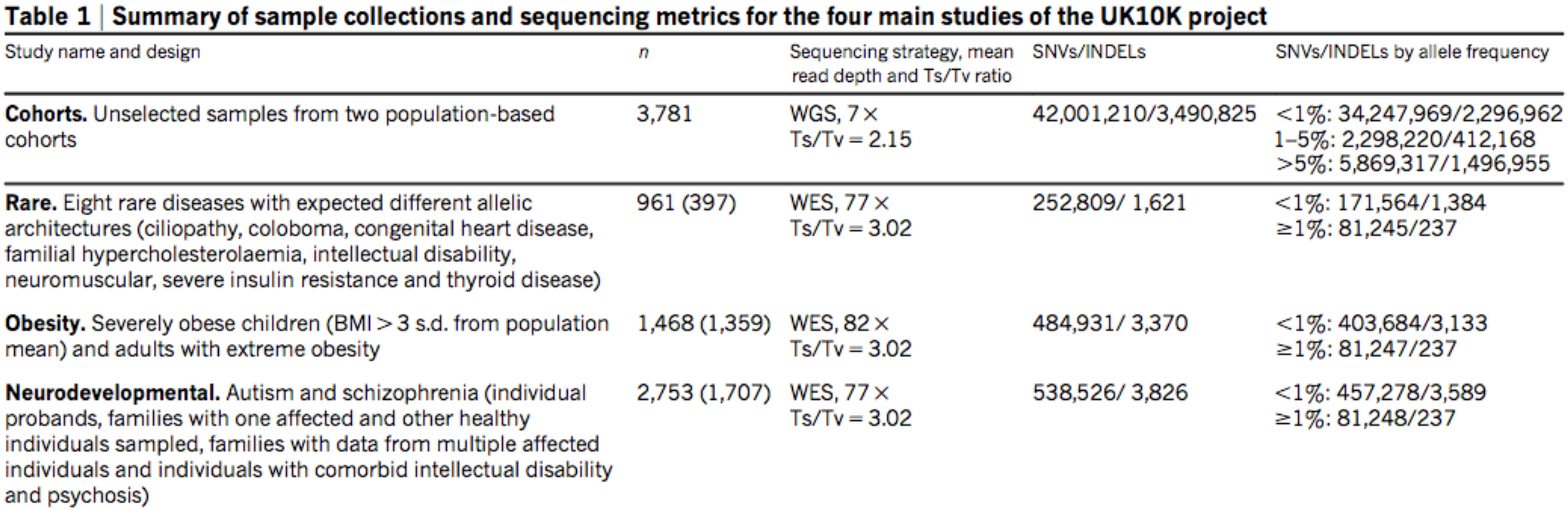


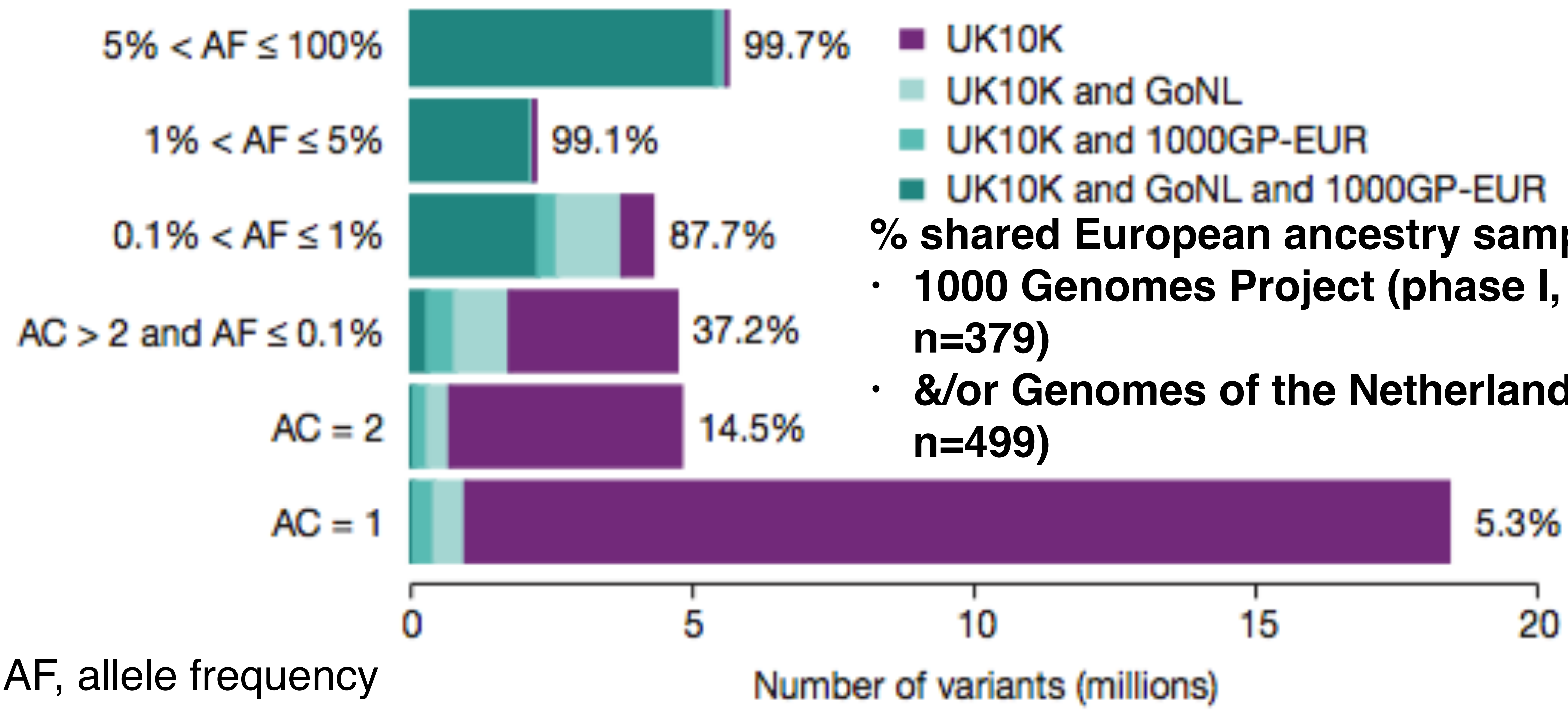
Table 1 | Summary of sample collections and sequencing metrics for the four main studies of the UK10K project

Study name and design	*n*	Sequencing strategy, mean read depth and Ts/Tv ratio	SNVs/INDELs	SNVs/INDELs by allele frequency
Cohorts. Unselected samples from two population-based cohorts	3,781	WGS, 7 × Ts/Tv = 2.15	42,001,210/3,490,825	<1%: 34,247,969/2,296,962 1–5%: 2,298,220/412,168 >5%: 5,869,317/1,496,955
Rare. Eight rare diseases with expected different allelic architectures (ciliopathy, coloboma, congenital heart disease, familial hypercholesterolaemia, intellectual disability, neuromuscular, severe insulin resistance and thyroid disease)	961 (397)	WES, 77 × Ts/Tv = 3.02	252,809/ 1,621	<1%: 171,564/1,384 ≥1%: 81,245/237
Obesity. Severely obese children (BMI > 3 s.d. from population mean) and adults with extreme obesity	1,468 (1,359)	WES, 82 × Ts/Tv = 3.02	484,931/ 3,370	<1%: 403,684/3,133 ≥1%: 81,247/237
Neurodevelopmental. Autism and schizophrenia (individual probands, families with one affected and other healthy individuals sampled, families with data from multiple affected individuals and individuals with comorbid intellectual disability and psychosis)	2,753 (1,707)	WES, 77 × Ts/Tv = 3.02	538,526/ 3,826	<1%: 457,278/3,589 ≥1%: 81,248/237

The UK10K Consortium, Nature 526, 82–90 (2015)

The UK10K Project

SNVs in all autosomal regions (Allele Frequency bins)



% shared European ancestry samples from

- **1000 Genomes Project (phase I, EUR n=379)**
- **&/or Genomes of the Netherlands (GoNL, n=499)**

