Completion of the 1000 Genomes Project

Goncalo Abecasis

University of Michigan School of Public Health

Project Goals (2008)

- >95% of accessible genetic variants
  with a frequency of >1%
  in each of multiple continental regions

- Extend discovery effort to lower frequency variants in coding regions of the genome

- Define haplotype structure in the genome

Pilot Projects (2010)

- 2 deeply sequenced trios
- 179 whole genomes sequenced at low coverage
- 8,820 exons deeply sequenced in 697 individuals

- 15M SNPs, 1M indels, 20,000 structural variants

Phase I (2012)

- More diverse set of populations sequenced
  - Total >1,092 individuals (EUR, ASN, AFR, AMR groupings)

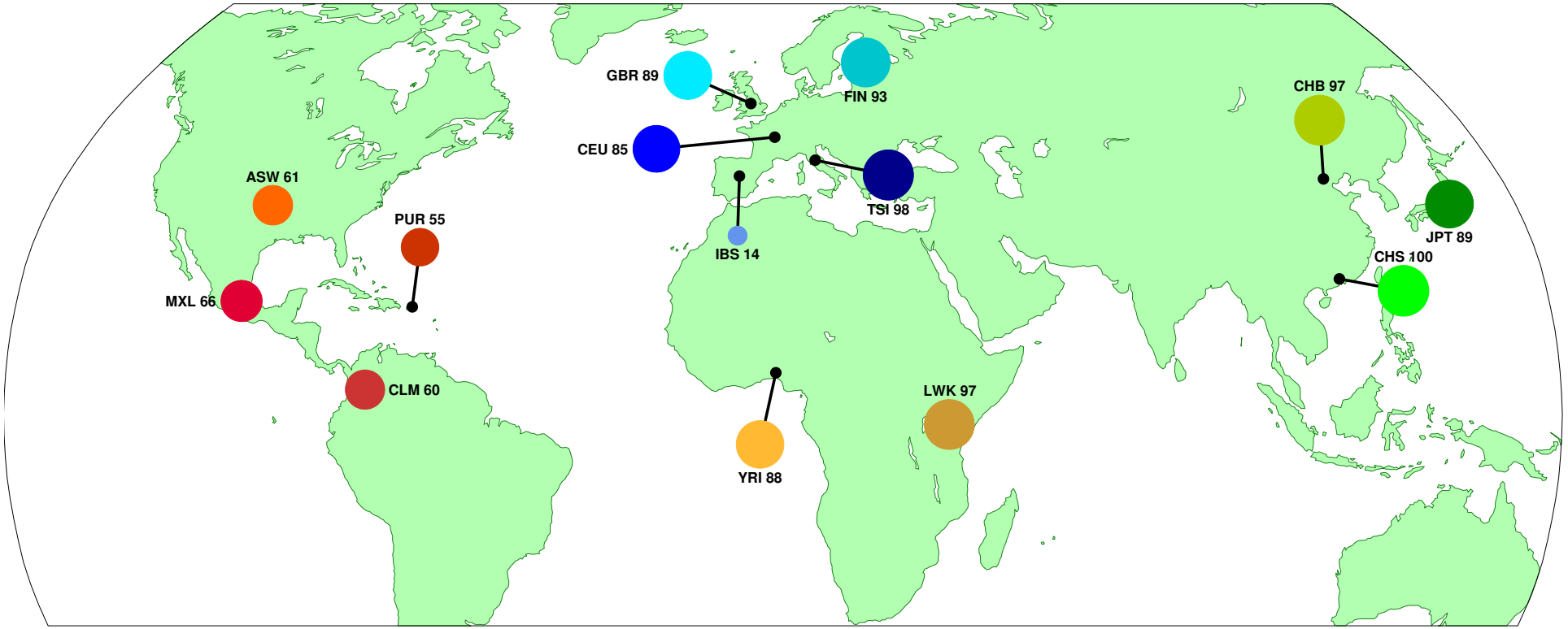
- >38.5 million SNP
  - 8.5M sites discovered before project (dbSNP 129)
  - 30M sites newly discovered
  - 98.9% of HapMap III sites rediscovered
  - Transition/transversion ratio of 2.16 vs 2.04 in pilot

- ~1.5M insertion deletion polymorphisms

- ftp://ftp.1000genomes.ebi.ac.uk
- ftp://ftp.ncbi.nlm.nih.gov/1000genomes/

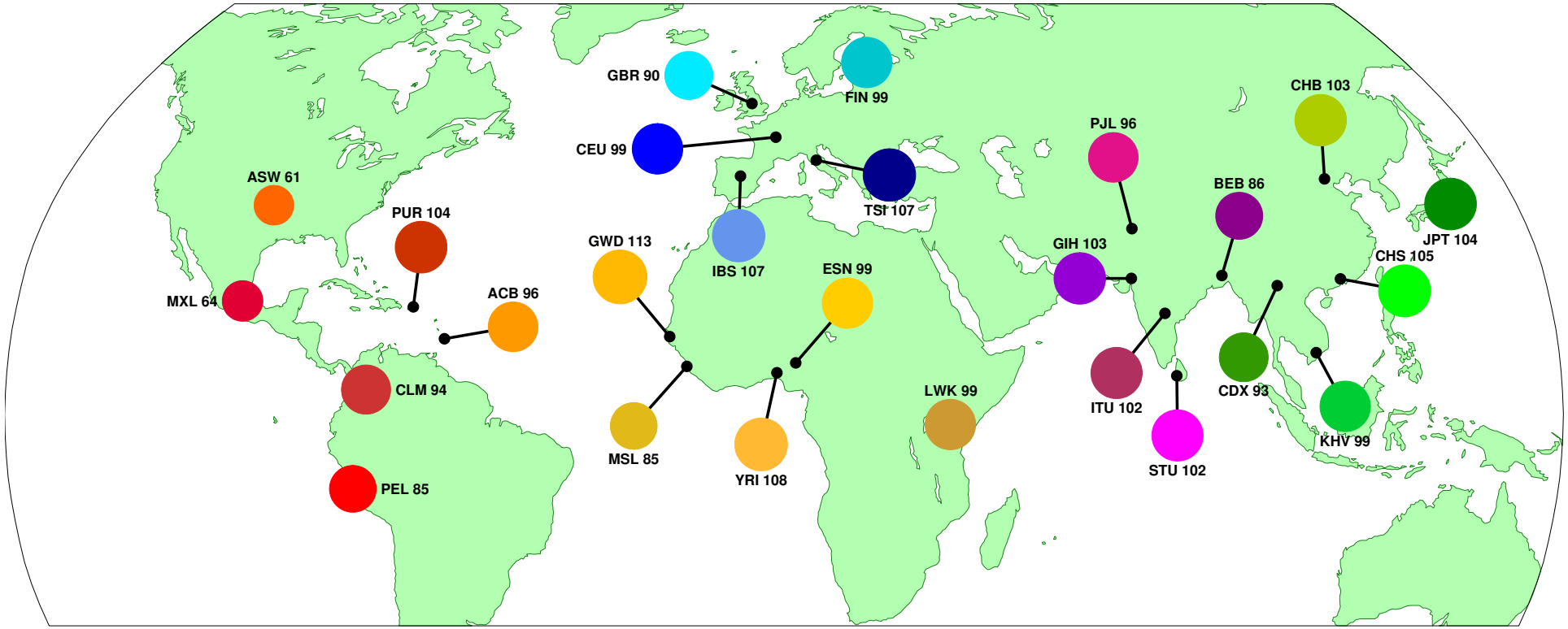
The 1000 Genomes Project (Nature, 2012)

