The 1000 Genomes Project

Mark DePristo

Co-director of Medical and Population Genetics Program

Broad Institute of MIT and Harvard

Representing and presenting slides from the 1000 Genomes Project Consortium

1000 Genomes project goals

- A public database of essentially all SNPs, indels, and detectable CNVs with allele frequency >1% in each of multiple human population samples

- Pioneer and evaluate methods for:
  - Generating data from next-generation sequencing platforms
  - Exchanging and combining data and analytical methods
  - Discovering and genotyping of SNPs, indels, and CNVs from next-generation DNA sequencing data
  - Imputation with and from next generation sequencing data

David Altshuler

Strategy: use next-generation DNA sequencing to discover common variation

- Collect shotgun DNA sequencing reads using next-generation DNA sequencers
  - Only shallow (4x) coverage per sample
- Map the reads to the reference genome
- Detect variants based on the multiple alignment of reads
  - Statistical analysis across all samples together

5 years ago little of this could be done efficiently, accurately or at scale

The 1000 Genomes Project is a trilogy in four parts

Pilot (2008-2010, published Nature Oct. 2010):

- Deep sequence for two trios (CEU and YRI)
- Low coverage (~2x) of 180 individuals in 3 populations
- Capture of 1000 genes in ~700 individuals

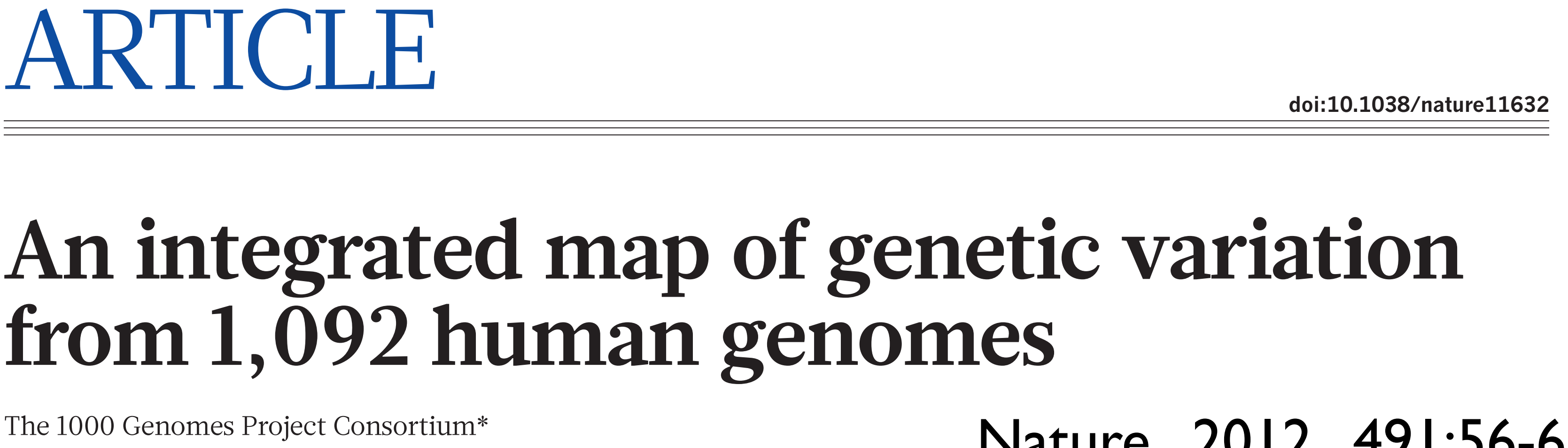
Phase 1 (2010-2012, published Nature Nov. 2012)

- 1100 individuals with ~3x low-coverage, many with exomes
- OMNI 2.5M genotyping
- Paper published last week in Nature

Phase 2+3 (2012-2013, publish final Paper TBD)

- ~2500 samples at >4X coverage, all with exomes and many genotyping arrays
- High coverage Complete Genomics data for 50 samples (with plans for 500)
- Total data size of 25-30 times from original plan, 2.5 more samples in more populations