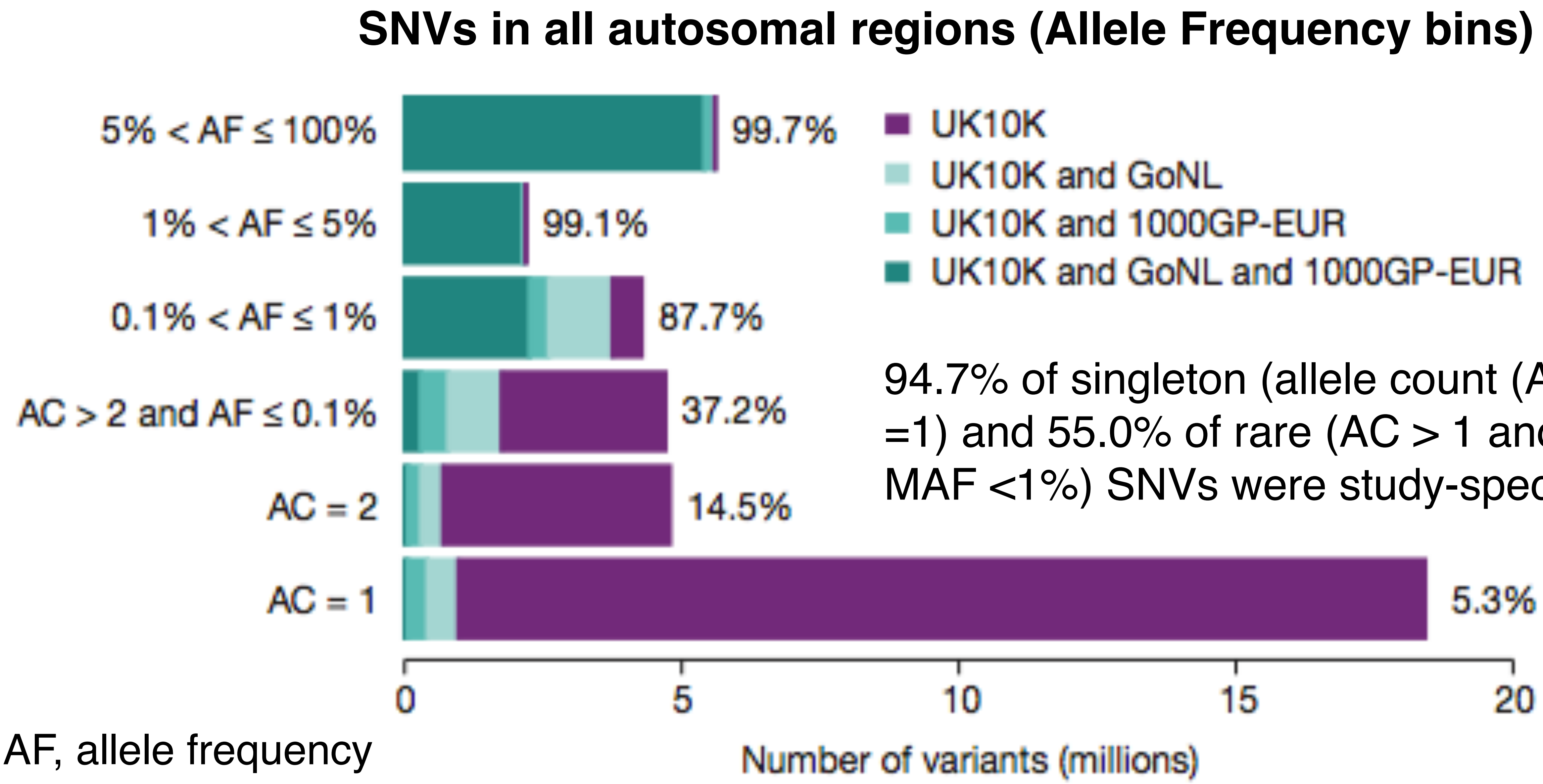
AF, allele frequency

AC,allele count

MAF, minor allele frequency

The UK10K Consortium, Nature 526, 82–90 (2015)

The UK10K Project



94.7% of singleton (allele count (AC) =1) and 55.0% of rare (AC > 1 and MAF <1%) SNVs were study-specific.

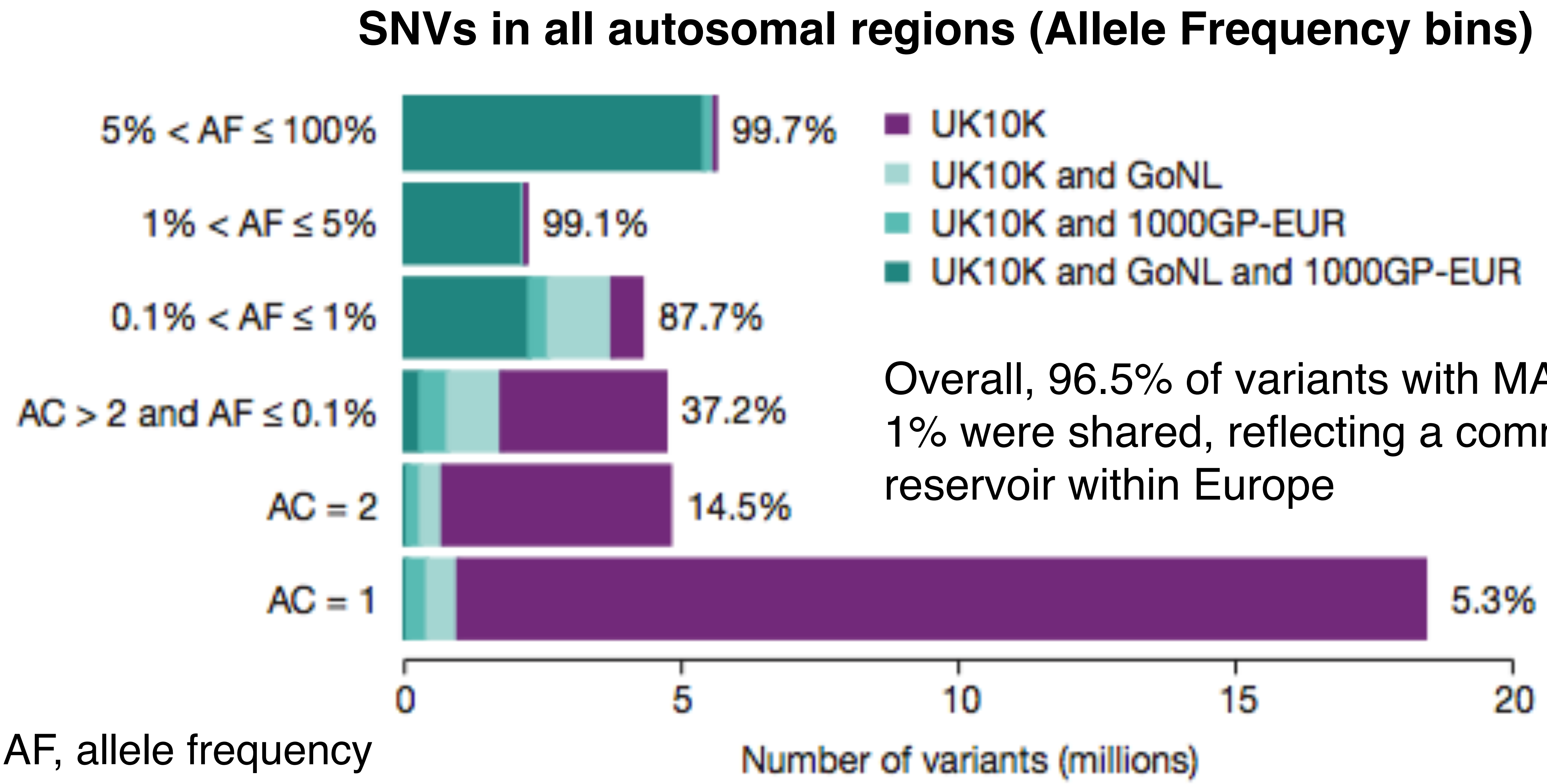
AF, allele frequency

AC,allele count

MAF, minor allele frequency

The UK10K Consortium, Nature 526, 82–90 (2015)

The UK10K Project



Overall, 96.5% of variants with MAF > 1% were shared, reflecting a common reservoir within Europe