Phase 1 samples



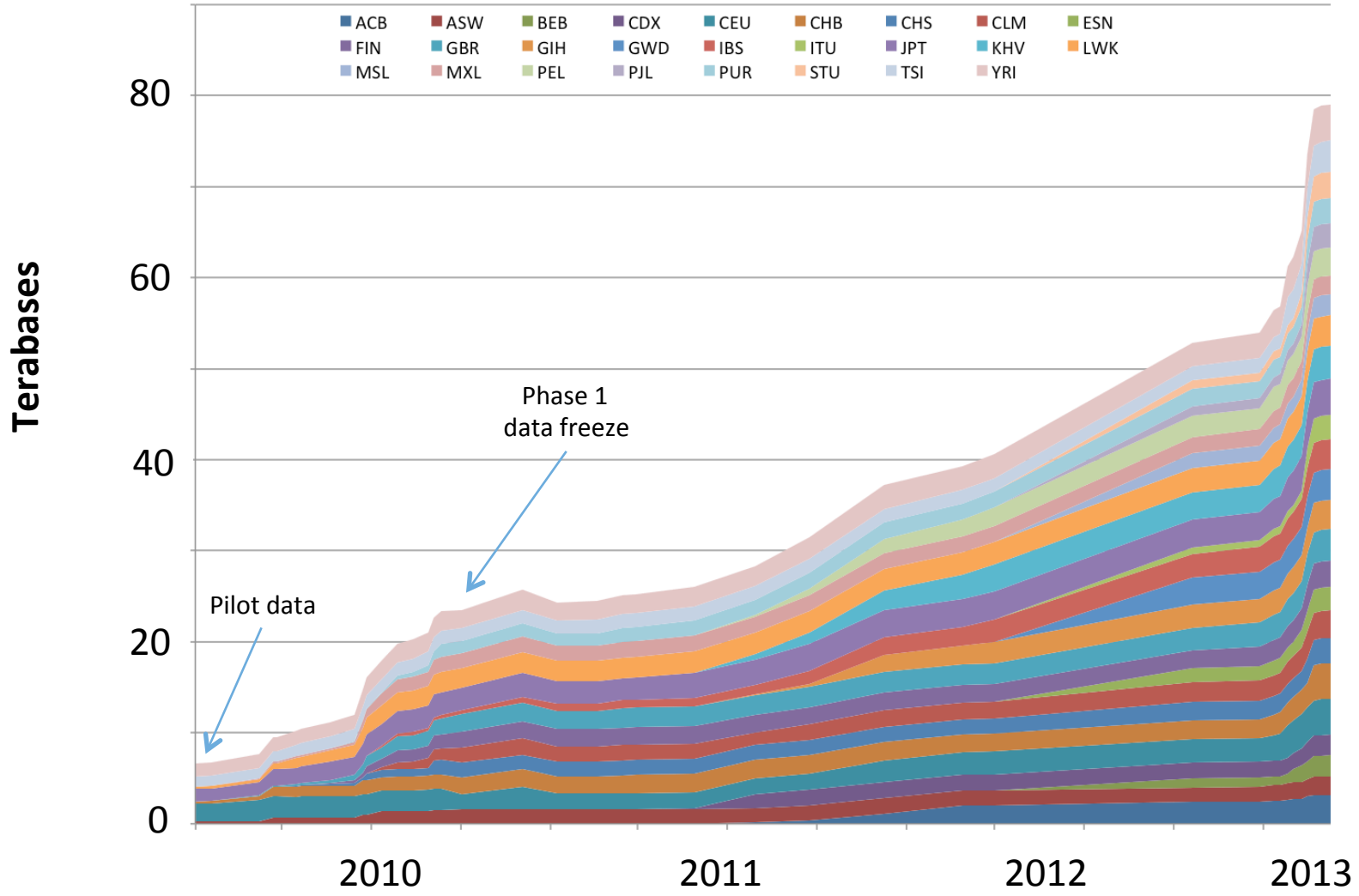
Bubble size = sample size

Samples in the final phase



Bubble size = sample size

1000 Genomes data generation



1000 Genomes Data

Total Dataset:
84 TB of BAM Files

Data Generation Complete:
May 2013

Multiple input variant callsets

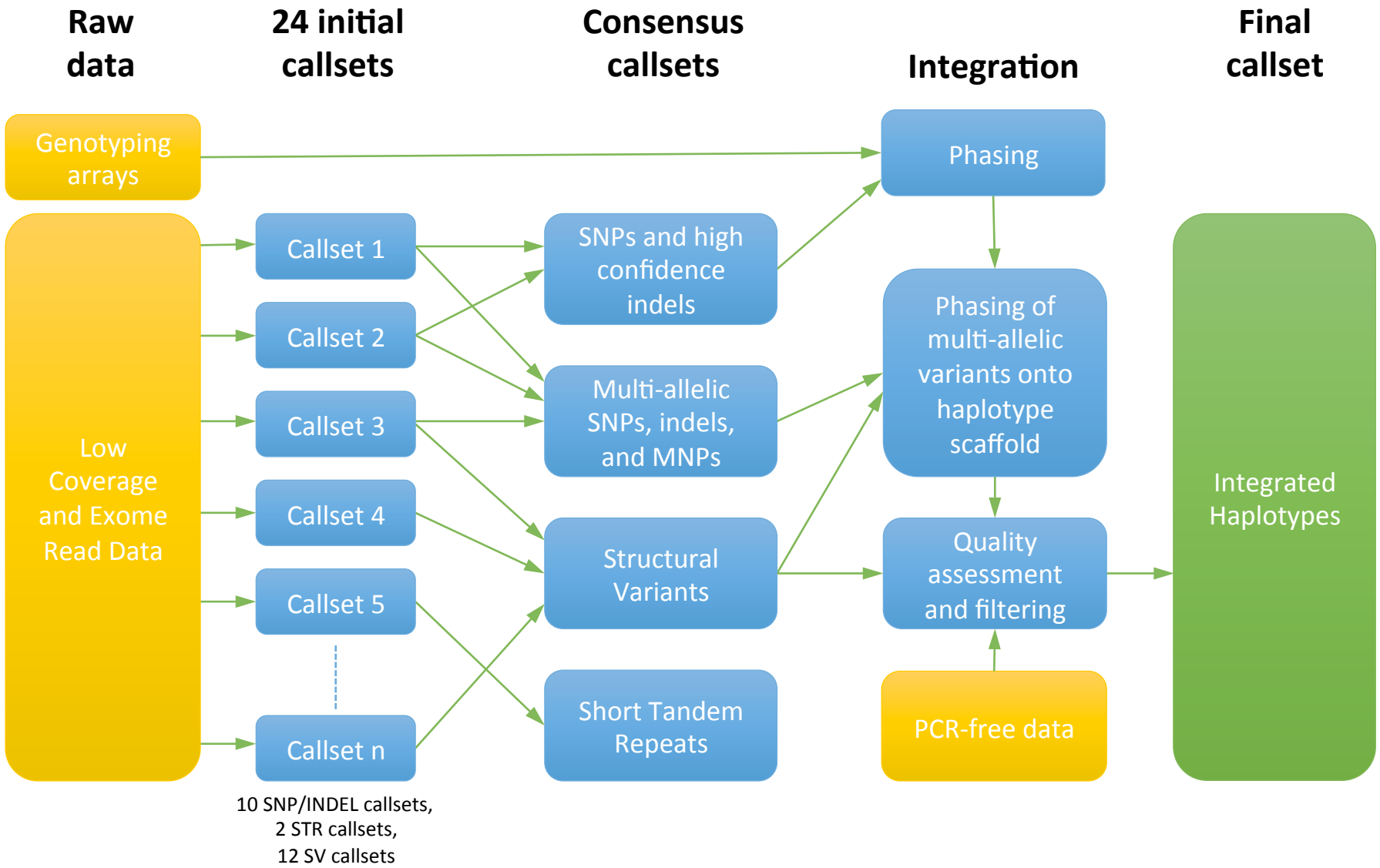
Short Variants (SNPs, indels)