Phase I complete; paper just published



ARTICLE

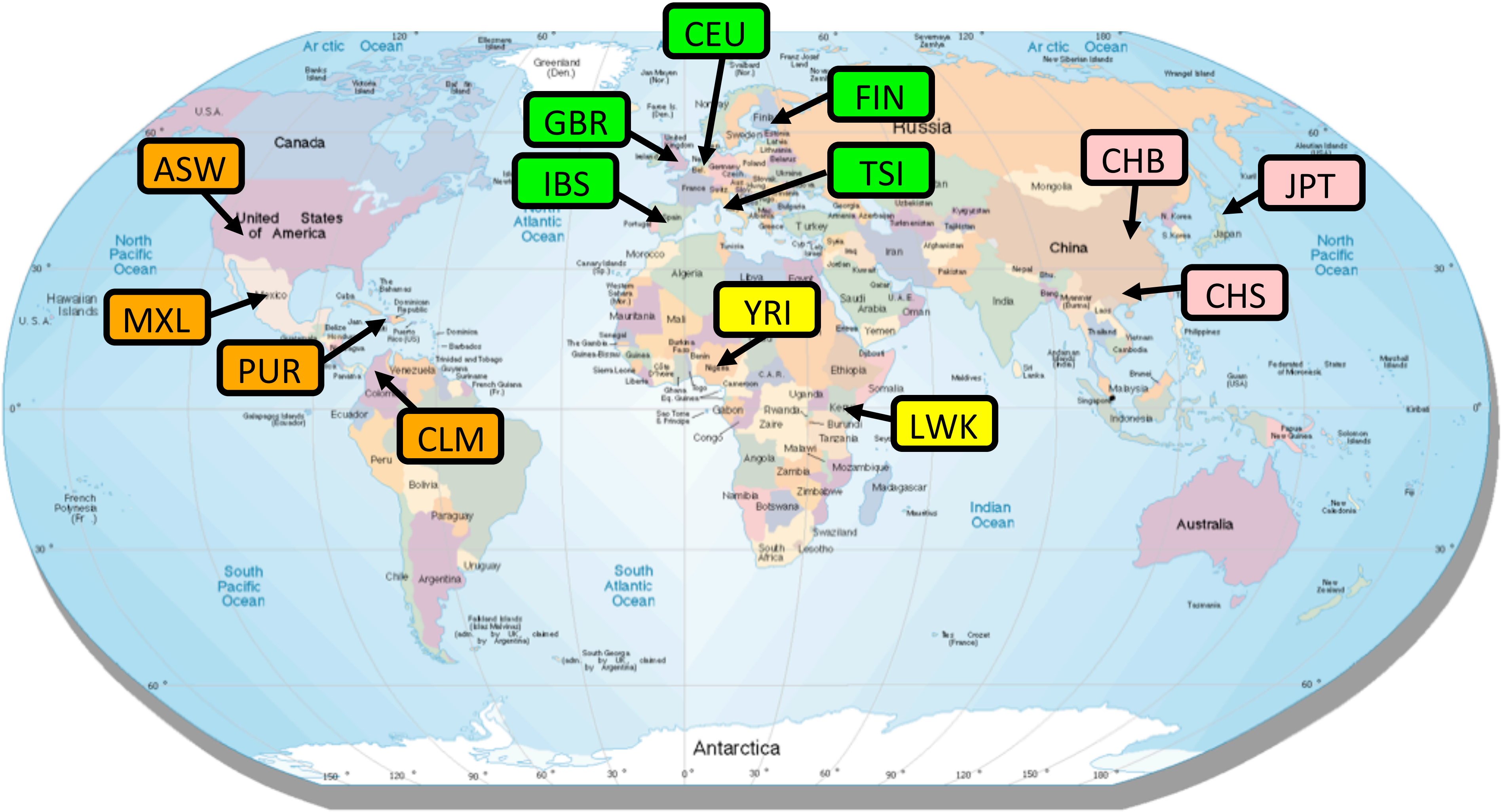
doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

Nature. 2012. 491:56-65

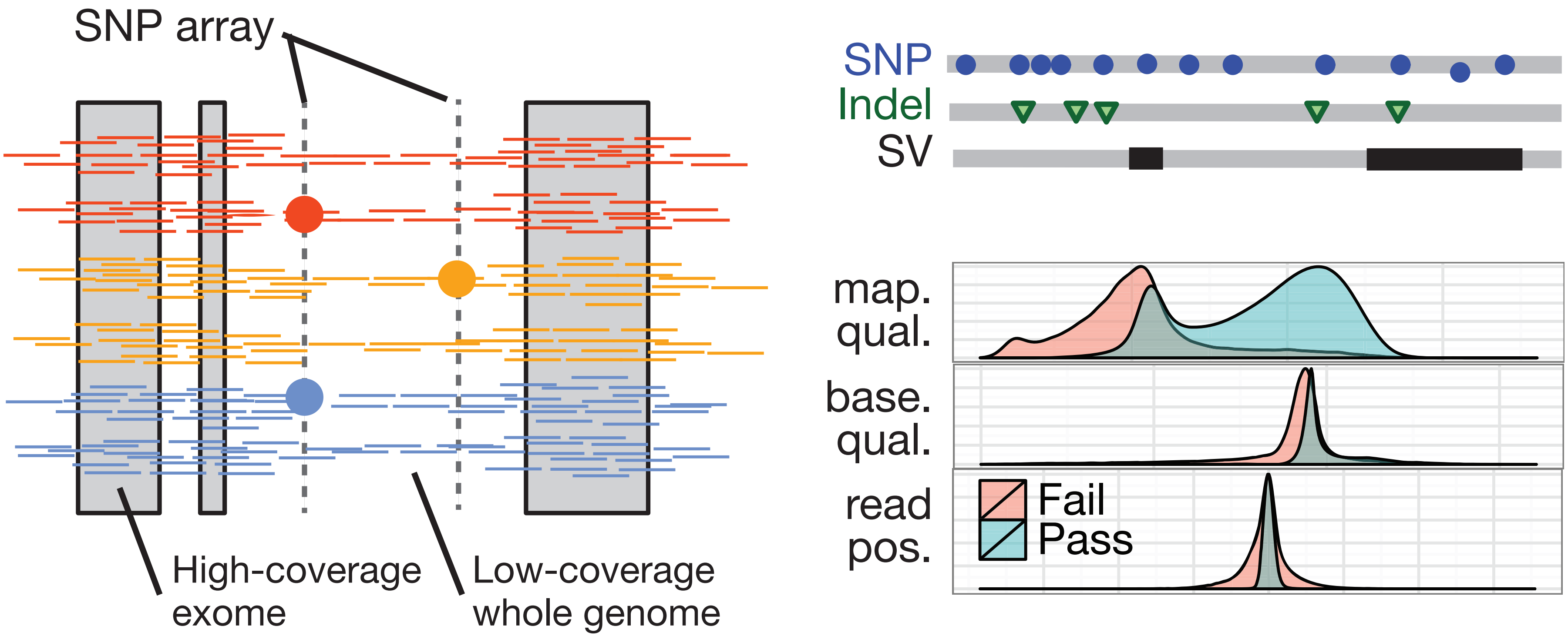
Phase1: 1092 samples from 14 populations