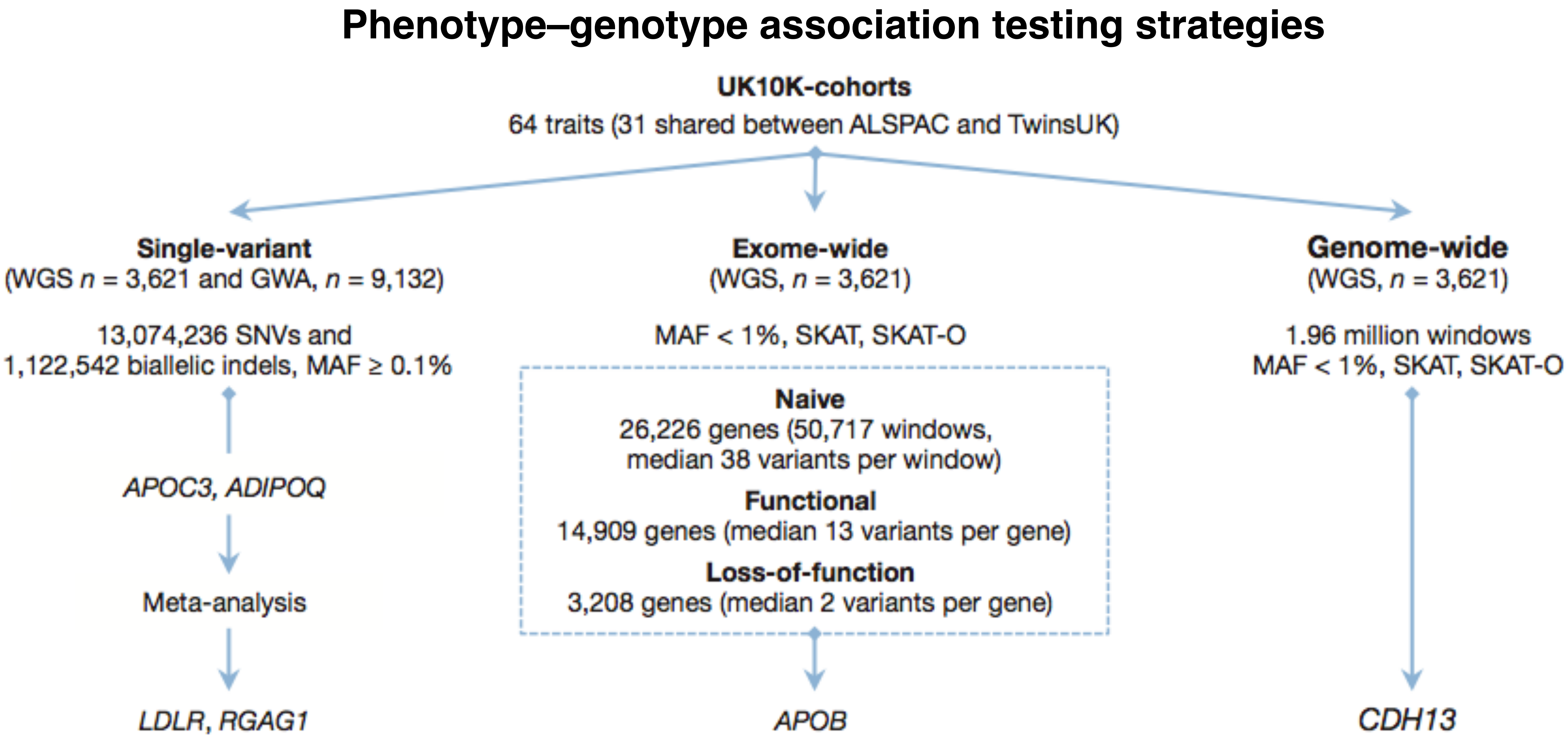
AF, allele frequency

AC,allele count

MAF, minor allele frequency

The UK10K Consortium, Nature 526, 82–90 (2015)

The UK10K Project



Phenotype–genotype association testing strategies

UK10K-cohorts
64 traits (31 shared between ALSPAC and TwinsUK)

Single-variant
(WGS $n = 3{,}621$ and GWA, $n = 9{,}132$)

13,074,236 SNVs and
1,122,542 biallelic indels, MAF ≥ 0.1%

APOC3, ADIPOQ

Meta-analysis

LDLR, RGAG1

Exome-wide
(WGS, $n = 3{,}621$)

MAF < 1%, SKAT, SKAT-O

Naive
26,226 genes (50,717 windows,
median 38 variants per window)

Functional
14,909 genes (median 13 variants per gene)

Loss-of-function
3,208 genes (median 2 variants per gene)

APOB

Genome-wide
(WGS, $n = 3{,}621$)

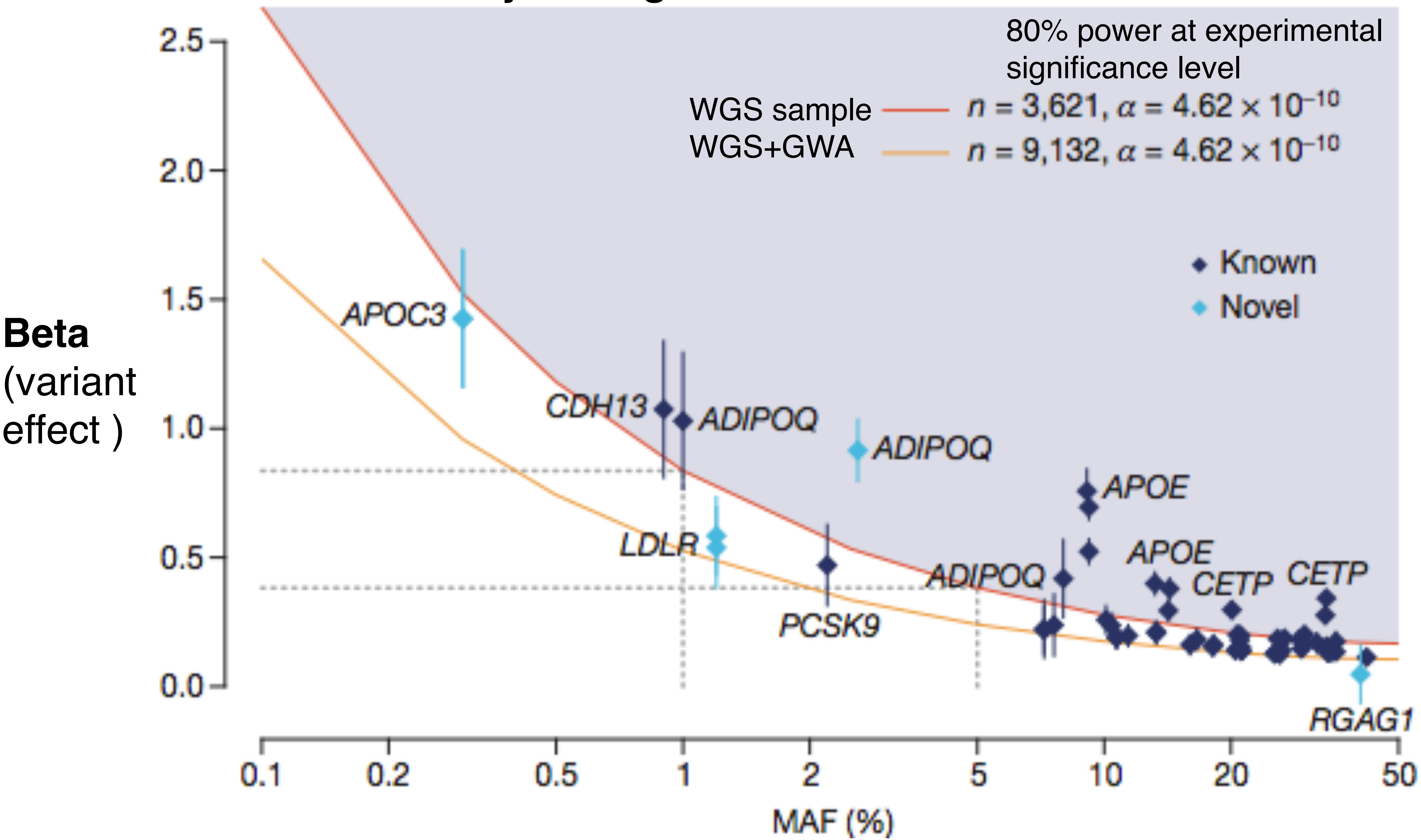
1.96 million windows
MAF < 1%, SKAT, SKAT-O

CDH13

The UK10K Consortium, Nature 526, 82–90 (2015)