- Baylor College of Medicine - SNPTOOLs
- Sanger - SAMtools
- Sanger - SGA-Dindel
- Broad - UnifiedGenotyper
- Broad - HaplotypeCaller
- Boston College - Freebayes
- Umich - GotCloud
- Oxford - Platypus
- Oxford - Cortex
- Stanford - Realtime Genomics

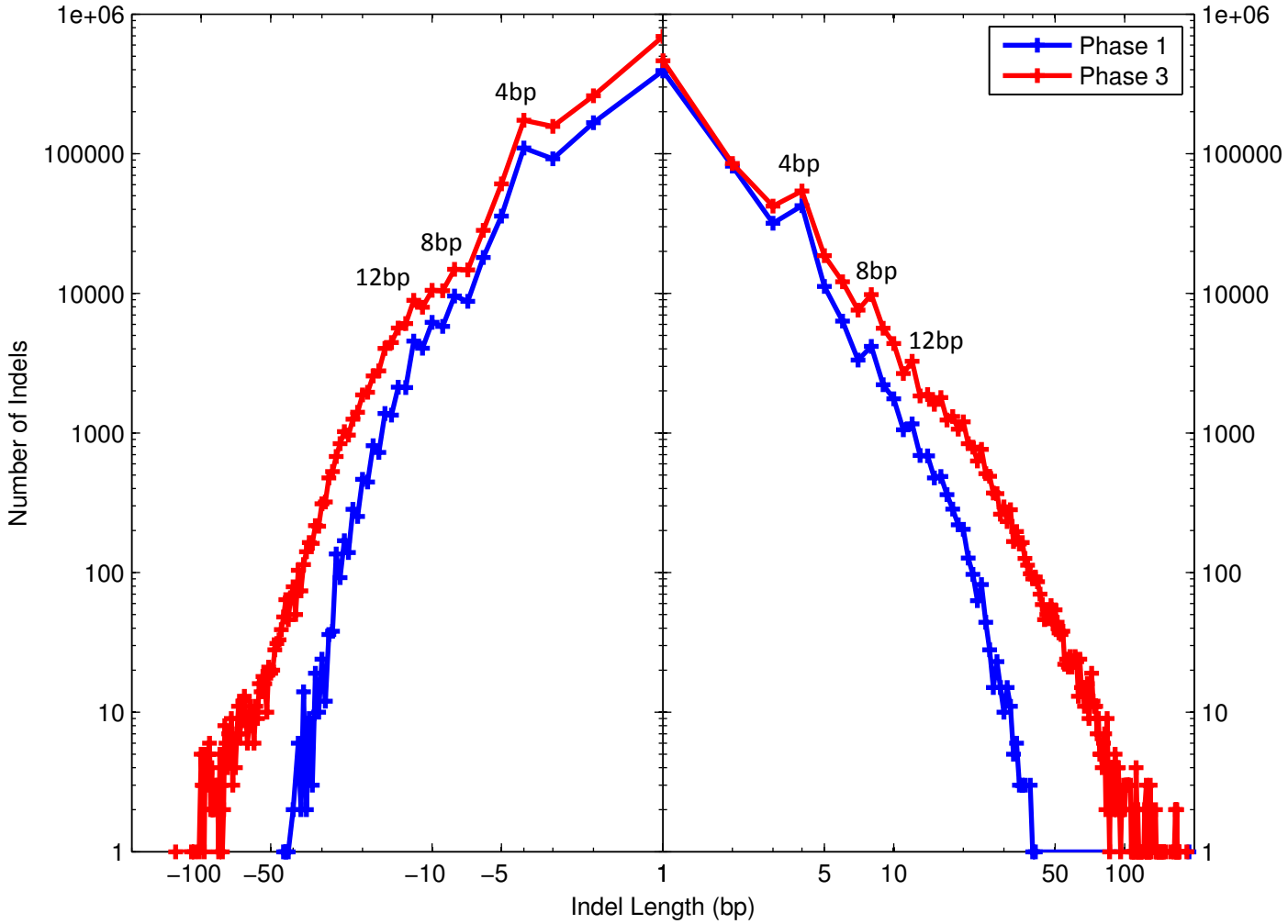
Structural Variants

- Variation Hunter
- MD Anderson TIGRA
- Delly
- Invy
- Genome STRiP
- Breakdancer
- Dinumt
- CNVnator
- Boston College ALUs
- Pindel
- UMaryland MELT
- UW Deletions

Phase 3 variant calling pipeline



More and longer indels



Quality Control of Short Variants

- For short variants, the high coverage PCR-free data from 26 individuals was used to assess the false discovery rate for each variant type.
- An allele is considered ‘validated’ if multiple supporting reads can be identified in PCR-free data.
- Sites included in the Phase 3 haplotypes have been selected to control the allele False Discovery Rate at 5%.

	Haplotype scaffold		MVNcall variants	
Variant Type	Bi-allelic SNPs	Bi-allelic Indels	Multi-allelic SNPs	Multi-allelic indels
Per-allele FDR	4.07%	0.59%	4.91%	4.95%

Quality Control of Structural Variants

- **CNVs (deletions, duplications)**
  - Custom 400K aCGH array
  - Custom 1M aCGH array
  - Omni 2.5 probe intensities
  - Affy 6.0 probe intensities
  - Complete Genomics data

- **Mobile element insertions (MEIs)**
  - PCR (N ~ 100 per MEI class)
  - Custom aCGH, incl. targeted breakpoint probes

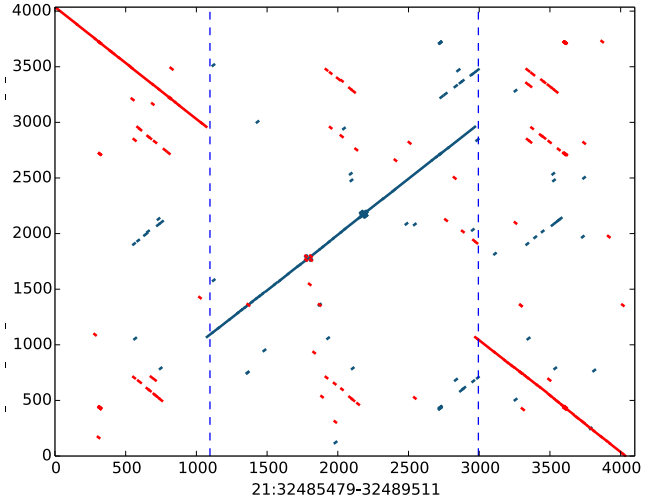
- **NUMTS (Nuclear mitochondrial sequence insertions)**
  - PCR (N = 50)
  - PacBio Amplicon-Seq.

