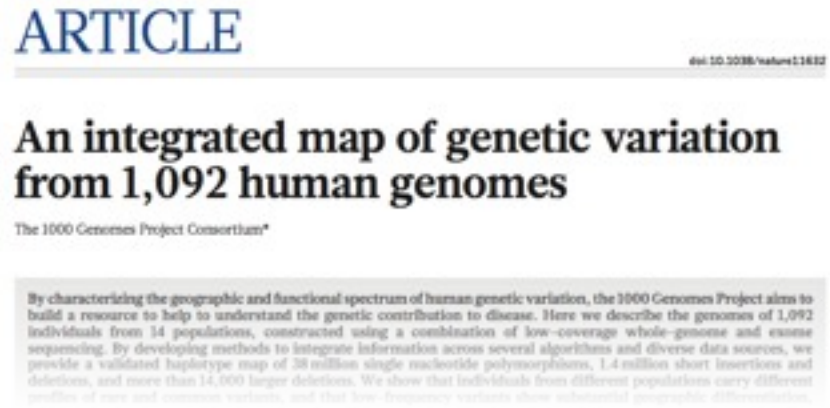
The 1000 Genomes Phase 3 Release

Shane McCarthy
Wellcome Trust Sanger Institute

Since Phase 1 of the 1000 Genomes

- Published in late 2012
  - Combined low coverage whole genome and targeted exome sequencing.
  - 1,092 individuals from 14 populations.
  - ~39.4 million variants.



ARTICLE

doi:10.1038/nature11632

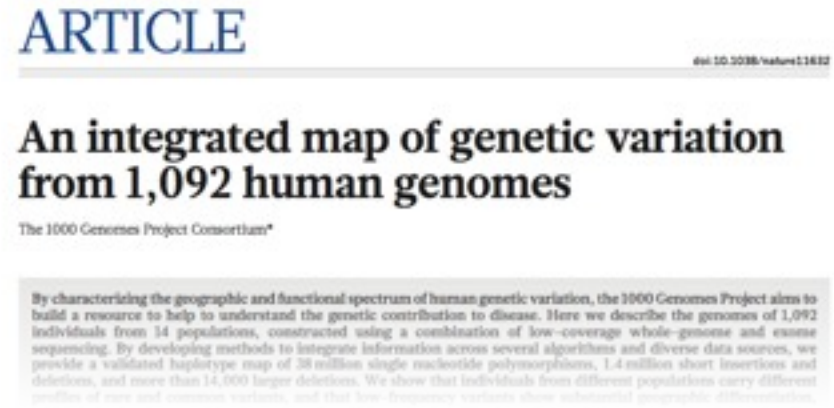
An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation,

Since Phase 1 of the 1000 Genomes

- Published in late 2012
  - Combined low coverage whole genome and targeted exome sequencing.
  - 1,092 individuals from 14 populations.
  - ~39.4 million variants.