The UK10K Project

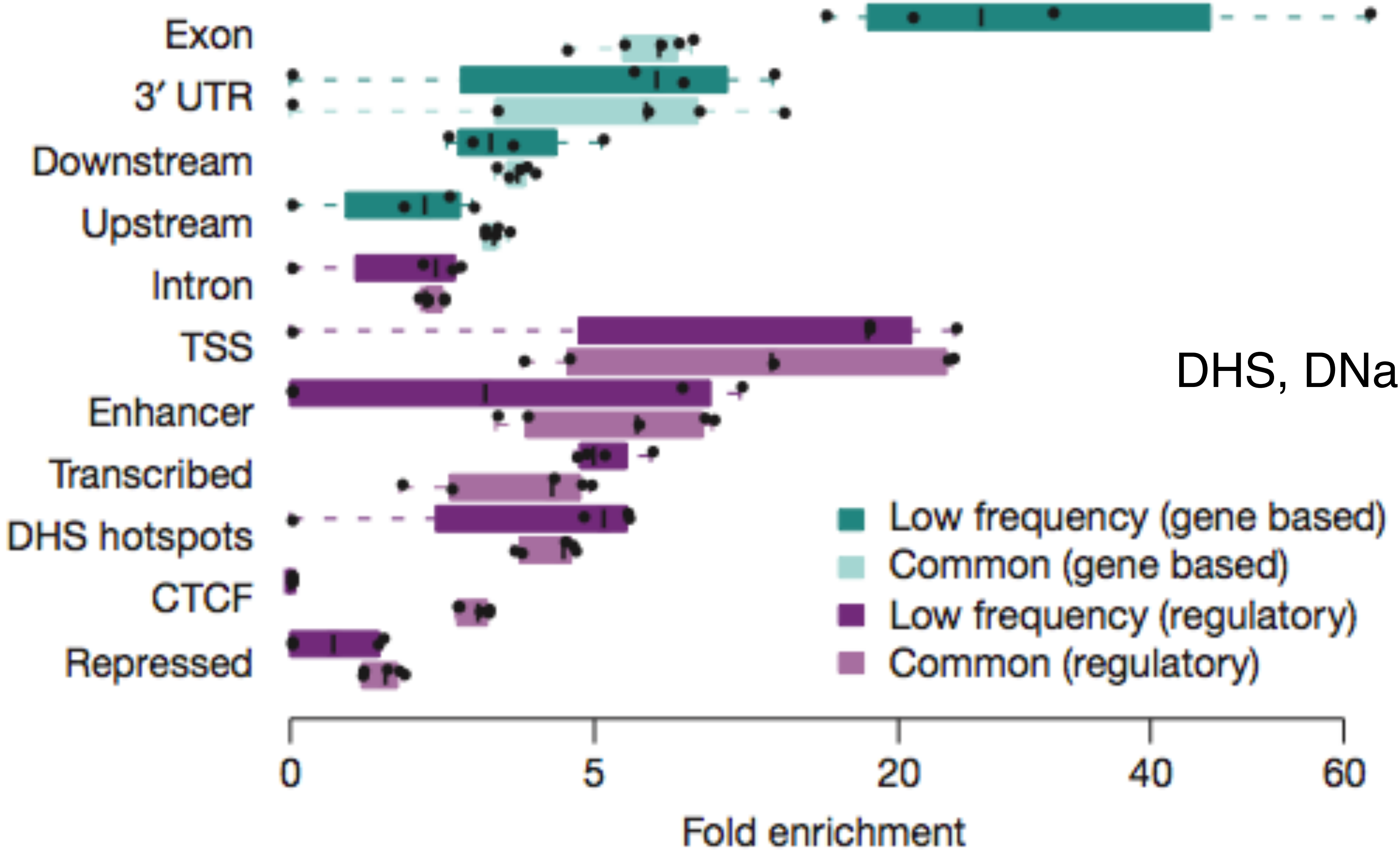
Summary of single-marker association results



The UK10K Consortium, Nature 526, 82–90 (2015)

The UK10K Project

Enrichment of single-marker association functional annotation



DHS, DNase I hotspots

Fold enrichment estimated across five (of 31 core) traits

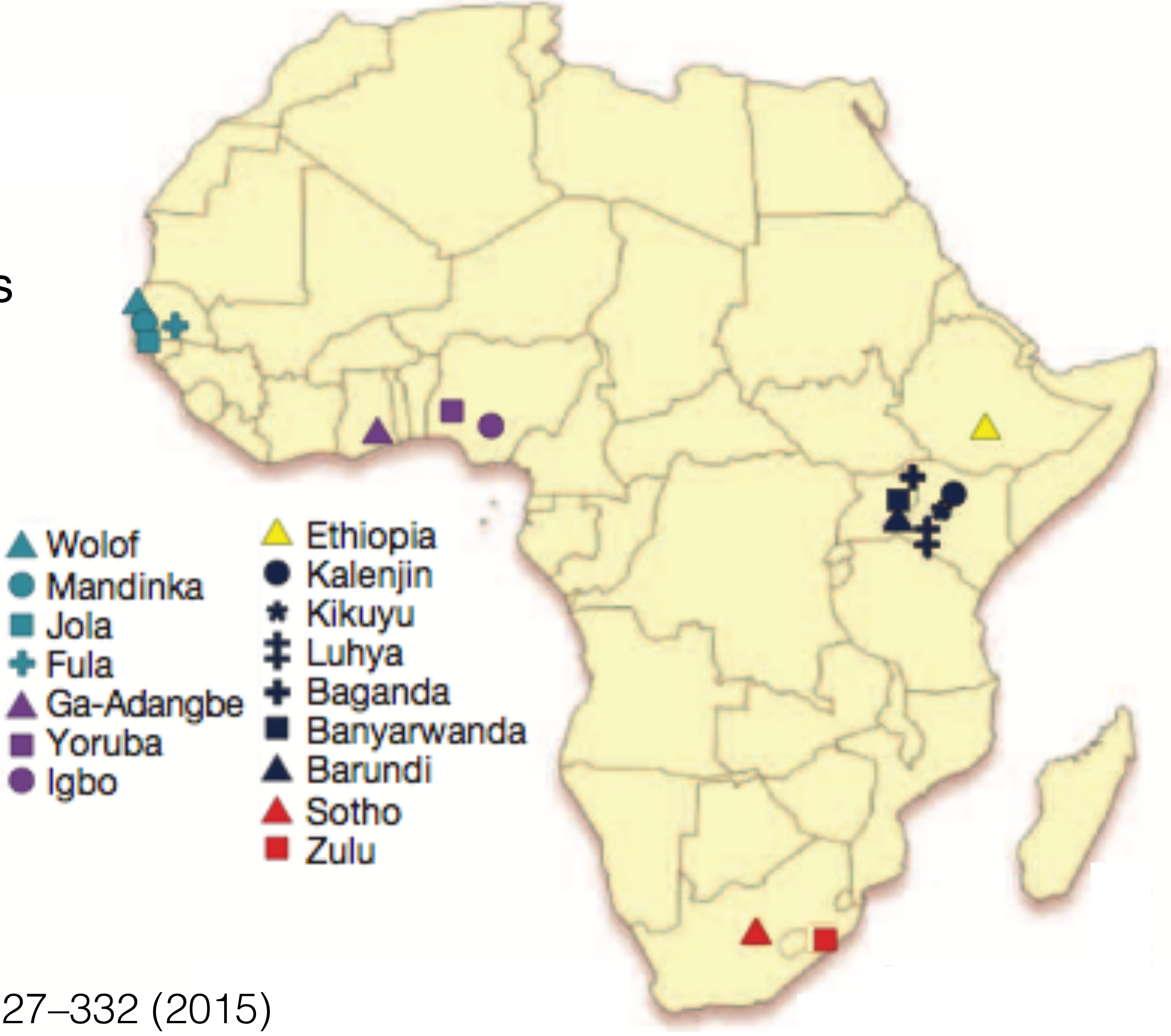
min 10 independent SNVs associated with the trait at 10^{-7} P-value (permutation test) (HDL, LDL, TC, APOA1 and APOB).

The UK10K Consortium, Nature 526, 82–90 (2015)