Integrated analysis strategy

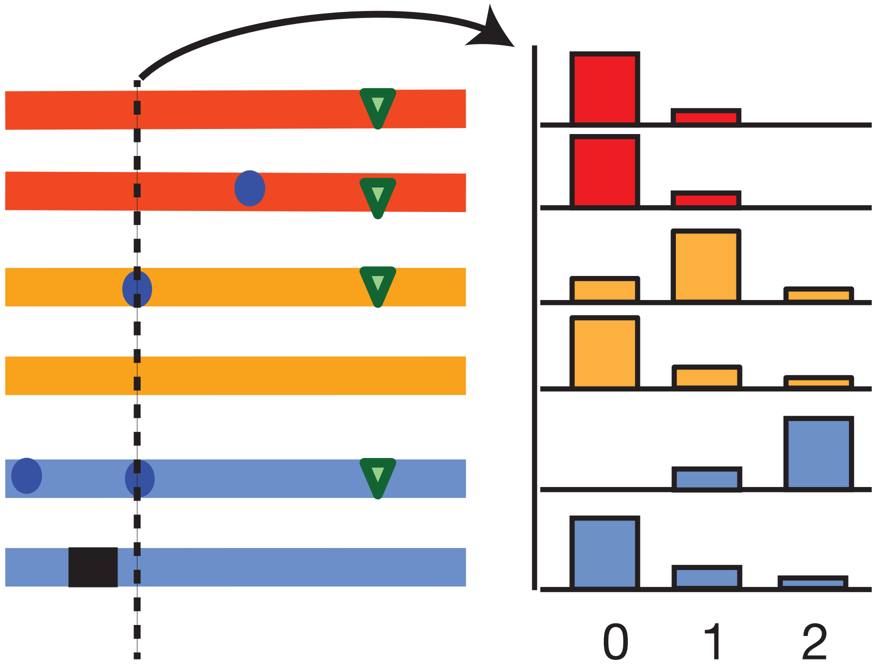
a Primary data
Sequencing, array genotyping

b Candidate variants and quality metrics
Read mapping, quality score recalibration