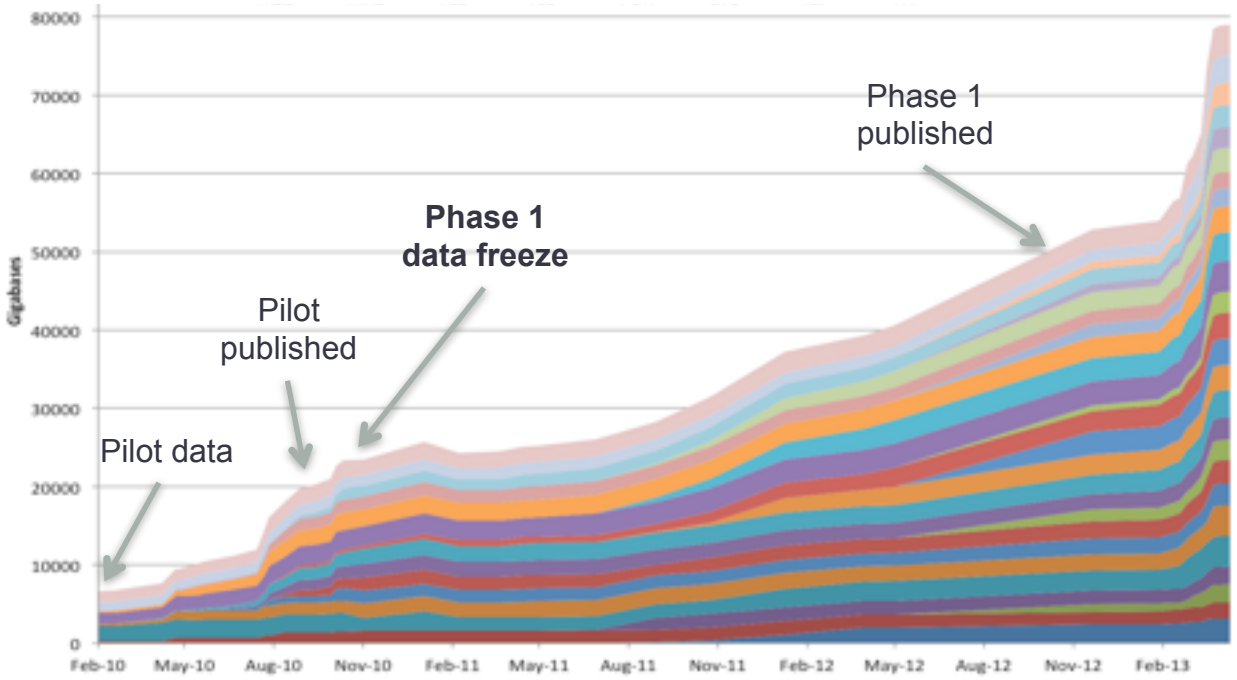


ARTICLE

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

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation.