African Genome Variation Project

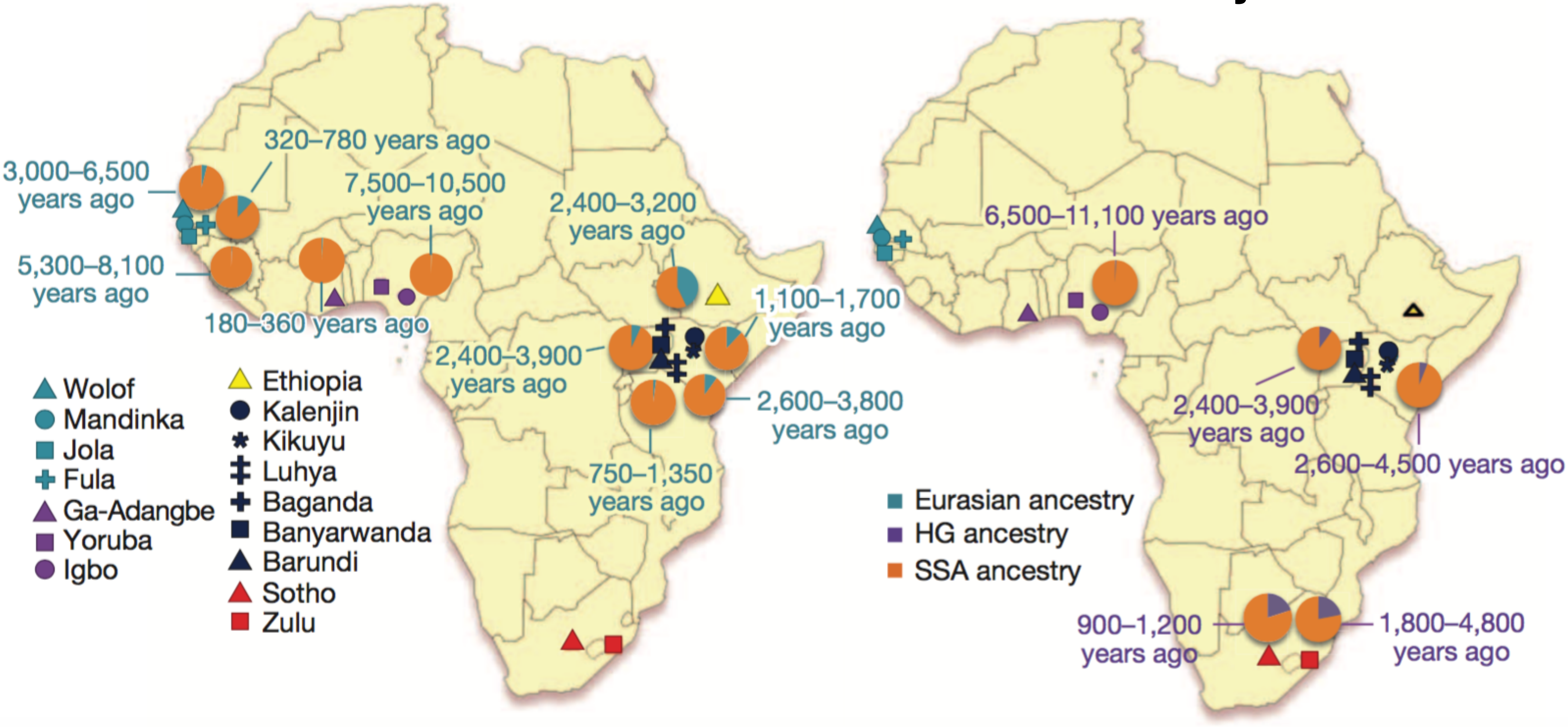
African Genome Variation Project

- Dense genotypes from 1,481 individuals
- Whole-genome sequences from 320 individuals
- 18 African populations (2 populations from 1000 Genomes Project)



Gurdasani et al.,Nature 517, 327–332 (2015)

African Genome Variation Project



Dating and proportion of Eurasian HG admixture among African populations.

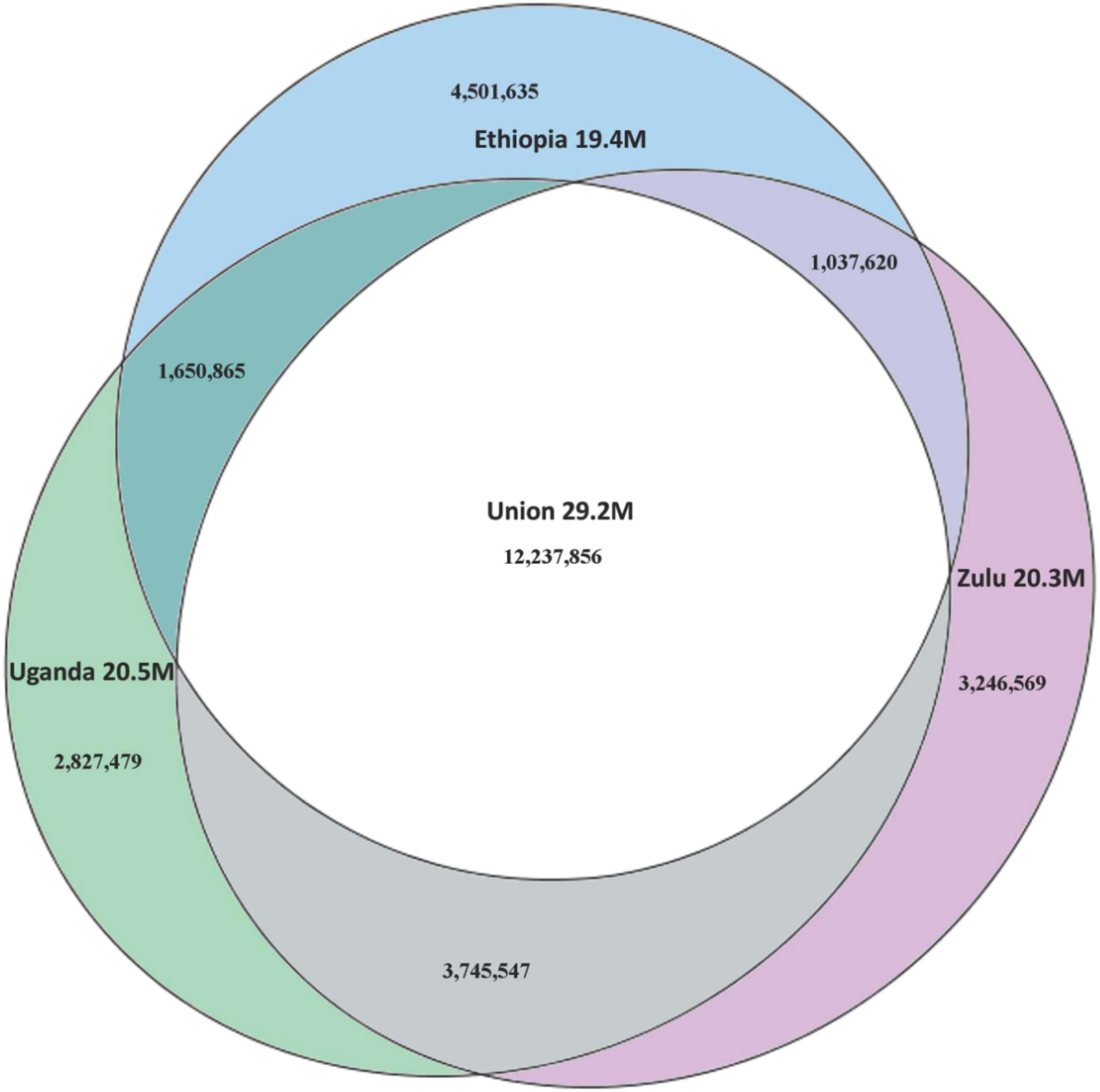
HG, Hunter Gatherer

SSA, sub-Saharan Africa

Gurdasani et al.,Nature 517, 327–332 (2015)

African Genome Variation Project

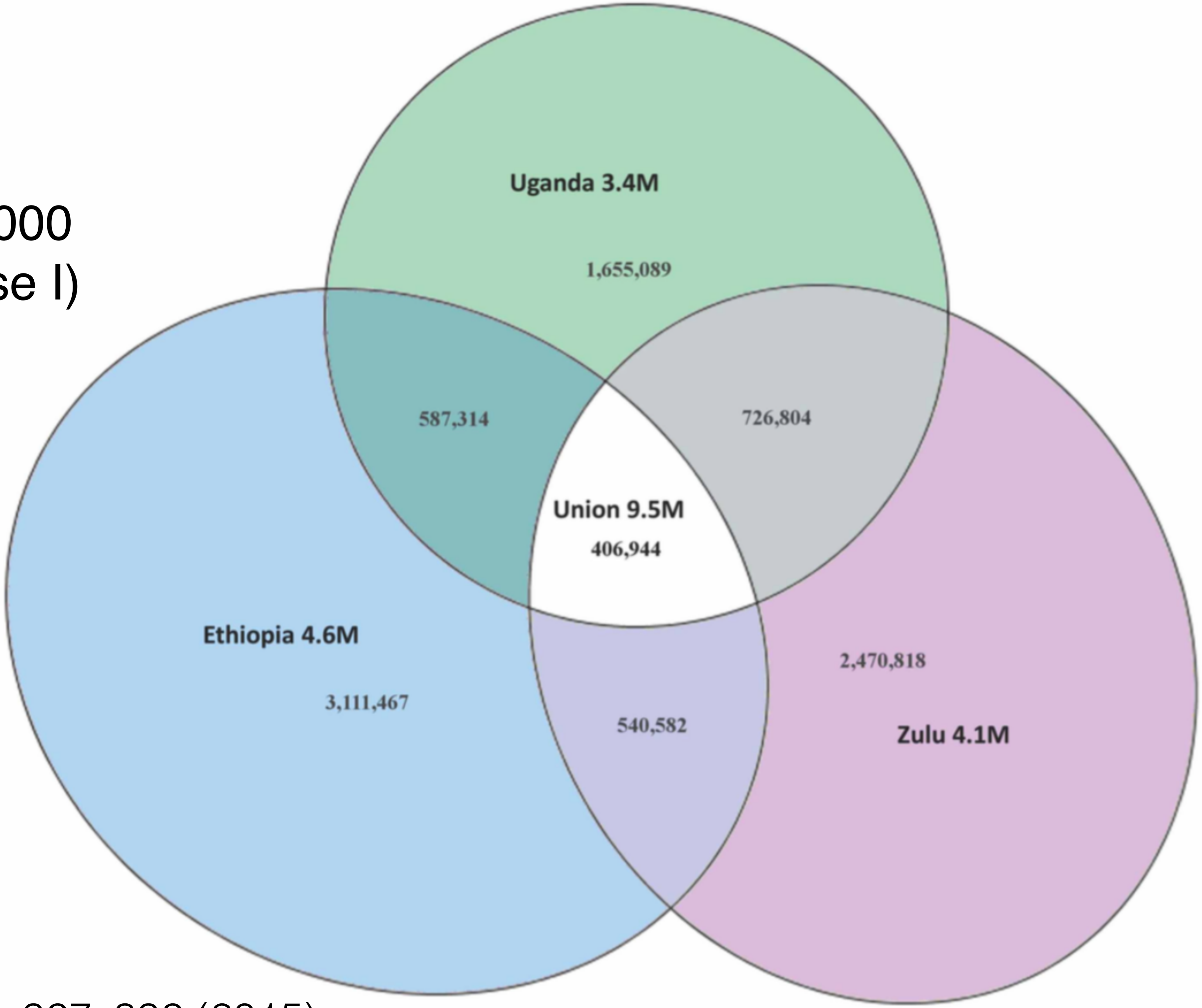
- 29.8 Million SNPs
- 4xWGS data from Zulu, Ugandan and Ethiopian individuals (subsampled to 100 samples each).
- 10-23% unshared (private variants) of the total number of variants in a given population.