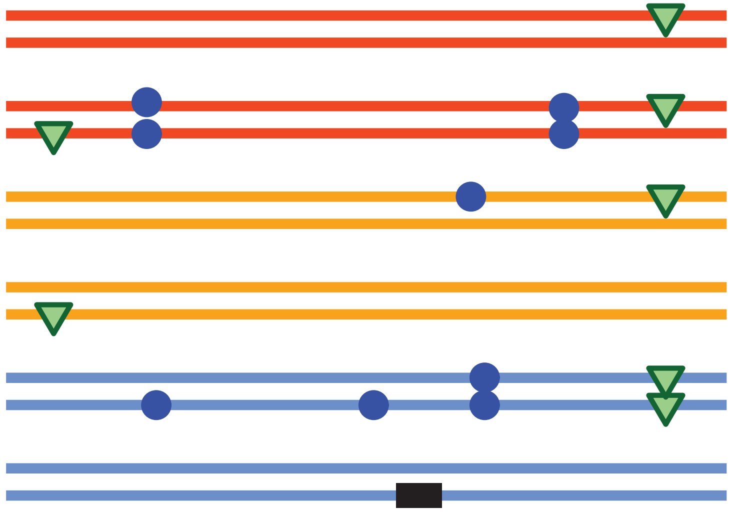


Integrated analysis strategy

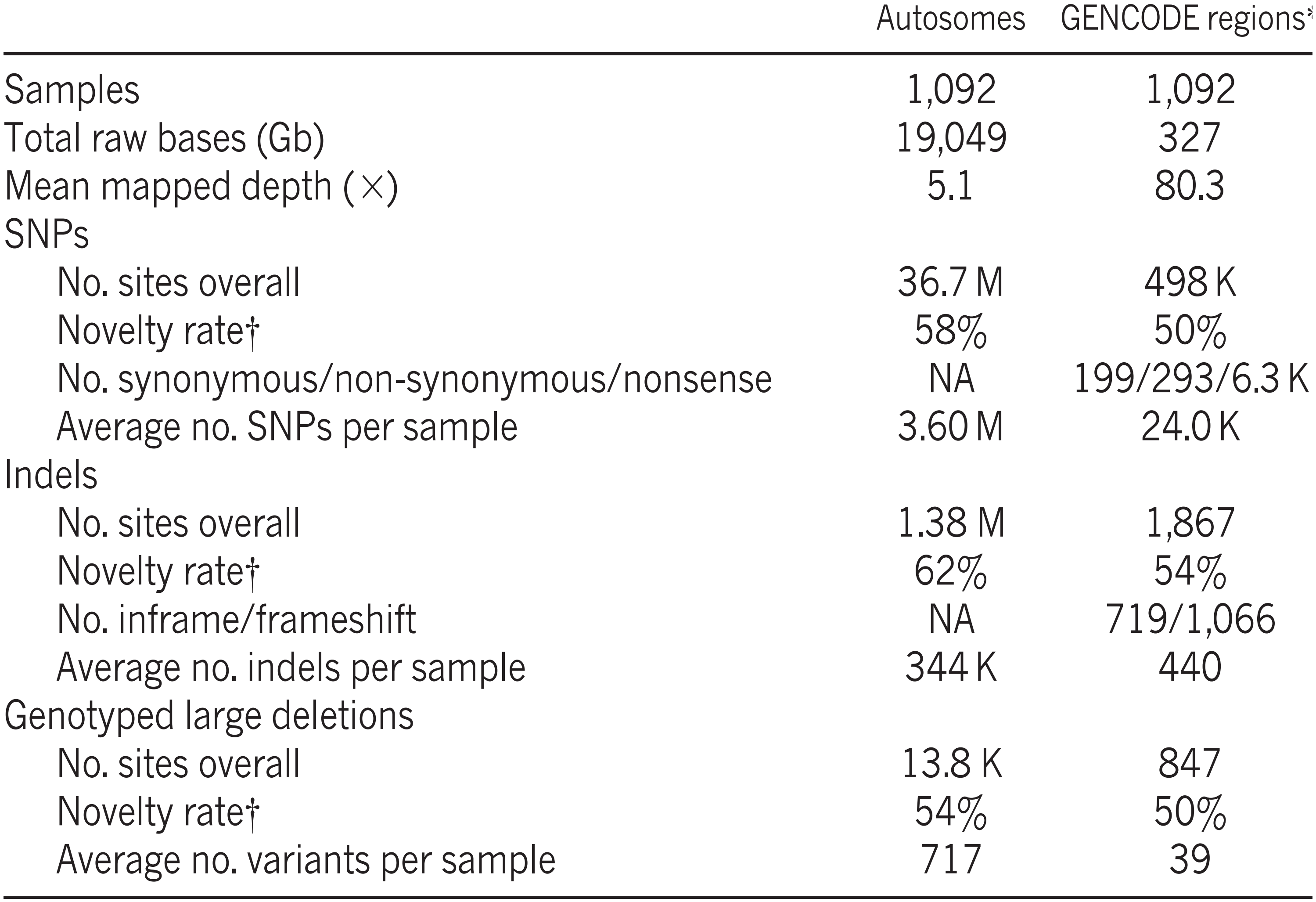
c Variant calls and genotype likelihoods
Variant calling, statistical filtering