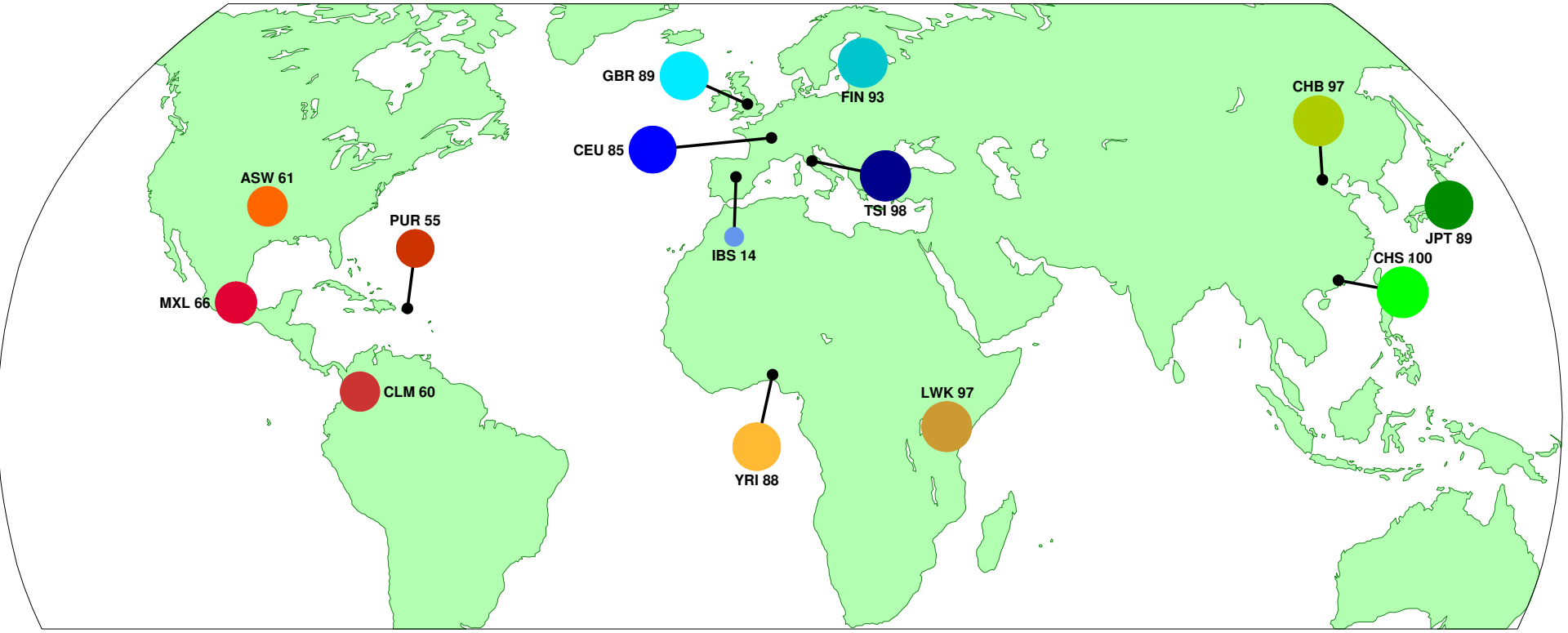
Data as of today

60Tb WGS BAMs
24Tb Exome BAMs

~3 fold more data than Phase 1

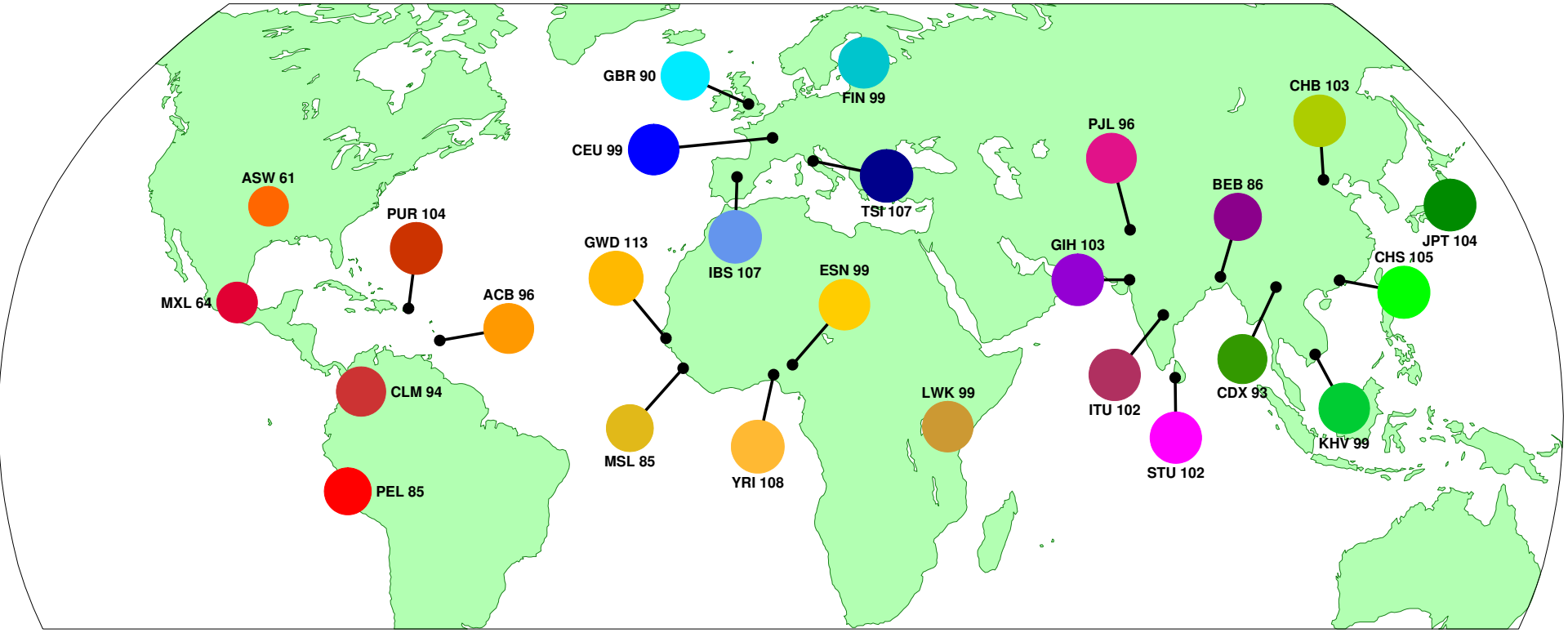
24TB 8-bin CRAM

Phase 1 samples



Bubble size = sample size

Samples in the final phase

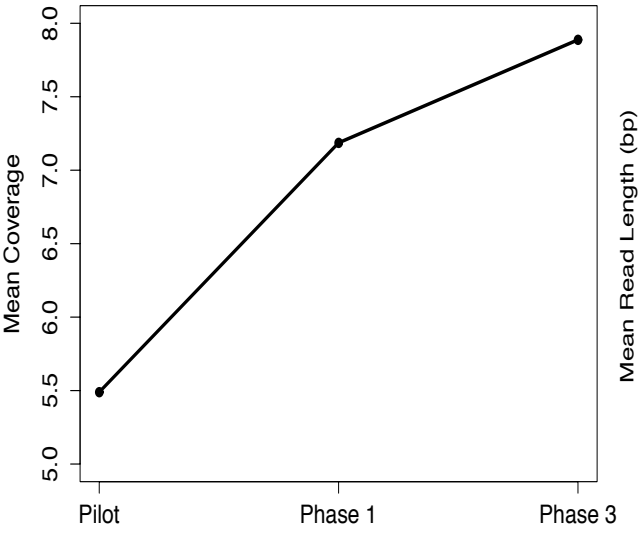


Bubble size = sample size

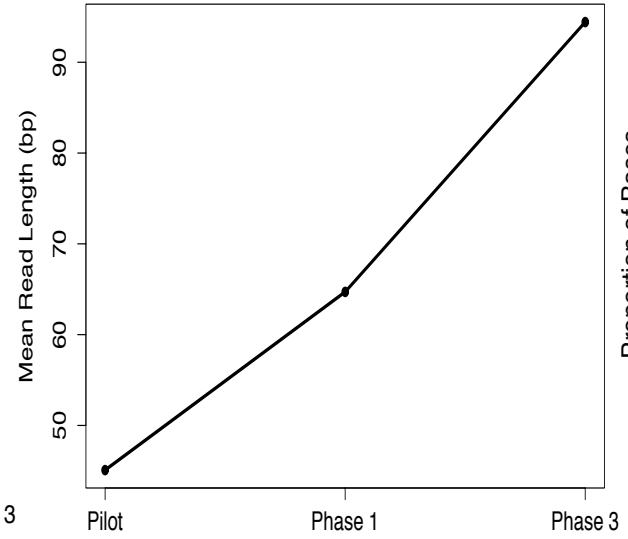
Data improvements

- 2,504 unrelated individuals from 26 populations.
  - Low-coverage whole genome and targeted exome sequencing.

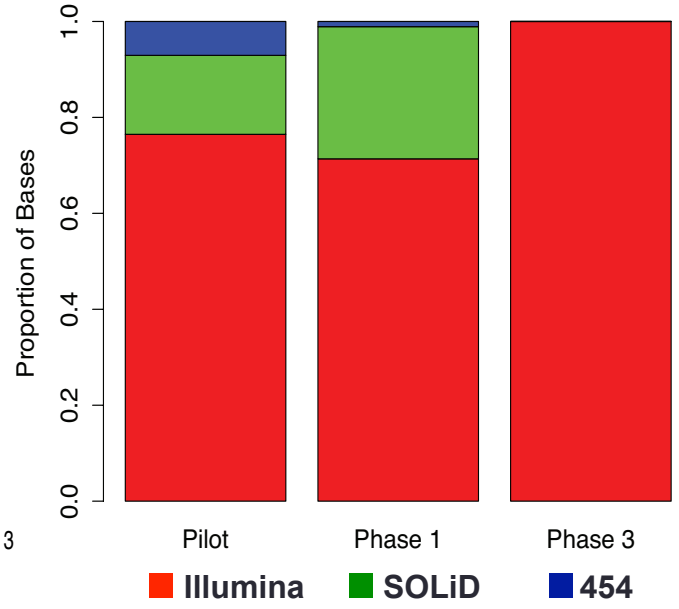
Sequencing Depth



Read Length



Sequencing Technology



Additional data

- 2,504 unrelated individuals from 26 populations.
  - Low-coverage whole genome and targeted exome sequencing.
  - Genotype data from high-density microarray (either OMNI or Affy 6.0)
- 424 high coverage Complete Genomics genomes
  - 129 family trios, 6 duos, and 25 unrelated individuals at
  - ~45X.
- High-coverage 2x250bp PCR-free sequencing for 30 individuals
  - One from each population plus updated data for the CEU and YRI trios from the 1000 Genomes Pilot.

