BLENDED GENOME EXOME

Capturing genomic diversity with a novel whole-exome plus low-pass whole genome product

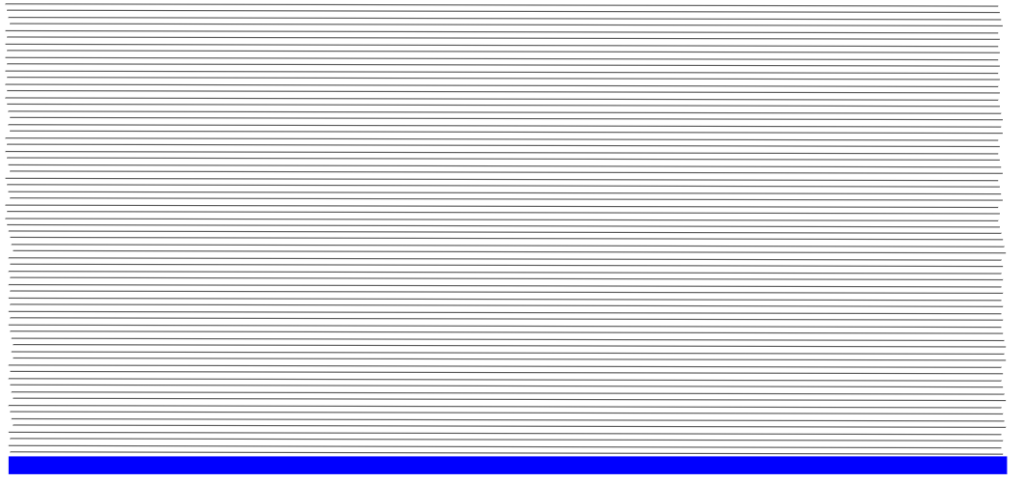
Daniel Howrigan, PhD
Data Group Leader, Neale Lab

BROAD INSTITUTE

STANLEY CENTER FOR PSYCHIATRIC RESEARCH

How do we effectively capture the diversity of the genome at scale?

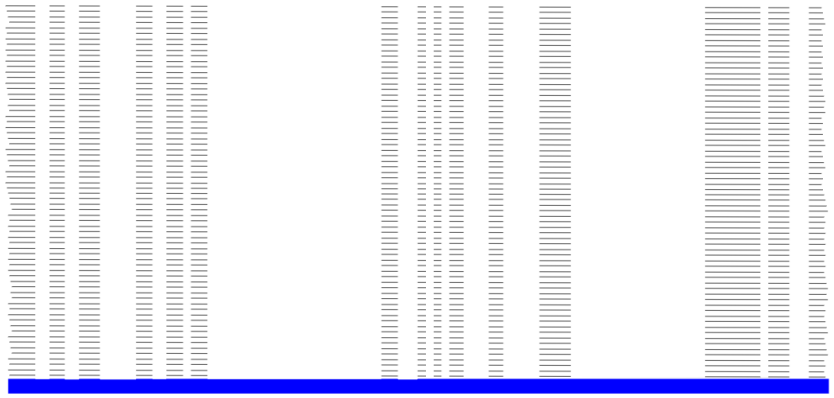
- **Option 1**: Deep whole genomes



- **Problem:** not yet cost-effective to sequence high-coverage whole genomes for large cohorts

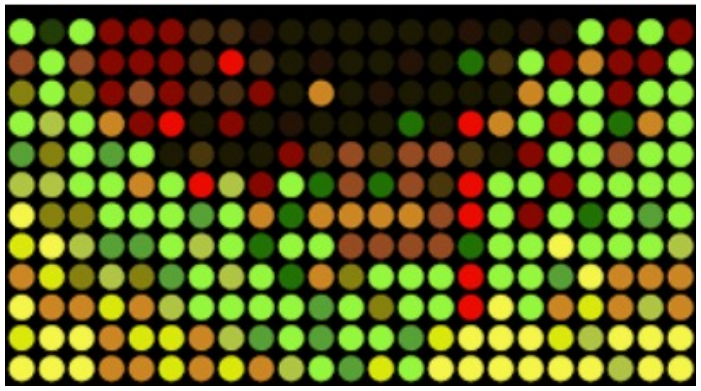
How do we effectively capture the diversity of the genome at scale?

- **Option 2**: Deep exome + GWAS array imputation



Rare coding variants

+



Common variant backbone for imputation

- **Problems**
  - All but the most expensive GWAS arrays biased towards SNPs discovered/common in European ancestries
  - Logistical challenges in harmonizing analyses from separate technologies

Blended Genome Exome (BGE) technology offers a new solution

Solutions:

- Unbiased common variant capture
- Exome and genome in the same sequence run
- Single CRAM/gVCF
- Cost effective alternative to deep WGS or exome + array

$150 per sample

*discounts at bulk sample size

The name? BGE won over:

- *GenEx Hybrid*
- *BEST capture*
- *Blendome*
- *Genxome*
- *Genome McExome Face*

The high-throughput technology behind BGE

Can run over 60 samples through a single lane of sequencing!