d Integrated haplotypes
Probabilistic haplotype estimation

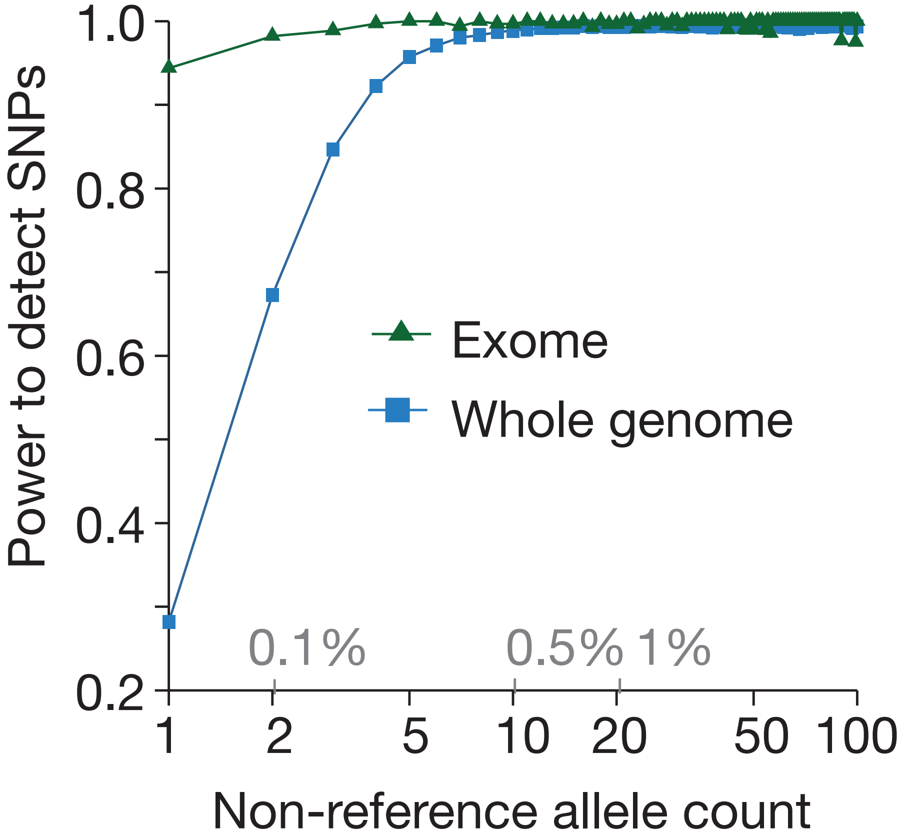


Phase I discovered ~40M SNPs, indels, and structural variants



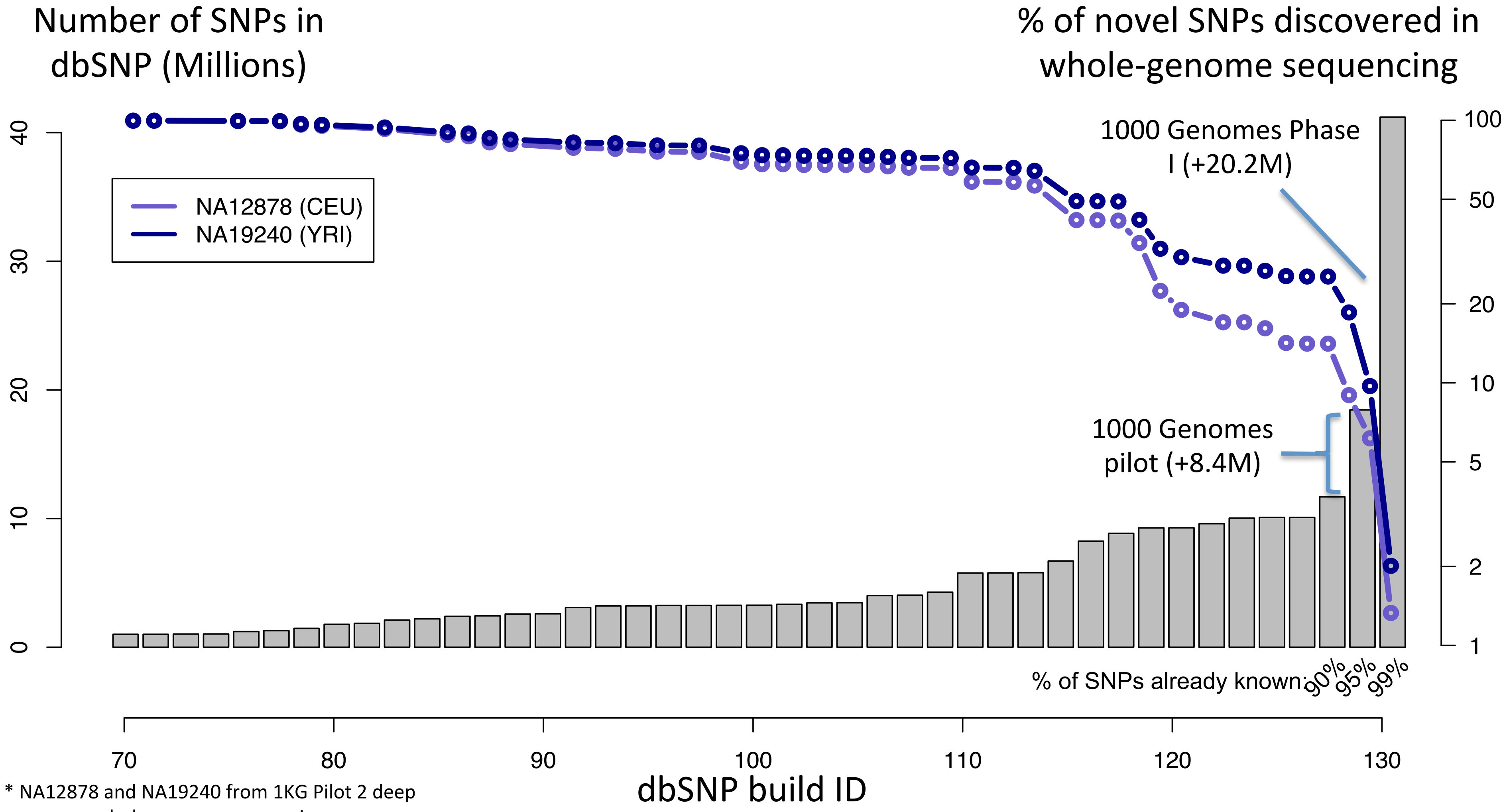
	Autosomes	GENCODE regions*
Samples	1,092	1,092
Total raw bases (Gb)	19,049	327
Mean mapped depth (×)	5.1	80.3
SNPs		
No. sites overall	36.7 M	498 K
Novelty rate†	58%	50%
No. synonymous/non-synonymous/nonsense	NA	199/293/6.3 K
Average no. SNPs per sample	3.60 M	24.0 K
Indels		
No. sites overall	1.38 M	1,867
Novelty rate†	62%	54%
No. inframe/frameshift	NA	719/1,066
Average no. indels per sample	344 K	440
Genotyped large deletions		
No. sites overall	13.8 K	847
Novelty rate†	54%	50%
Average no. variants per sample	717	39

Phase I achieves our goal of essentially complete discovery of all variants with >1% frequency



- For 1% frequency SNPs
  - – 99.3% genome
  - – 99.8% exome
- For 0.1% frequency SNPs
  - – 70% genome
  - – 90% exome