- **Inversions**
  - PCR *(*N ~ 50)
  - PacBio Amplicon-Seq

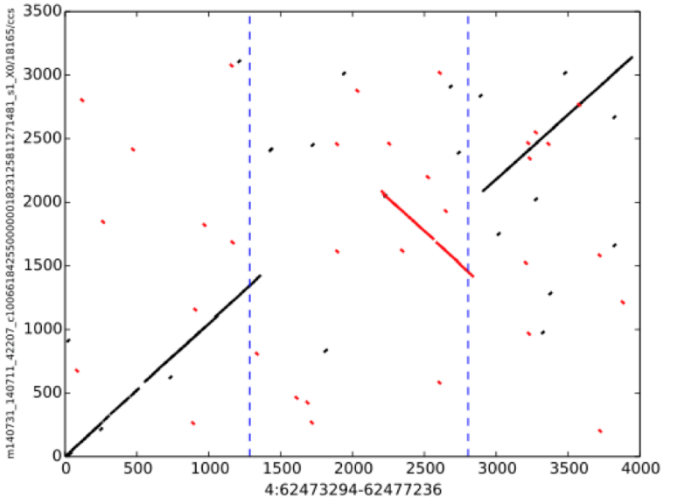
- **Multiple SV types**
  - PacBio NA12878 WGS data to resolve breakpoints & verify sites
  - PCR-free trios to verify sites and genotypes