Gurdasani et al.,Nature 517, 327–332

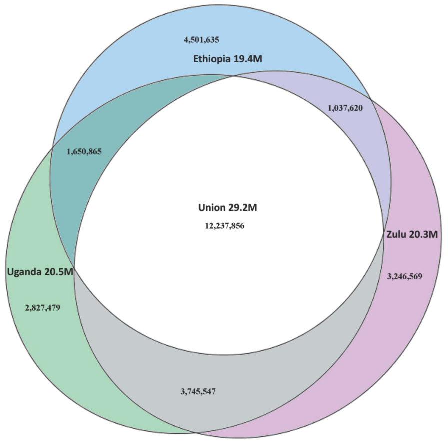
African Genome Variation Project

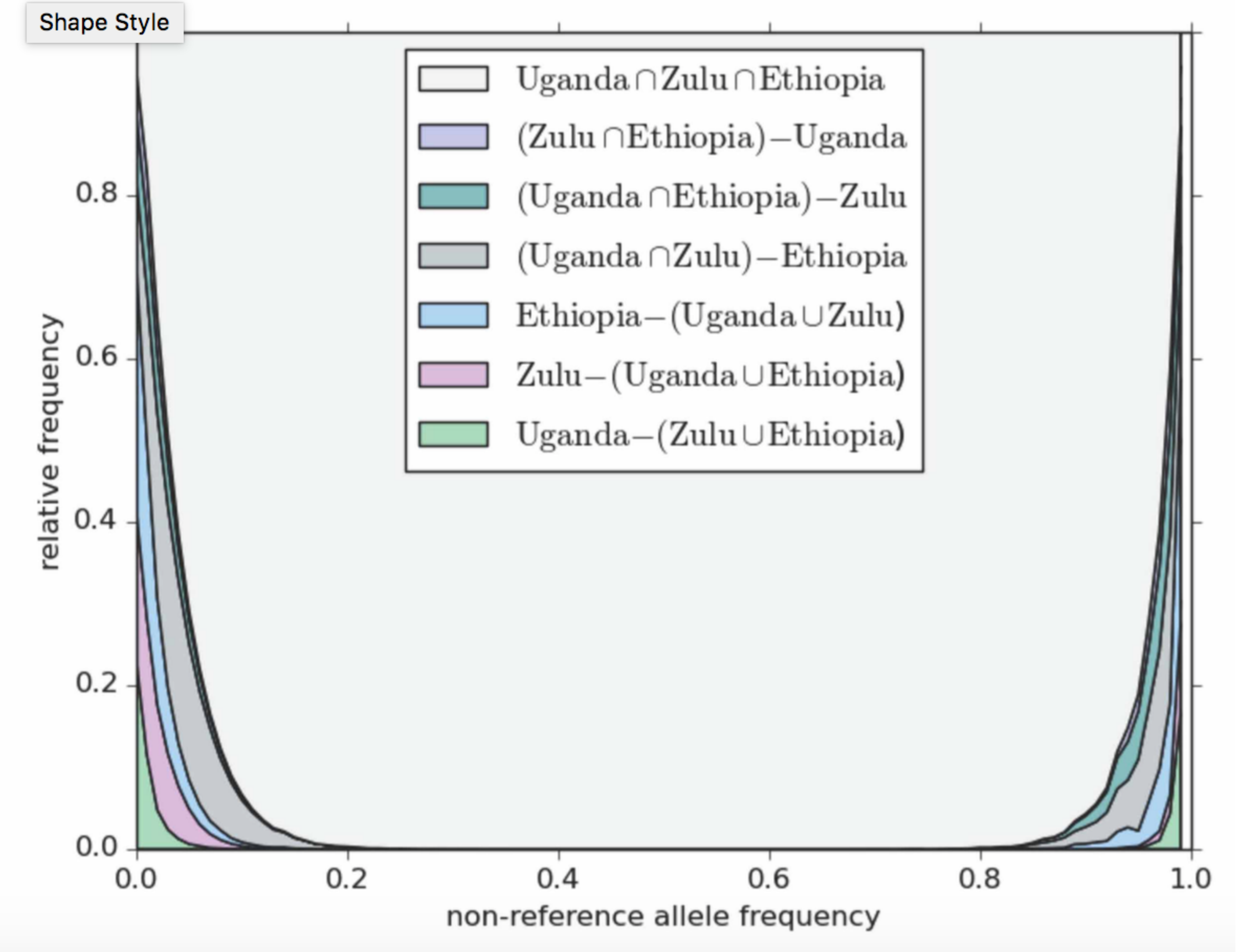
- novel variants (not in1000 Genomes Project phase I)
- Ethiopia has highest



Gurdasani et al.,Nature 517, 327–332 (2015)