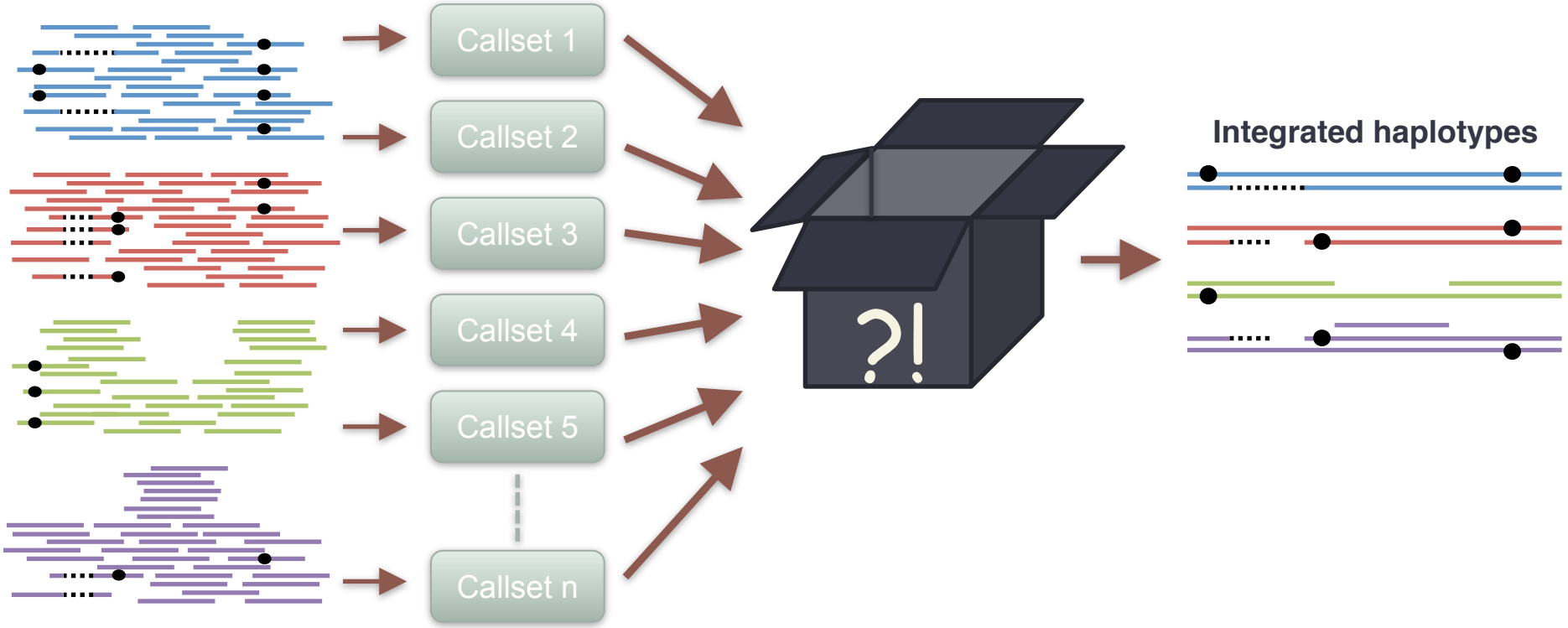
Data integration

New calling methods:

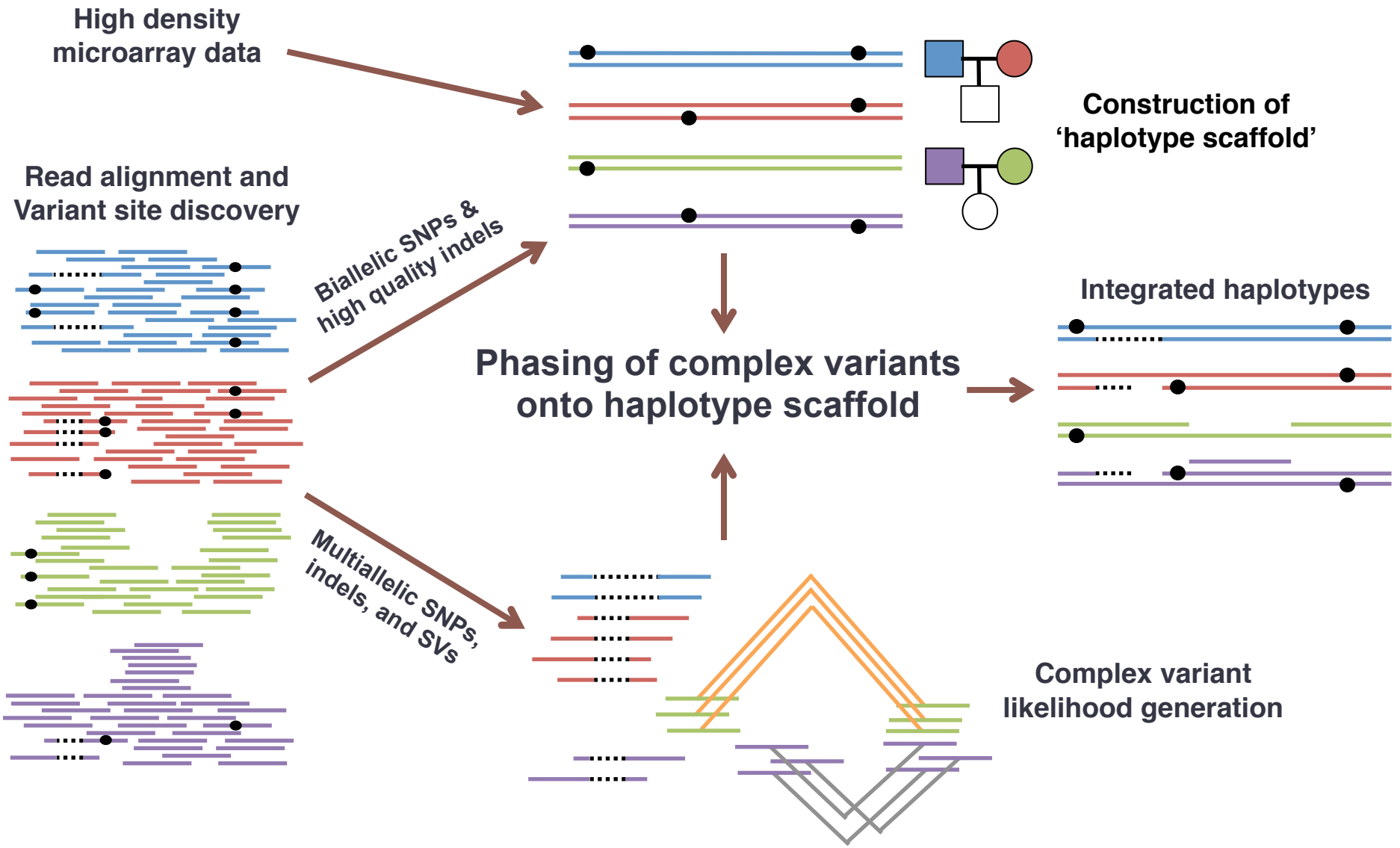
- Alignment based
- Assembly based
- Local re-assembly based