Verification & further characterization of inversions by PacBio sequencing

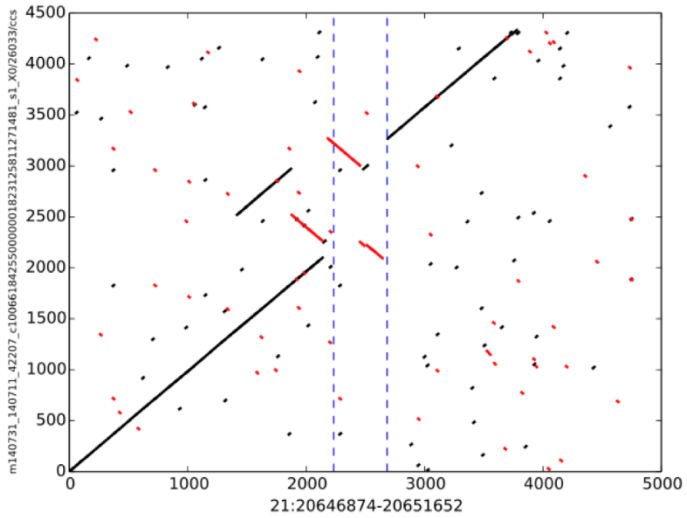
Regular ("simple") inversion



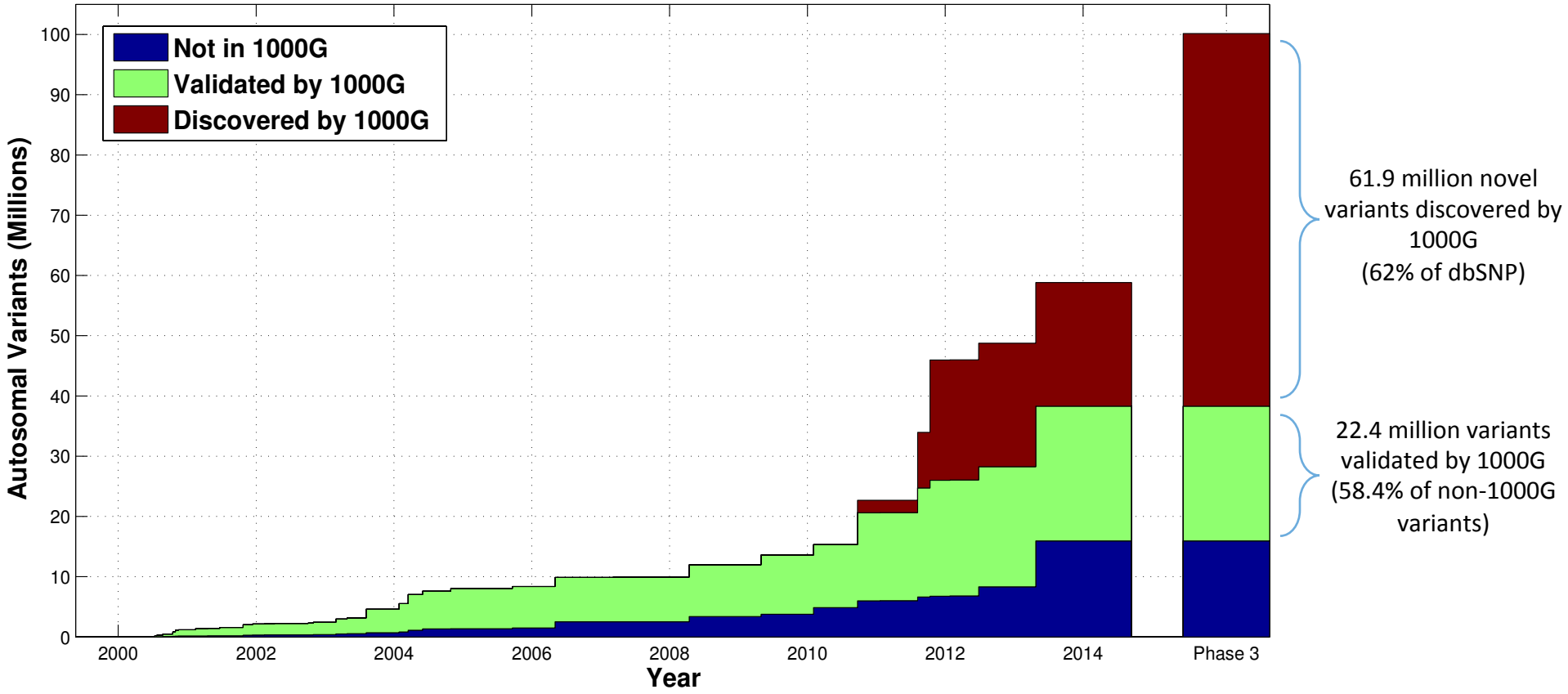
Inversion with flanking deletion



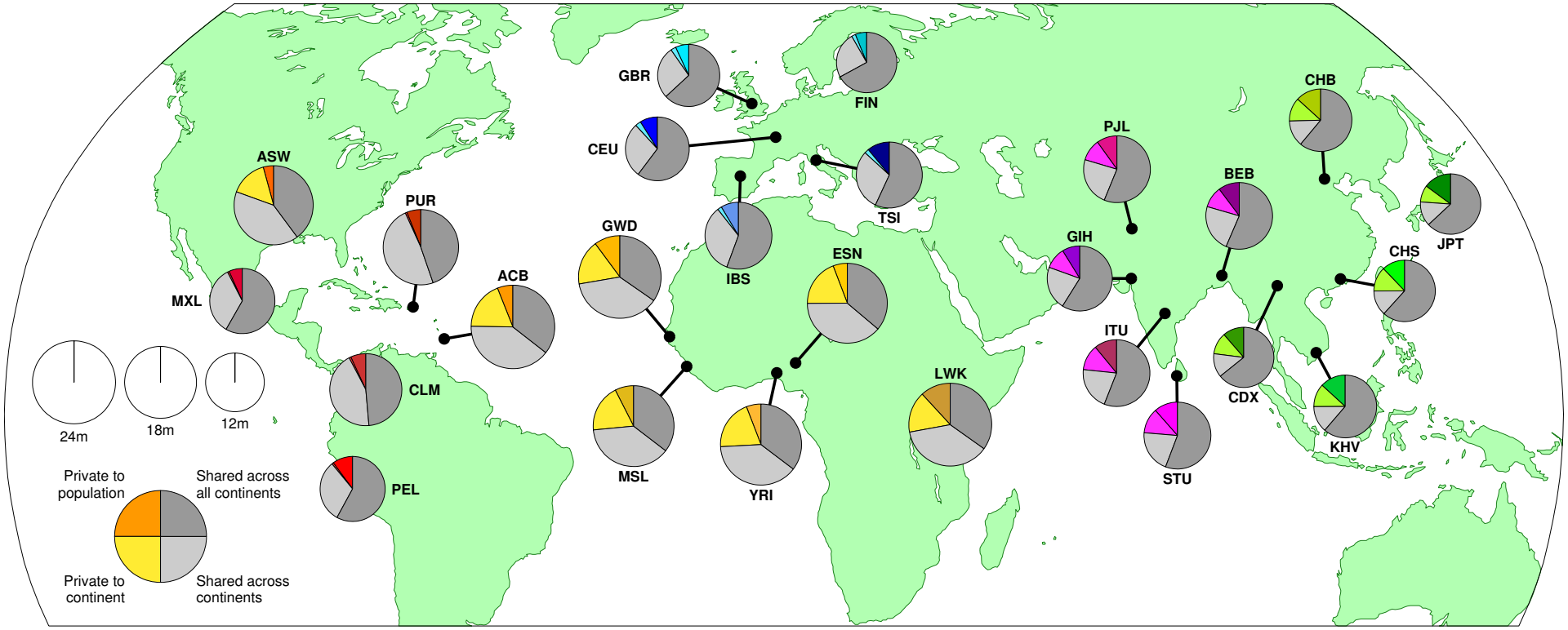
Complex SVs with inverted sequences



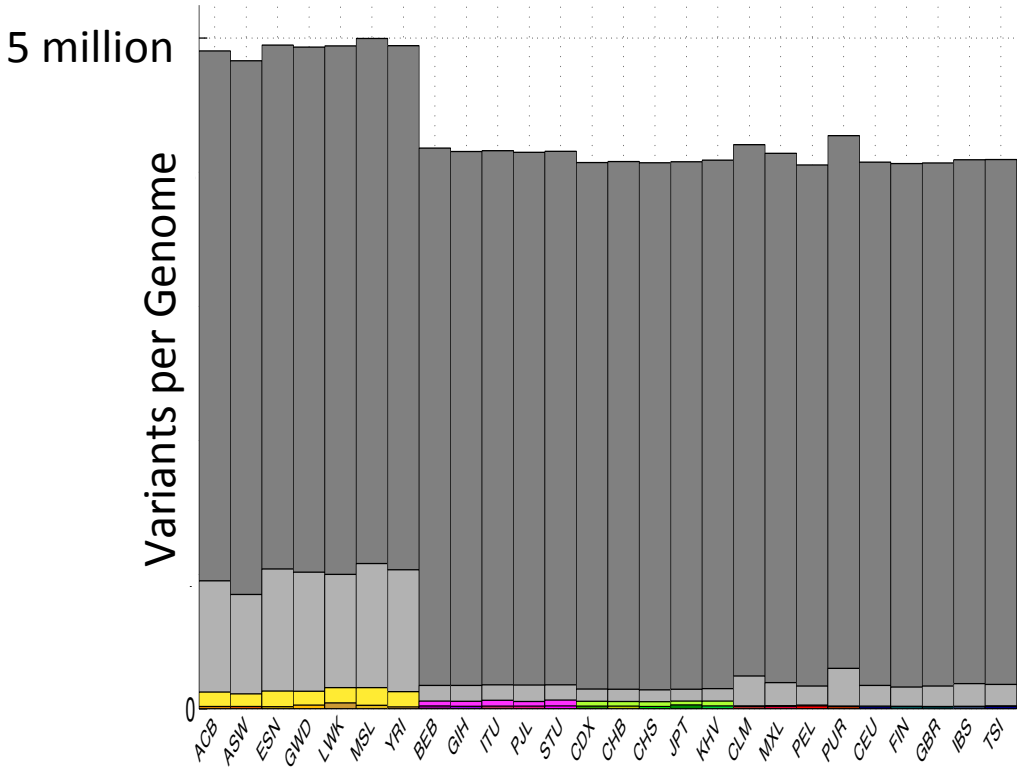
Contribution of the 1000G to dbSNP

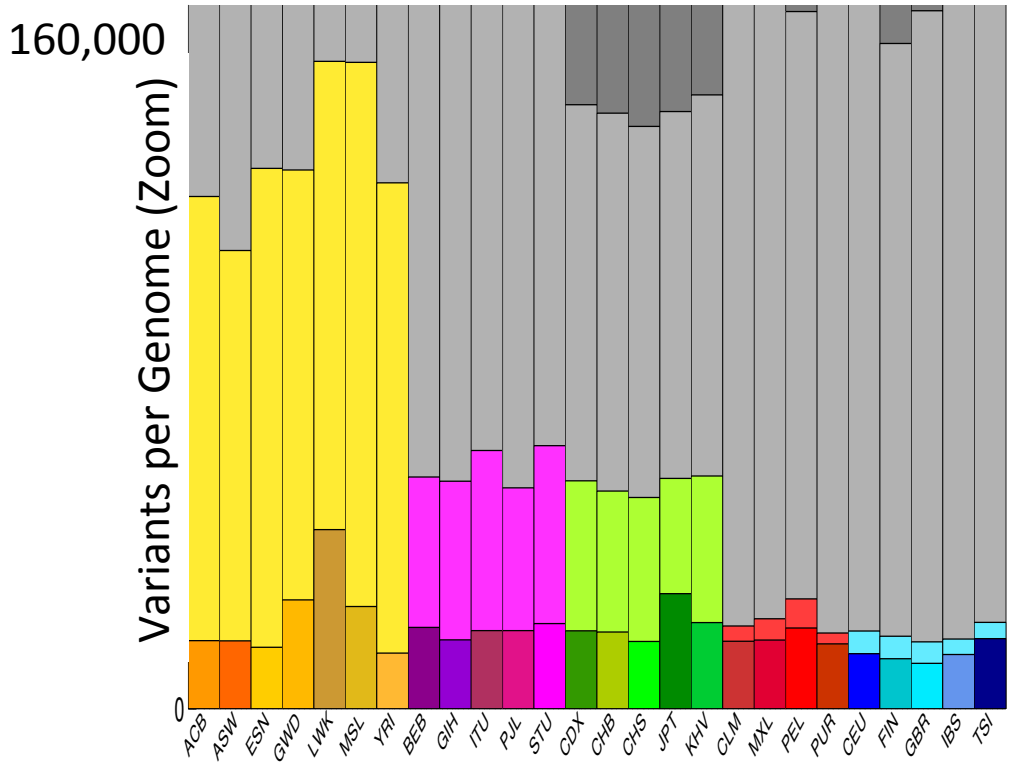


Private vs. Shared Variation (Population View)