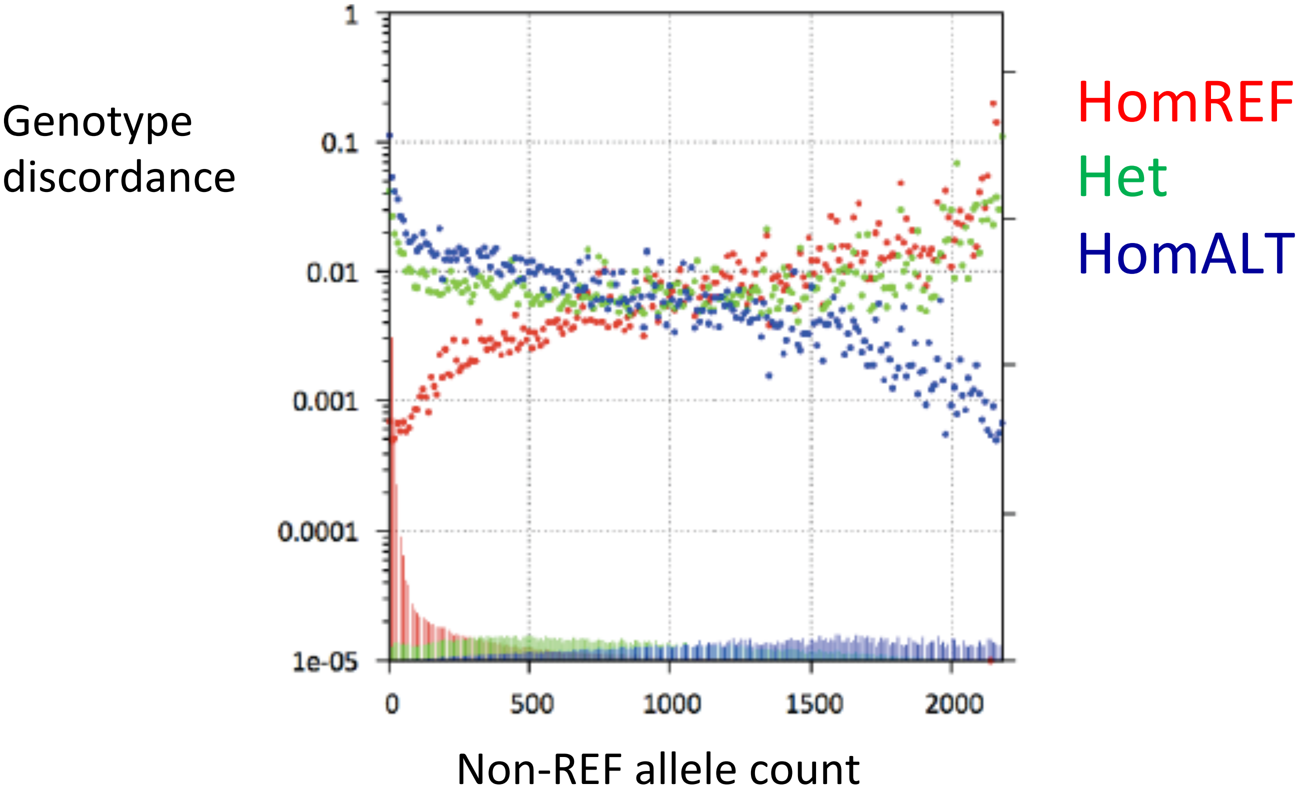
~99% of variation in each person has already been cataloged in 1000 Genomes Phase I



* NA12878 and NA19240 from 1KG Pilot 2 deep whole genome sequencing

Genotype accuracy at >99% at chip heterozygous sites

Sites	METHOD	#chr20 Variants	#OMNI Overlaps	HET (OMNI)	NREF-EITHER	OVER-ALL
LC SNPs/INDELs/SVs + EX SNPs	Beagle +MaCH	907,452	52,329	0.95%	**1.11%**	0.36%



Hyun Min Kang

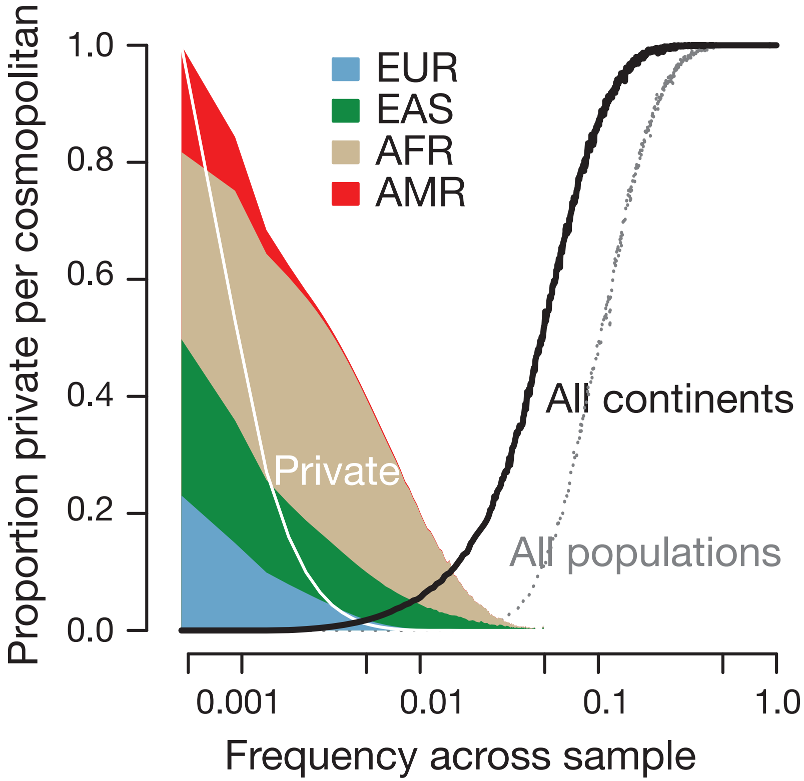
Scientific insights: many rare functional variants

- 3-4,000,000 variants per individual
- 10-11,000 nonsynonymous changes
- 220-250 in frame indels
- 80-100 premature stop codons
- 40-50 splice site disruptions
- 50-100 HGMD "recessive disease causing" mutations