24 initial callsets

Read alignment and Variant site discovery



Integration of complex variant types



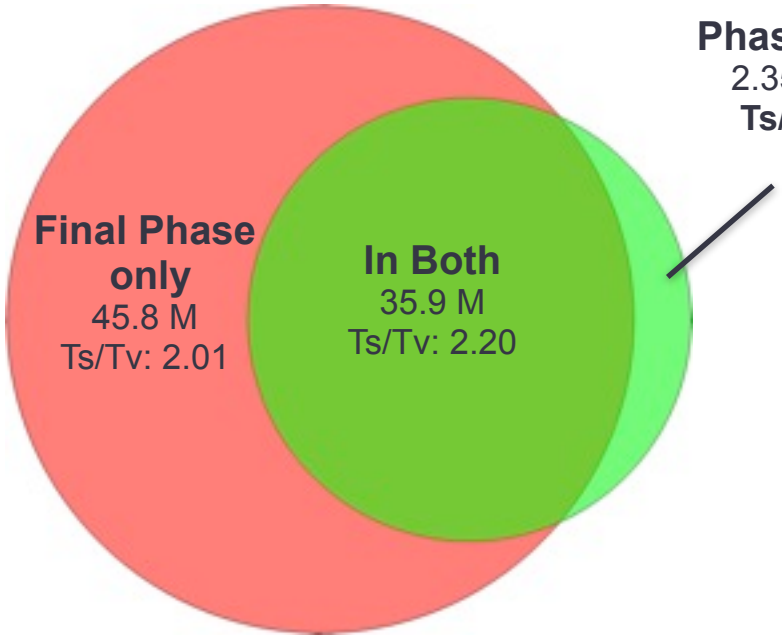
Variant Call Format

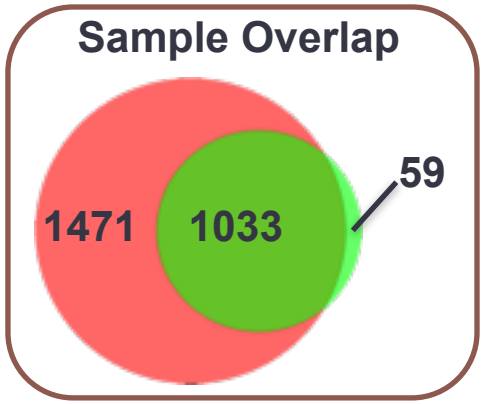
- Integrated haplotypes released in Variant Call Format (VCF)

```
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=20140730
##reference=ftp://ftp.1000genomes.ebi.ac.uk//vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
##source=1000GenomesPhase3Pipeline
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
...
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1)">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=EAS_AF,Number=A,Type=Float,Description="Allele frequency in the EAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=EUR_AF,Number=A,Type=Float,Description="Allele frequency in the EUR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AFR_AF,Number=A,Type=Float,Description="Allele frequency in the AFR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AMR_AF,Number=A,Type=Float,Description="Allele frequency in the AMR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=SAS_AF,Number=A,Type=Float,Description="Allele frequency in the SAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele. Format: AA|REF|ALT|IndelType. AA: Ancestral allele, REF:Reference Allele, ALT:Alternate Allele, IndelType:Type of Indel (REF, ALT and
IndelType are only defined for indels)">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096 HG00097 HG00099
20      60343   .       G       A       100     PASS    AC=1;AF=0.000199681;AN=5008;NS=2504;DP=20377;EAS_AF=0;AMR_AF=0.0014;AFR_AF=0;EUR_AF=0;SAS_AF=0;AA=.      GT      0|0     0|0     0|0
20      60419   .       A       G       100     PASS    AC=1;AF=0.000199681;AN=5008;NS=2504;DP=19865;EAS_AF=0;AMR_AF=0;AFR_AF=0;EUR_AF=0;SAS_AF=0.001;AA=.      GT      0|0     0|0     0|0
20      60479   rs149529999     C       T       100     PASS    AC=17;AF=0.00339457;AN=5008;NS=2504;DP=20218;EAS_AF=0;AMR_AF=0.0043;AFR_AF=0.0106;EUR_AF=0;SAS_AF=0;AA=.        GT      0|0     0|0     0|0