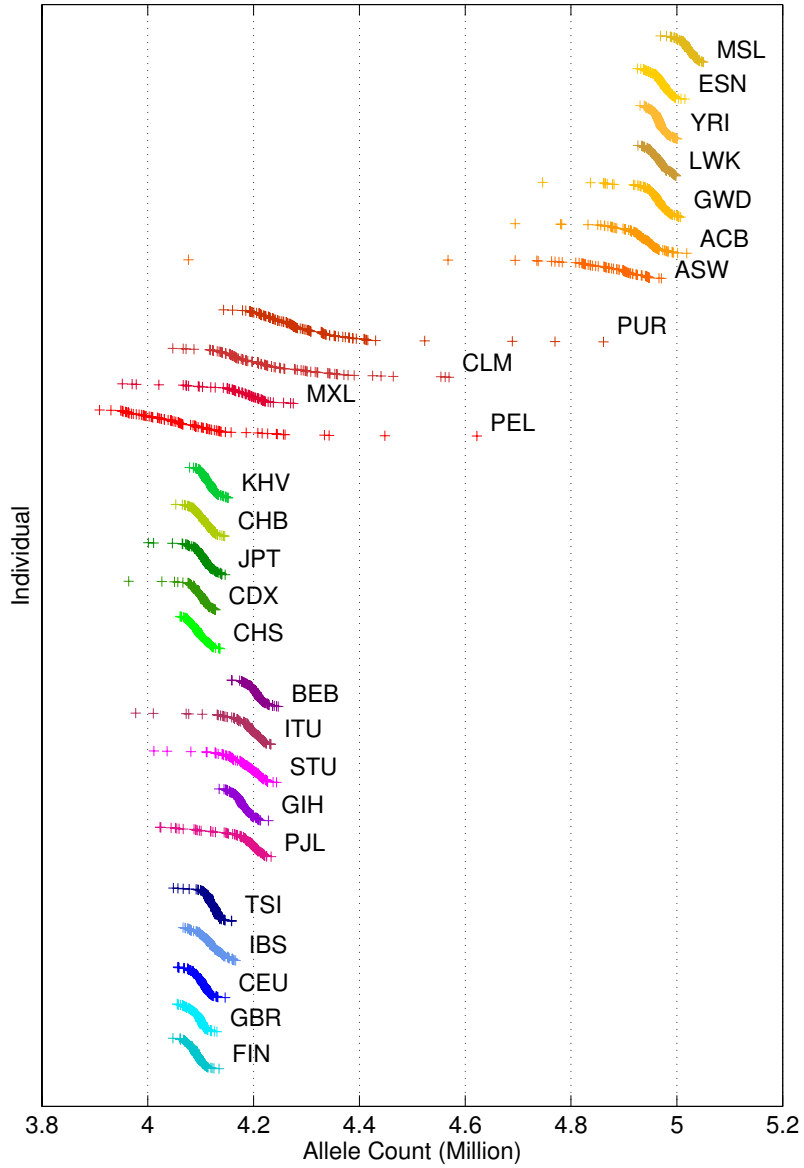
Private vs. Shared Variation (Individual View)