African Genome Variation Project

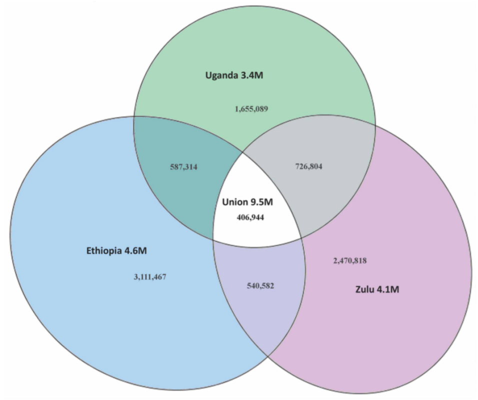




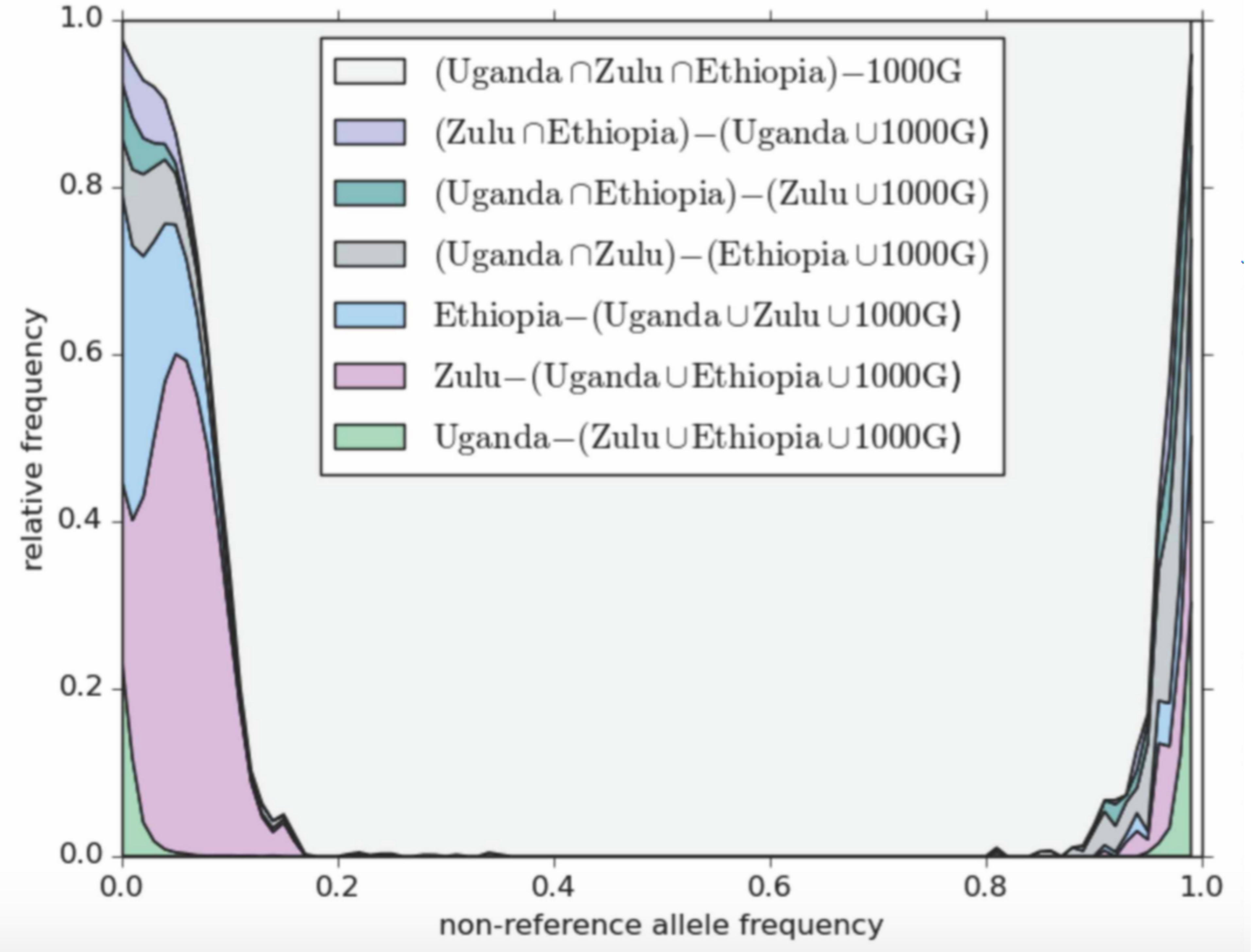
- relative allele frequencies

Gurdasani et al.,Nature 517, 327–332 (2015)

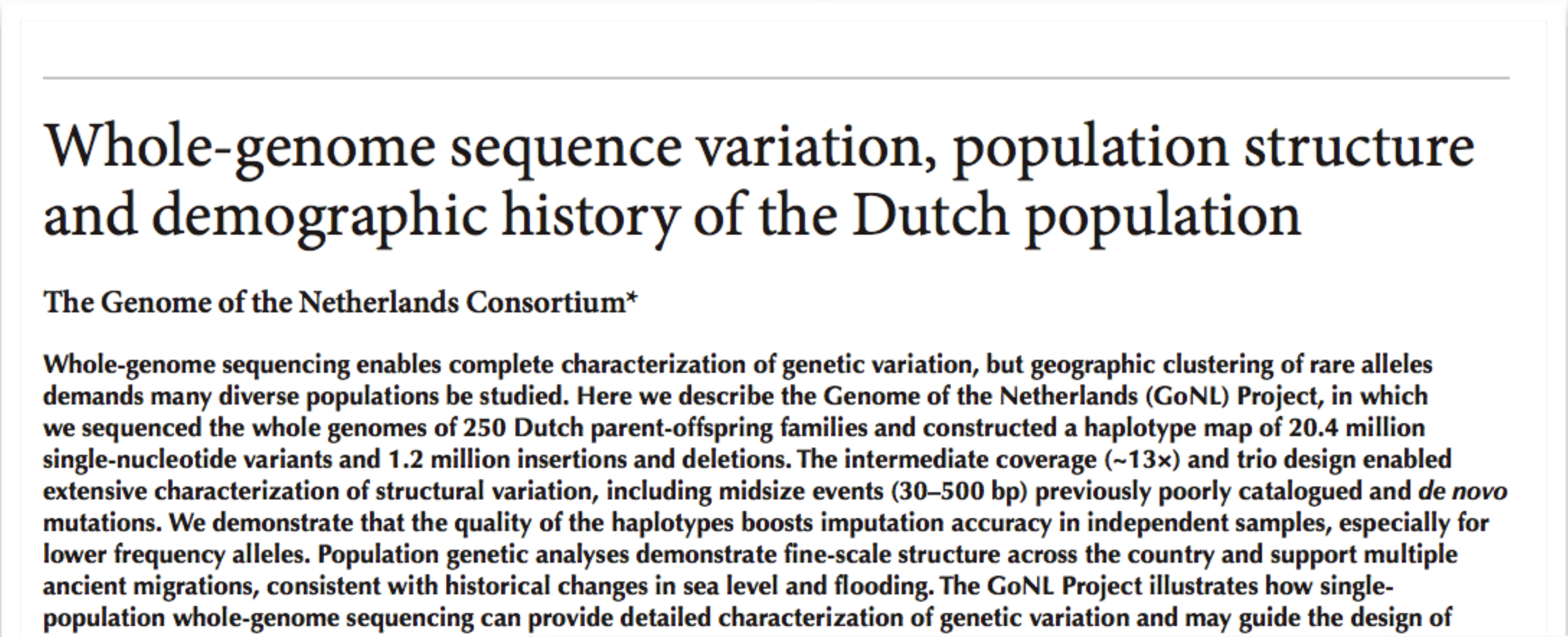
African Genome Variation Project



- relative allele frequencies



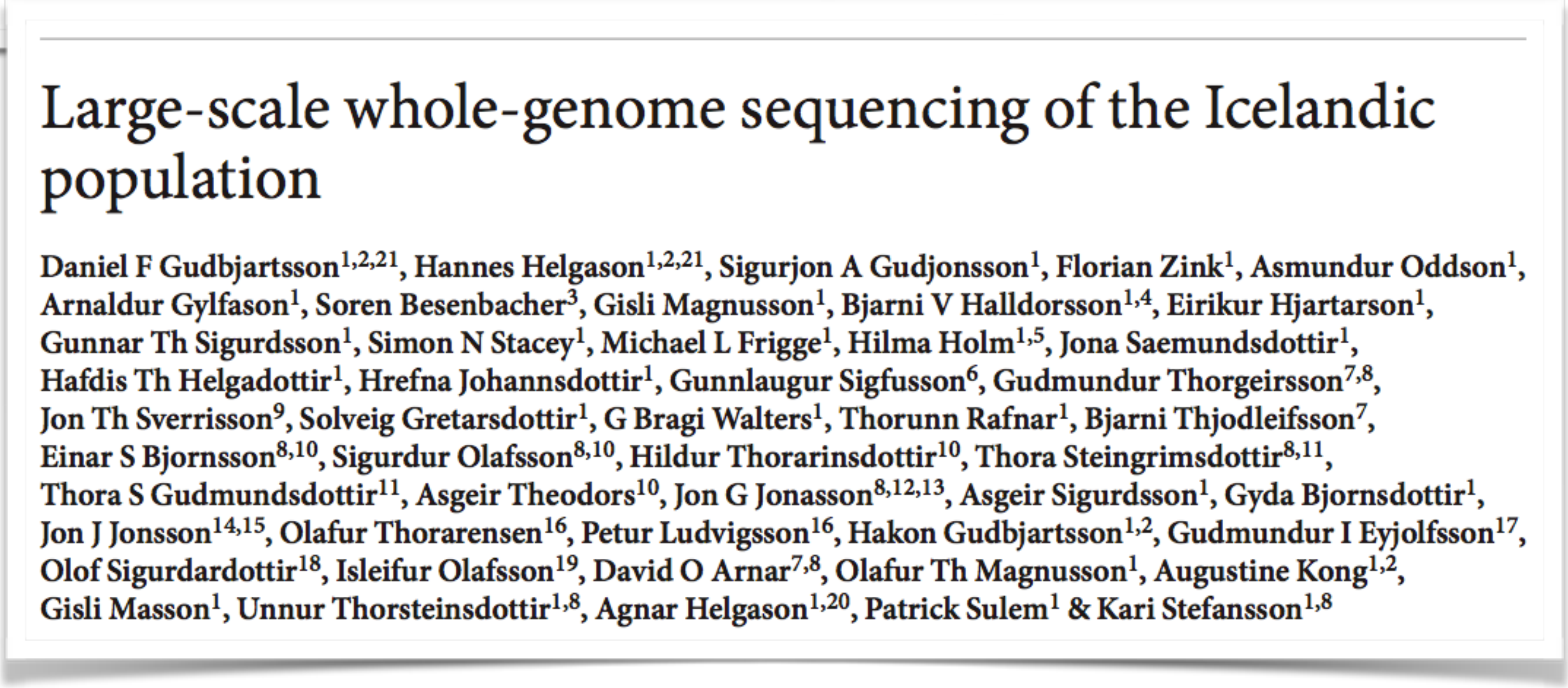
Gurdasani et al.,Nature 517, 327–332 (2015)