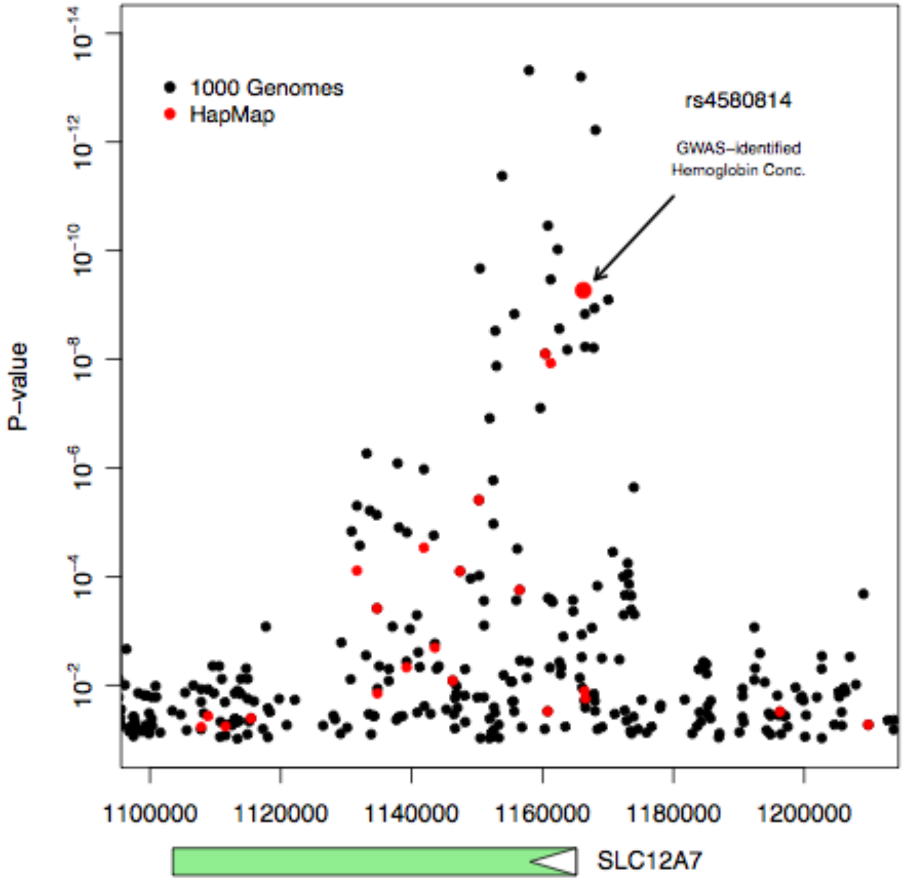
Pilot project analysis; Current data summarized in Table 2 of Phase 1 paper

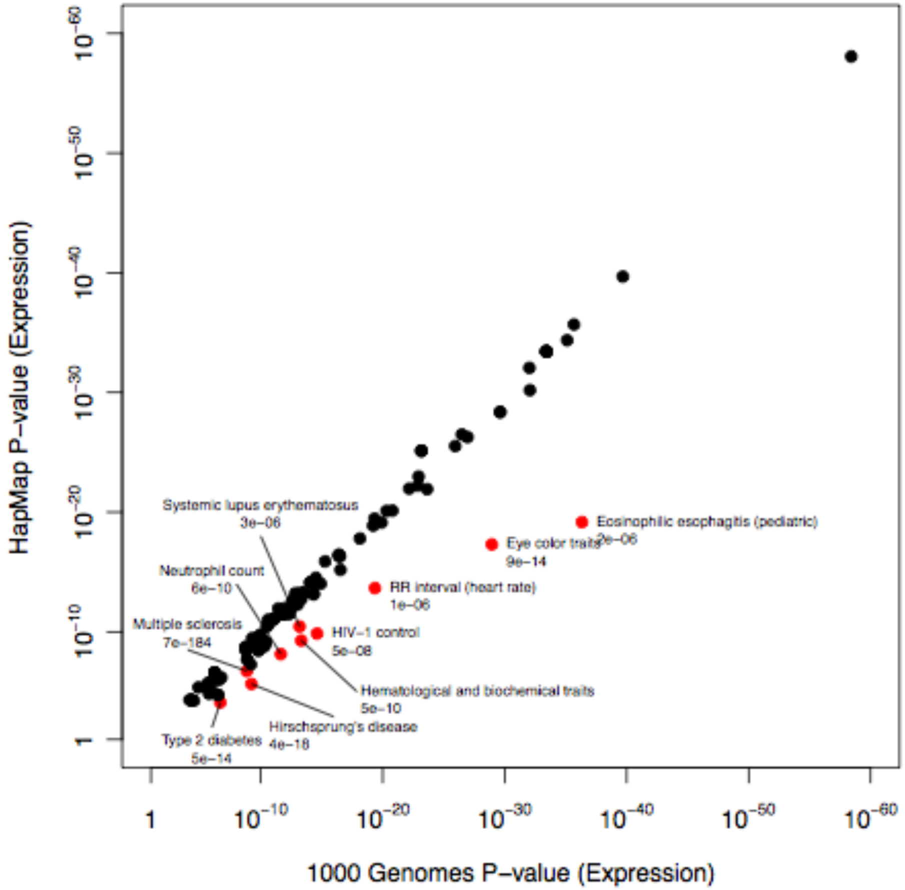
Scientific insights: rare variation is population specific

- 17% of low frequency (0.5-5%) in a single ancestry group
- 53% of less than 0.5% in a single population
- African populations have many more many low frequency variants due to bottleneck on other lineages
- All populations are enriched in rare variants
  - Explosive recent population growth



Scientific insights:
full sequence variation helps GWAS





Gaffney and Pritchard

What has 1000 Genomes given us?