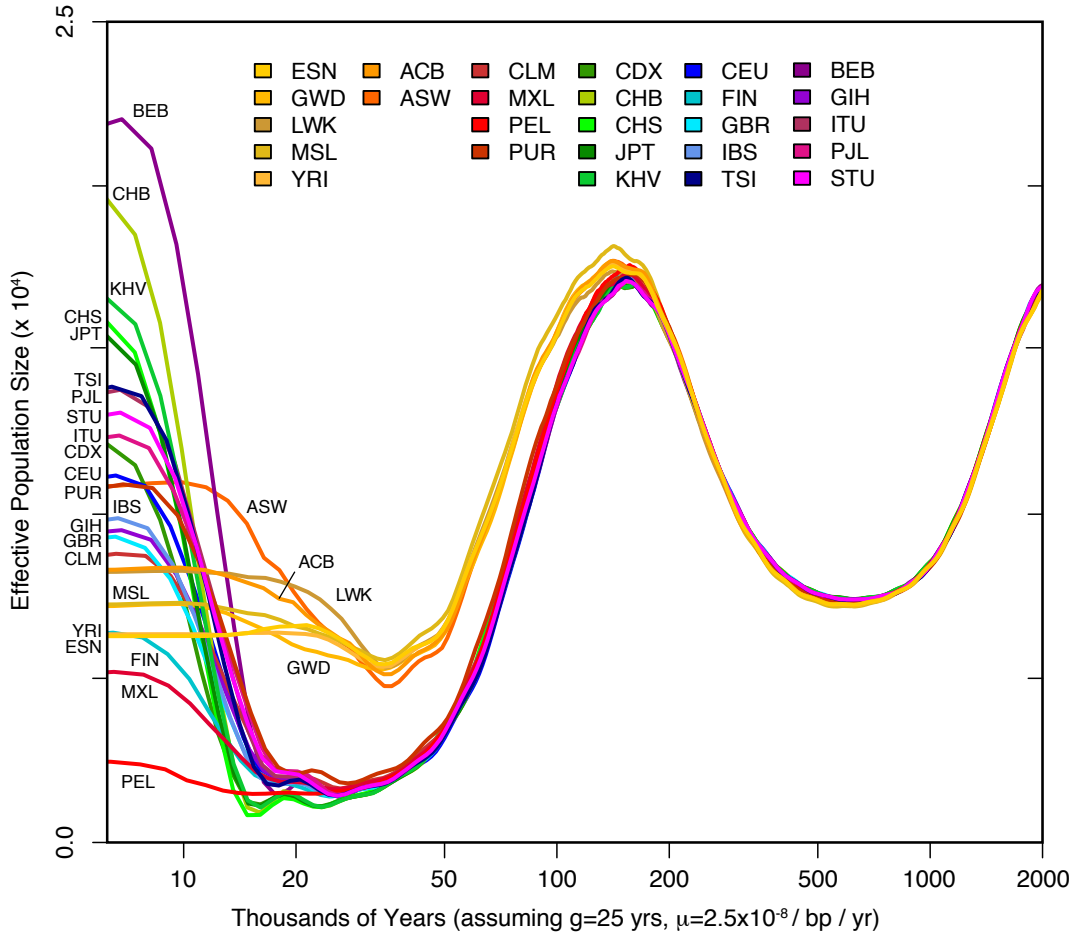


Variants per genome

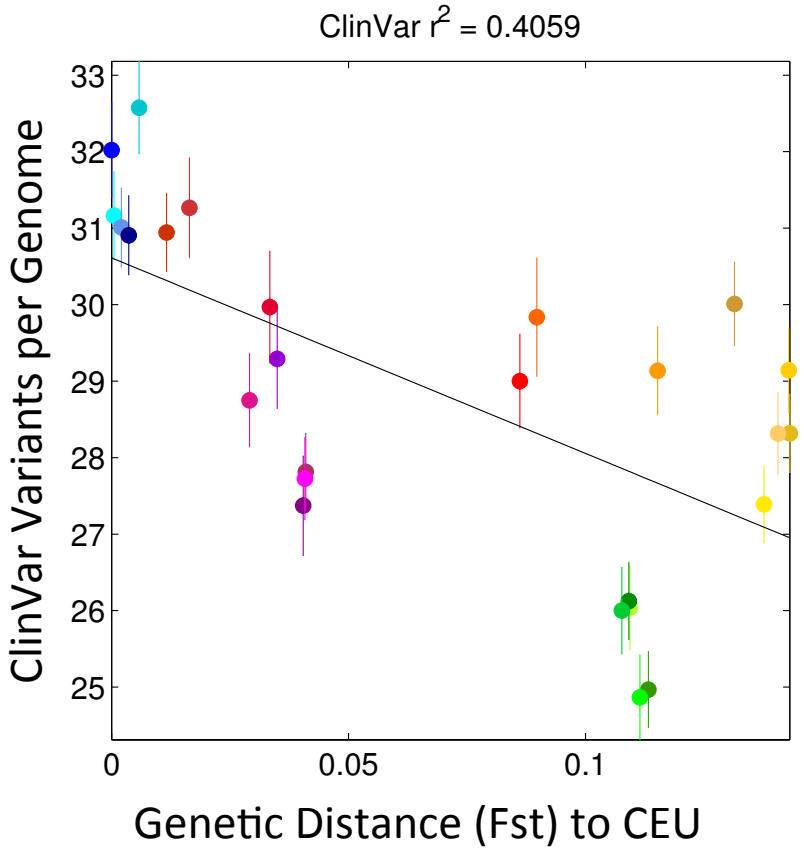
Type	Variant sites / genome
SNPs	$3.8 * 10^6$
Indels	$5.7 * 10^5$
Mobile Element Insertions	~1000
Large Deletions	~1000
CNVs	~150
Inversions	~11



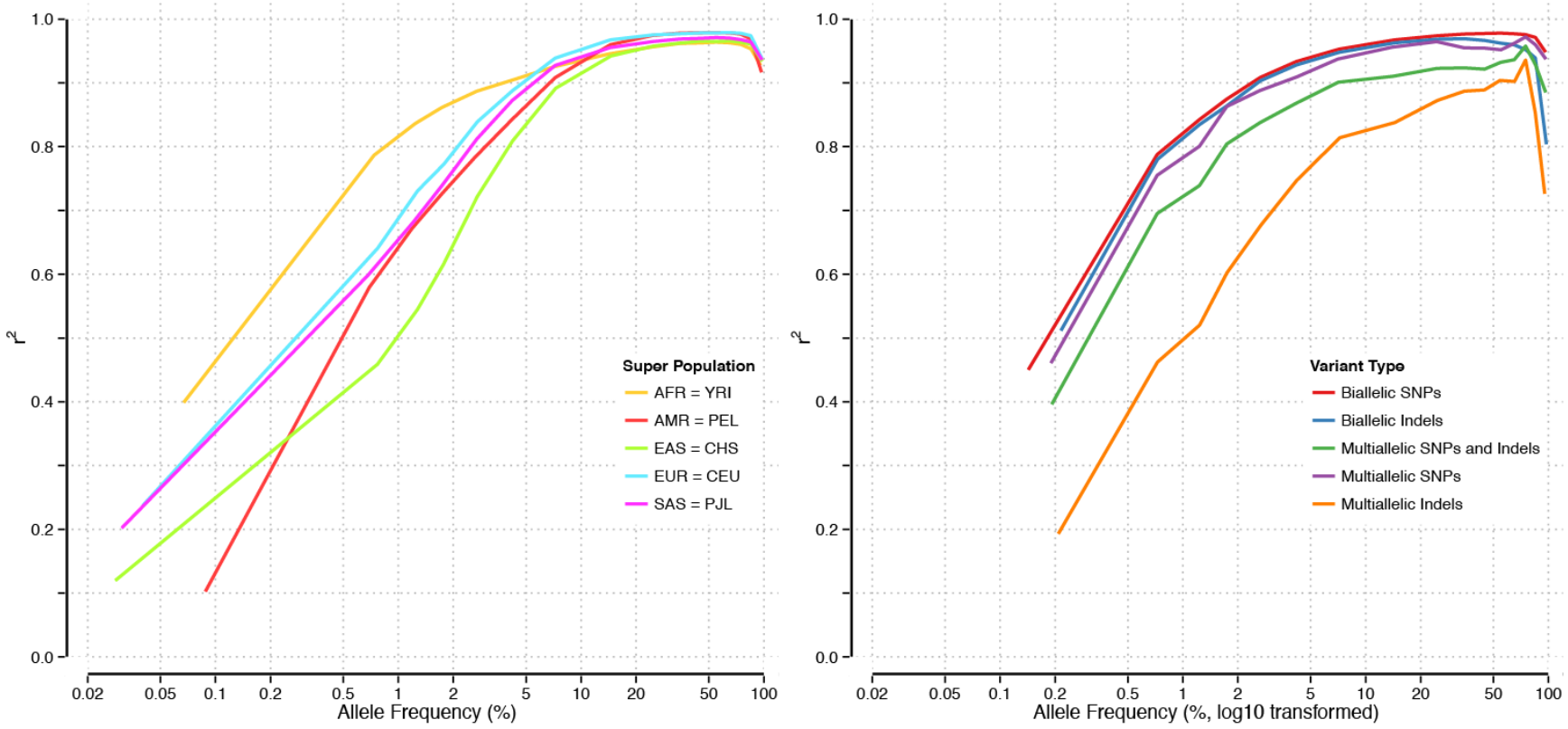
Population histories



Biases in Variation Databases?