20      60522   rs150241001     T       TC      100     PASS    AC=68;AF=0.0135783;AN=5008;NS=2504;DP=20754;EAS_AF=0;AMR_AF=0.0029;AFR_AF=0.0499;EUR_AF=0;SAS_AF=0;AA=unknown(NO_COVERAGE)      GT      0|0
0|0     0|0
20      60568   .       A       C       100     PASS    AC=1;AF=0.000199681;AN=5008;NS=2504;DP=20728;EAS_AF=0;AMR_AF=0;AFR_AF=0.0008;EUR_AF=0;SAS_AF=0;AA=.      GT      0|0     0|0     0|0
20      60571   rs116145529     C       A       100     PASS    AC=10;AF=0.00199681;AN=5008;NS=2504;DP=20683;EAS_AF=0;AMR_AF=0.0014;AFR_AF=0.0068;EUR_AF=0;SAS_AF=0;AA=.        GT      0|0     0|0     0|0
20      60579   .       G       A       100     PASS    AC=1;AF=0.000199681;AN=5008;NS=2504;DP=20396;EAS_AF=0.001;AMR_AF=0;AFR_AF=0;EUR_AF=0;SAS_AF=0;AA=.      GT      0|0     0|0     0|0
20      60649   .       A       G       100     PASS    AC=1;AF=0.000199681;AN=5008;NS=2504;DP=20484;EAS_AF=0;AMR_AF=0.0014;AFR_AF=0;EUR_AF=0;SAS_AF=0;AA=.      GT      0|0     0|0     0|0
20      60778   .       A       G       100     PASS    AC=1;AF=0.000199681;AN=5008;NS=2504;DP=21261;EAS_AF=0.001;AMR_AF=0;AFR_AF=0;EUR_AF=0;SAS_AF=0;AA=.      GT      0|0     0|0     0|0
20      60795   rs184056664     G       C       100     PASS    AC=1;AF=0.000199681;AN=5008;NS=2504;DP=21333;EAS_AF=0;AMR_AF=0;AFR_AF=0;EUR_AF=0.001;SAS_AF=0;AA=.      GT      0|0     0|0     0|0
20      60808   .       G       A       100     PASS    AC=1;AF=0.000199681;AN=5008;NS=2504;DP=21348;EAS_AF=0;AMR_AF=0;AFR_AF=0.0008;EUR_AF=0;SAS_AF=0;AA=.      GT      0|0     0|0     0|0
20      60810   .       G       GA      100     PASS    AC=4;AF=0.000798722;AN=5008;NS=2504;DP=21358;EAS_AF=0;AMR_AF=0.0058;AFR_AF=0;EUR_AF=0;SAS_AF=0;AA=unknown(NO_COVERAGE)  GT      0|0     0|0     0|0
20      60826   .       A       G       100     PASS    AC=1;AF=0.000199681;AN=5008;NS=2504;DP=21136;EAS_AF=0;AMR_AF=0;AFR_AF=0.0008;EUR_AF=0;SAS_AF=0;AA=.      GT      0|0     0|0     0|0
```