- Large, public NGS datasets
- Catalogues of variants
  - Functional candidates
  - Screening list for medical sequencing
  - Basis for imputation
  - Data for population genetics analysis
- File formats and tools for NGS analysis
  - Basis for large scale medical projects

Phase 2+3 will include more populations, deeper data, better calls

- Expand into 11 more populations
  - In Africa, Asia, and Indian sub-convenient
  - 2500 samples overall
- Deeper, better data
  - At least 3x coverage, minimum of paired end 76bp
  - Exomes and exome chips for all samples
- Better calls
  - New variant calling (local and *de novo* assembly)
  - Multi-allelic haplotype integration

Upcoming talks

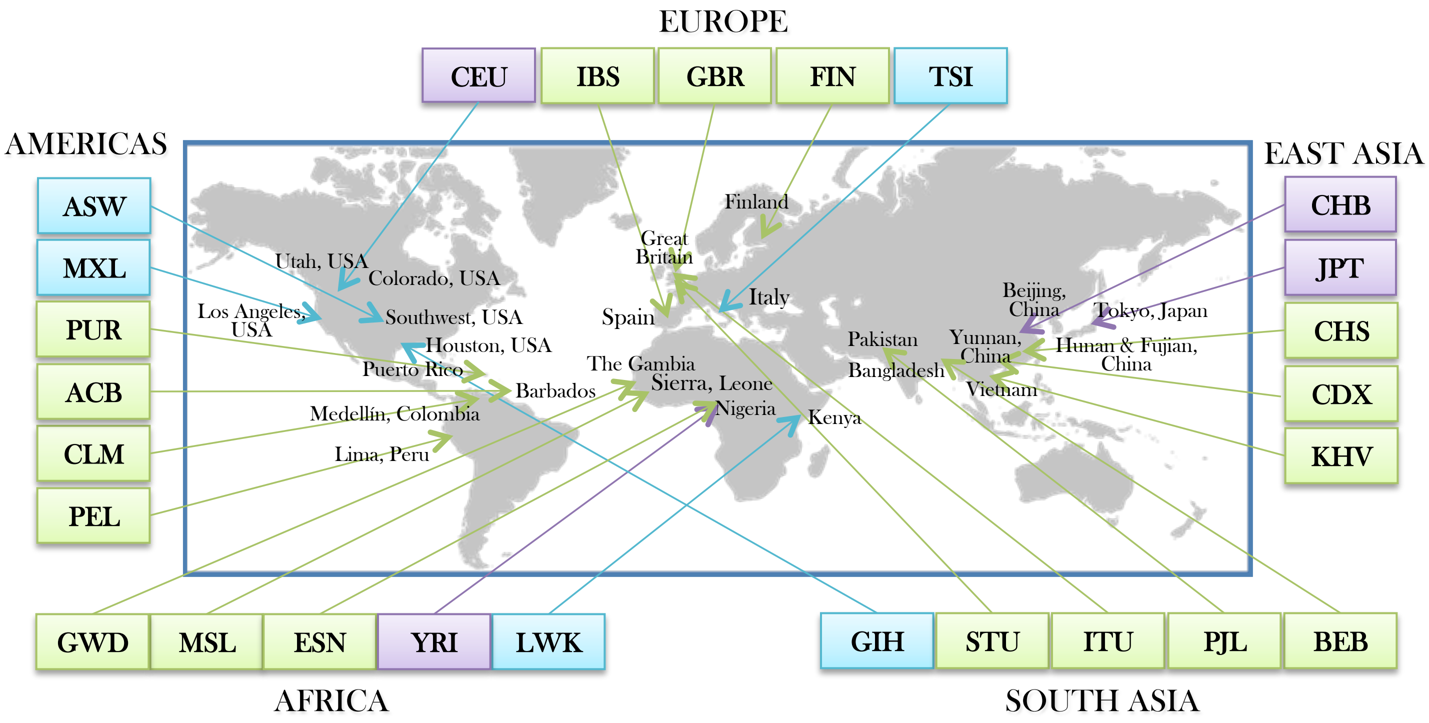
- How to access the data
  - Laura Clarke
- Structural variants
  - Ryan Mills
- Population genetic and admixture analyses
  - Eimear Kenny
- Functional analyses
  - Ekta Khurana
- How to use the data in disease studies
  - Stephan Ripke

Credits

More information at www.1000genomes.org

Paul Flicek for contributing so many slides

1000 Genomes Project Populations



EUROPE

CEU IBS GBR FIN TSI

AMERICAS

ASW MXL PUR ACB CLM PEL

EAST ASIA

CHB JPT CHS CDX KHV

AFRICA

GWD MSL ESN YRI LWK

SOUTH ASIA

GIH STU ITU PJL BEB

Utah, USA

Colorado, USA

Los Angeles, USA

Southwest, USA

Houston, USA

Puerto Rico

Barbados

Medellín, Colombia

Lima, Peru

Finland

Great Britain

Spain

Italy

The Gambia

Sierra, Leone

Nigeria

Kenya

Pakistan

Bangladesh

Yunnan, China

Vietnam

Beijing, China

Tokyo, Japan

Hunan & Fujian, China

International HapMap Population

HapMap 3 Population

New 1000 Genomes Population