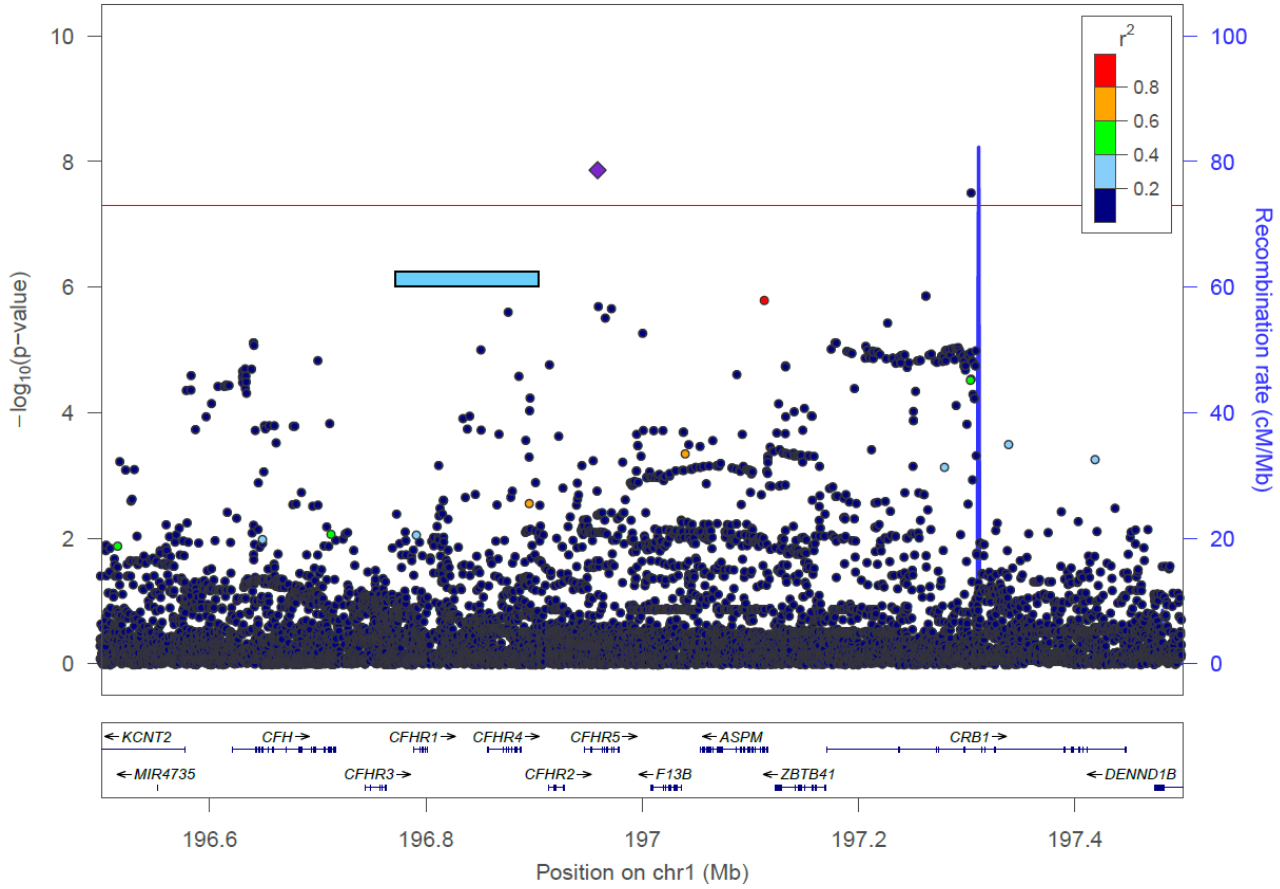


Imputation Accuracy

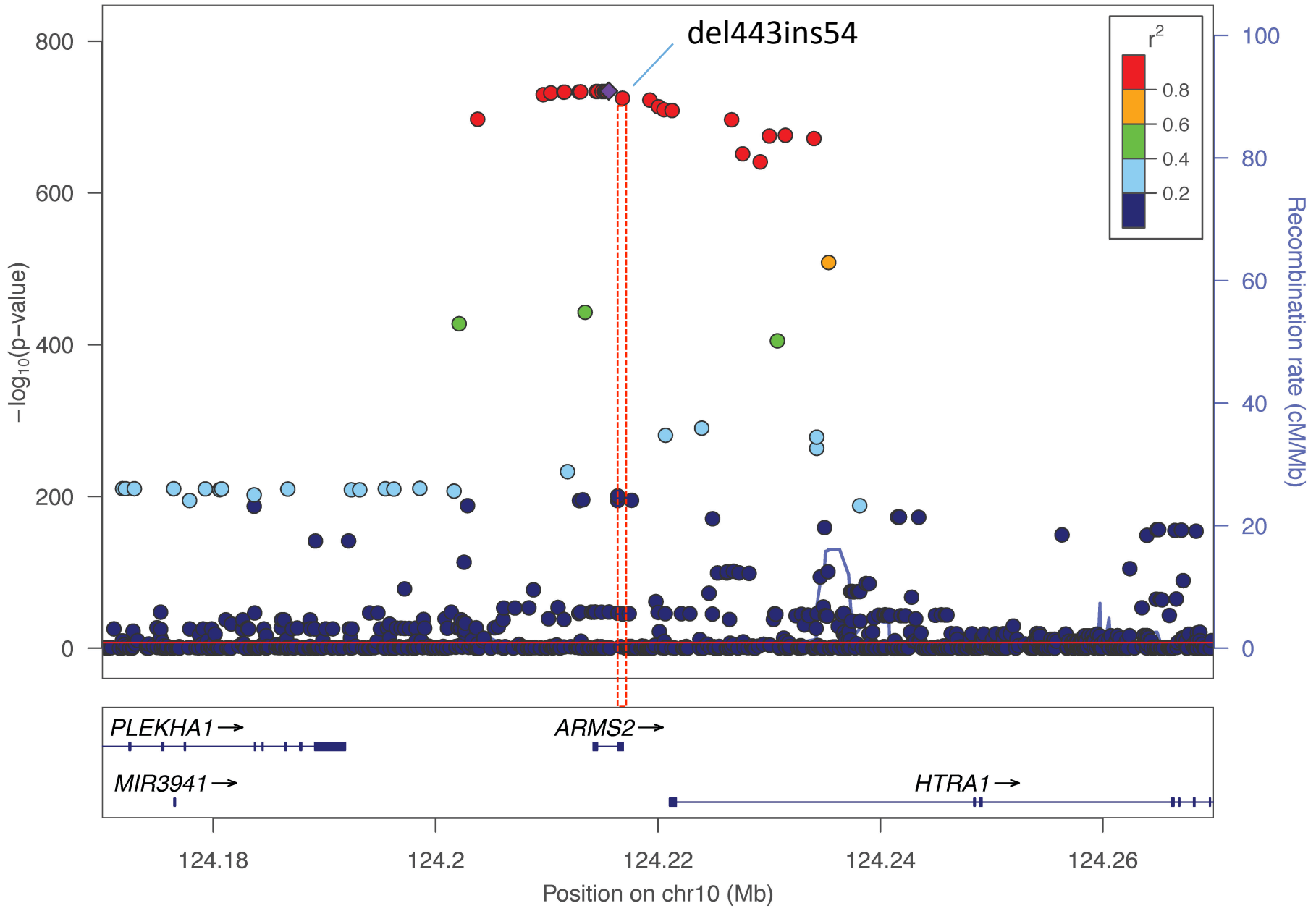


TODO: Multiallelic SNPs and indels to be renamed

AMD Imputation Example #1



Imputation
Example #2



Parting Thoughs

- Variation is extremely rare
  - In any one genome, nearly all variation is shared …
  - But almost all variants are unique to a population or continent
- Great benefits to integrated analyses
  - But analyses still requires time comparable to data generation
- Major improvements in genome coverage, variant quality and integration
- Advances can be transferred to disease studies through imputation

Phase 3 release

- The Phase 3 final release is available at www.1000genomes.org .