Whole-genome sequence variation, population structure and demographic history of the Dutch population

The Genome of the Netherlands Consortium*

Whole-genome sequencing enables complete characterization of genetic variation, but geographic clustering of rare alleles demands many diverse populations be studied. Here we describe the Genome of the Netherlands (GoNL) Project, in which we sequenced the whole genomes of 250 Dutch parent-offspring families and constructed a haplotype map of 20.4 million single-nucleotide variants and 1.2 million insertions and deletions. The intermediate coverage (~13×) and trio design enabled extensive characterization of structural variation, including midsize events (30–500 bp) previously poorly catalogued and *de novo* mutations. We demonstrate that the quality of the haplotypes boosts imputation accuracy in independent samples, especially for lower frequency alleles. Population genetic analyses demonstrate fine-scale structure across the country and support multiple ancient migrations, consistent with historical changes in sea level and flooding. The GoNL Project illustrates how single-population whole-genome sequencing can provide detailed characterization of genetic variation and may guide the design of future population studies.



Large-scale whole-genome sequencing of the Icelandic population

Daniel F Gudbjartsson[1,2,21], Hannes Helgason[1,2,21], Sigurjon A Gudjonsson[1], Florian Zink[1], Asmundur Oddson[1], Arnaldur Gylfason[1], Soren Besenbacher[3], Gisli Magnusson[1], Bjarni V Halldorsson[1,4], Eirikur Hjartarson[1], Gunnar Th Sigurdsson[1], Simon N Stacey[1], Michael L Frigge[1], Hilma Holm[1,5], Jona Saemundsdottir[1], Hafdis Th Helgadottir[1], Hrefna Johannsdottir[1], Gunnlaugur Sigfusson[6], Gudmundur Thorgeirsson[7,8], Jon Th Sverrisson[9], Solveig Gretarsdottir[1], G Bragi Walters[1], Thorunn Rafnar[1], Bjarni Thjodleifsson[7], Einar S Bjornsson[8,10], Sigurdur Olafsson[8,10], Hildur Thorarinsdottir[10], Thora Steingrimsdottir[8,11], Thora S Gudmundsdottir[11], Asgeir Theodors[10], Jon G Jonasson[8,12,13], Asgeir Sigurdsson[1], Gyda Bjornsdottir[1], Jon J Jonsson[14,15], Olafur Thorarensen[16], Petur Ludvigsson[16], Hakon Gudbjartsson[1,2], Gudmundur I Eyjolfsson[17], Olof Sigurdardottir[18], Isleifur Olafsson[19], David O Arnar[7,8], Olafur Th Magnusson[1], Augustine Kong[1,2], Gisli Masson[1], Unnur Thorsteinsdottir[1,8], Agnar Helgason[1,20], Patrick Sulem[1] & Kari Stefansson[1,8]

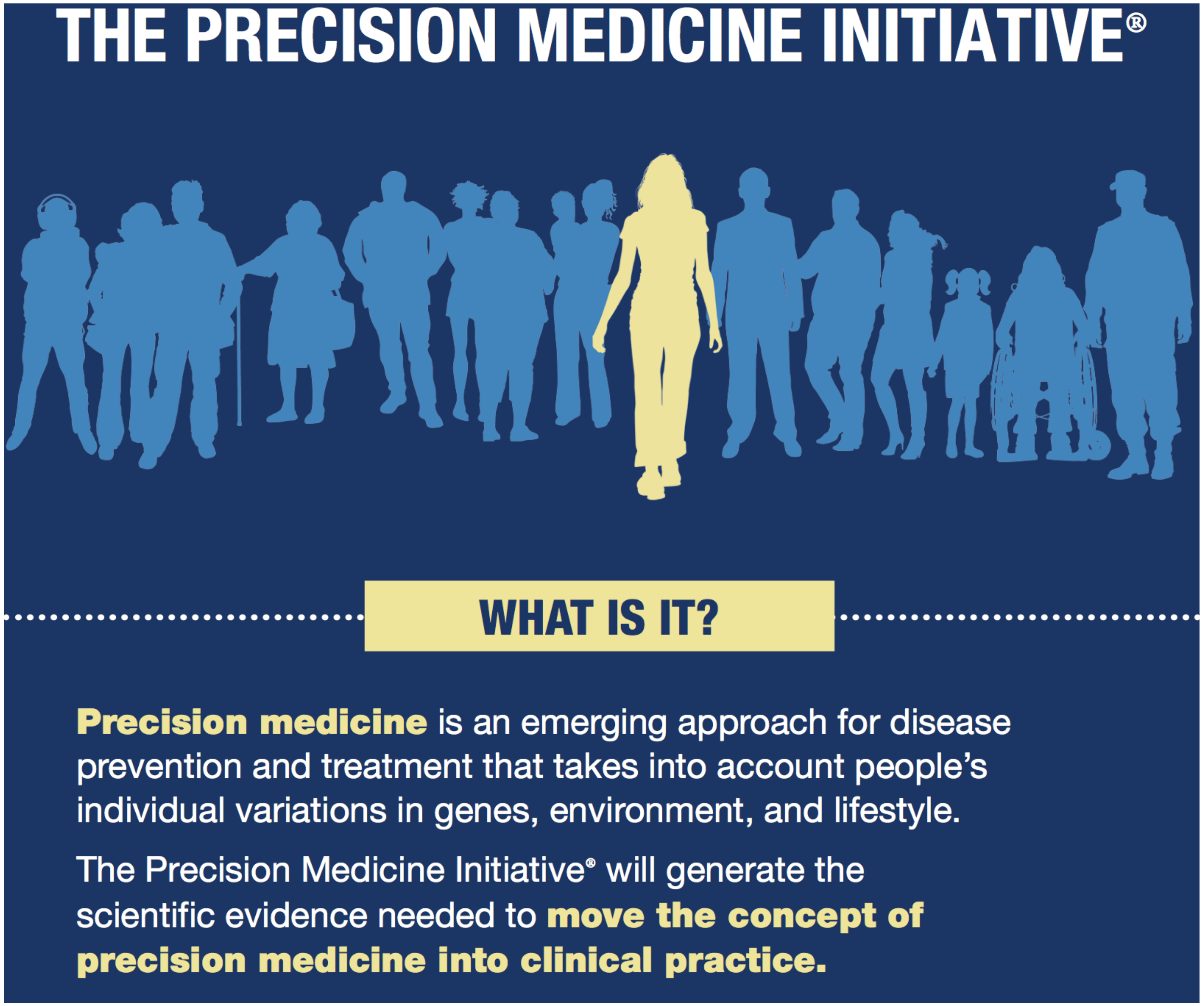
and more on the way…

Francioli et al.,Nature Genetics 46, 818–825 (2014)
Gudbjartsson et al., Nature Genetics 47, 435–444 (2015)

Precision Medicine Initiative

Precision Medicine Initiative

1 million participant research



THE PRECISION MEDICINE INITIATIVE®

WHAT IS IT?

Precision medicine is an emerging approach for disease prevention and treatment that takes into account people's individual variations in genes, environment, and lifestyle.

The Precision Medicine Initiative® will generate the scientific evidence needed to **move the concept of precision medicine into clinical practice.**

http://www.nih.gov/news-events/multimedia-nih-framework-points-way-forward-building-national-large-scale-research-cohort Credit: NIH

Precision Medicine Initiative

- Develop ways to measure risk for a range of diseases based on environmental exposures, genetic factors and interactions between the two;
- Identify the causes of individual differences in response to commonly used drugs (commonly referred to as pharmacogenomics);
- Discover biological markers that signal increased or decreased risk of developing common diseases;
- Use mobile health (mHealth) technologies to correlate activity, physiological measures and environmental exposures with health outcomes;
- Develop new disease classifications and relationships;
- Empower study participants with data and information to improve their own health; and
- Create a platform to enable trials of targeted therapies.

https://www.nih.gov/precision-medicine-initiative-cohort-program/scale-scope

Precision Medicine Initiative

more soon…