Many tools, including new and improved ones written by 1000G Consortium members: vcftools, bcftools/htslib, vcflib, vt, GATK

Autosomal variants

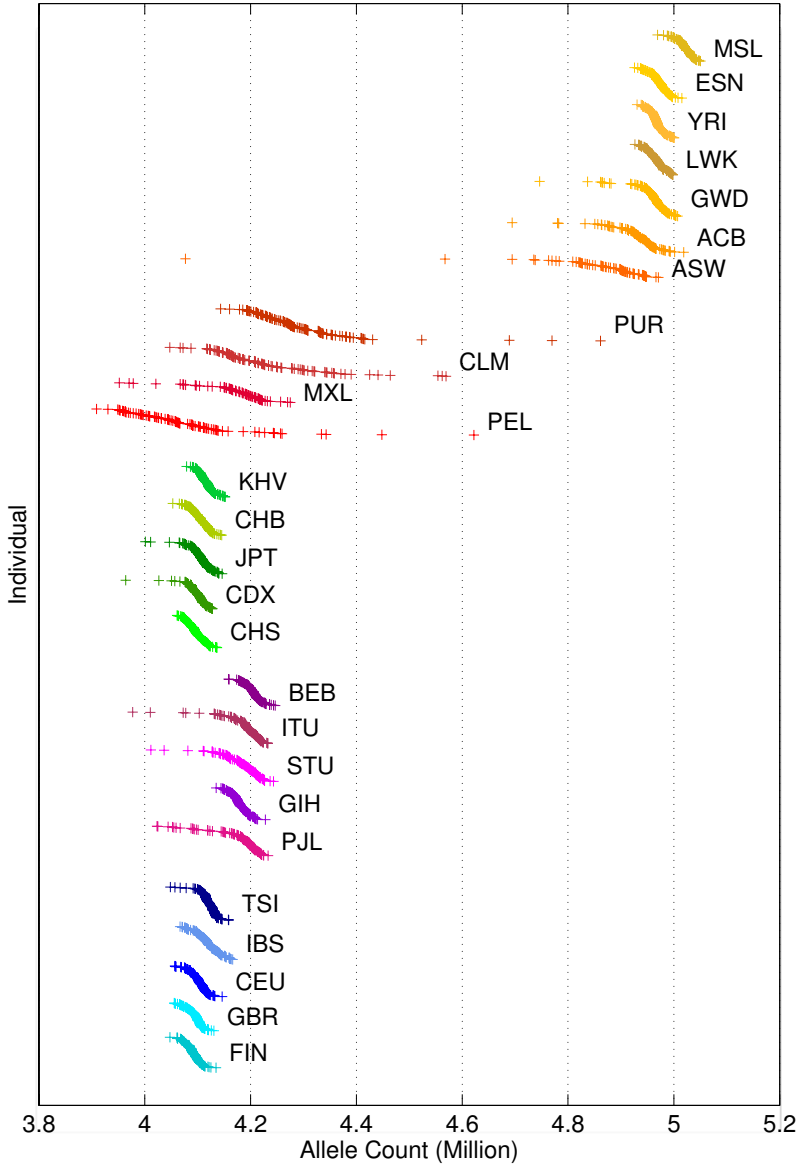




Type	Phase 1 Total	**Final Phase Total**
Biallelic SNPs	36.7 M	**77.8 M**
Multi-allelic SNPs	-	**259.4 k**
Biallelic indels	1.4 M	**3.0 M**
Multi-allelic indels	-	**154.9 k**
Mobile Element Insertions	-	**16.4 k**
Large Deletions	13.8 k	**32.4 k**
Copy Number Variants (CNVs)	-	**8.8 k**
Inversions	-	**100**
Total	38.1 M	**81.3 M**

Variants per genome

Type	Variant sites / genome
SNPs	3.78 million
Indels	573 thousand
Mobile Element Insertions	~1066
Large Deletions	~976
CNVs	~162
Inversions	~11



Phase 3 release

- The Phase 3 final release is available for download now.
- Autosome, chrX, chrY (soon), SNPs, INDELs, complex, SVs, STRs

  ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

  (or find the link at www.1000genomes.org)
- IMPUTE reference panel at

  https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference

Plenary Abstract, Tue. 8am, Hall B1
Goncalo Abecasis, *Completion of the 1000 Genomes Project: Results, lessons learned and open questions*