Gory details:

- Enzymatic fragmentation (NEBNext Ultra II FS kit)
  - NEB – New England Biosciences
- Quarter reaction volumes
- 384 sample batches (have 192 indexed adapters now)
- 384 well SPRI cleanups
  - SPRI – Solid Phase Reversible Immobilization
- Multiple additions of sample + bead to magnet
- Reduced cost exome capture
- Tempest for fast non-contact dispense destination normalization (384 in minutes!)

Lessons learned from Covid Dx and Covid Seq!

Low pass imputation using GLIMPSE software

GLIMPSE Home Overview Installation Documentation Benchmark GitHub

glimpse

Genotype Likelihoods IMputation and PhaSing mEthod

CATCH A GLIMPSE OF YOUR LOW DEPTH SEQUENCING DATA

GLIMPSE is a phasing and imputation method for large-scale low-coverage sequencing studies.

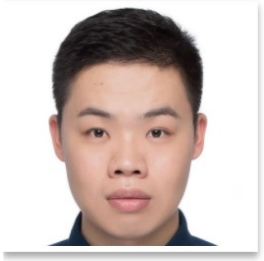
Main features of the method:

1. **Accurate imputed genotype calls.** Our method takes advantage of reference panels to produce high quality genotype calls.
2. **Accurate phasing.** GLIMPSE outputs accurate phased haplotypes for the low-coverage sequenced dataset.
3. **Low-coverage sequencing outperforms SNP arrays.** Imputation using low-coverage sequencing data is competitive to SNP array imputation. Results for European and African-American populations are interactively available on the website.
4. **A cost-effective paradigm.** GLIMPSE realises whole genome imputation from the HRC reference panel for less than 1$.

GLIMPSE tools is available under the MIT licence on the Github repository https://github.com/odelaneau/GLIMPSE.

HUGE thanks to Kai Yuan for being our GLIMPSE workflow expert

THE HUANG LAB



Kai Yuan

Research Fellow

Email: kyuan@broadinstitute.org

Kai Yuan is a postdoctoral fellow in the Massachusetts General Hospital and the Broad Institute, advised by Dr. Hailiang Huang. He obtained his PhD in computational biology from the Partner Institute for Computational Biology, Chinese Academy of Sciences. During his PhD training, he worked on population genetics, especially for the admixed populations and developed several methods to infer population

Low pass imputation using GLIMPSE software

"Variable position" = SNP in reference panel, not in sequence data

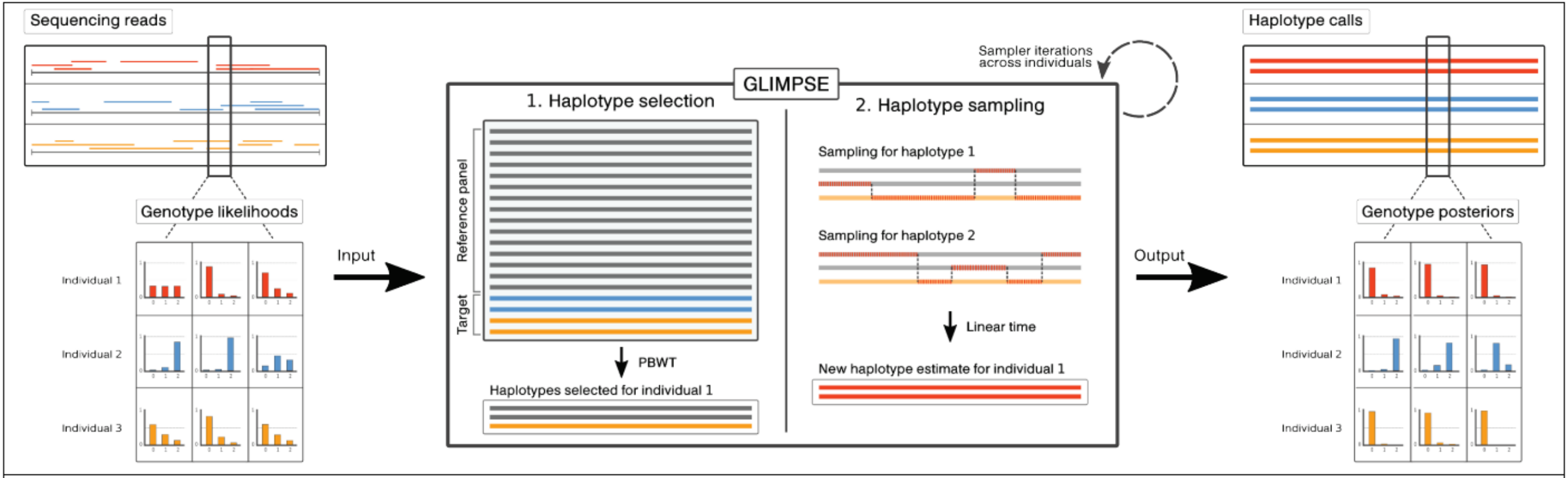
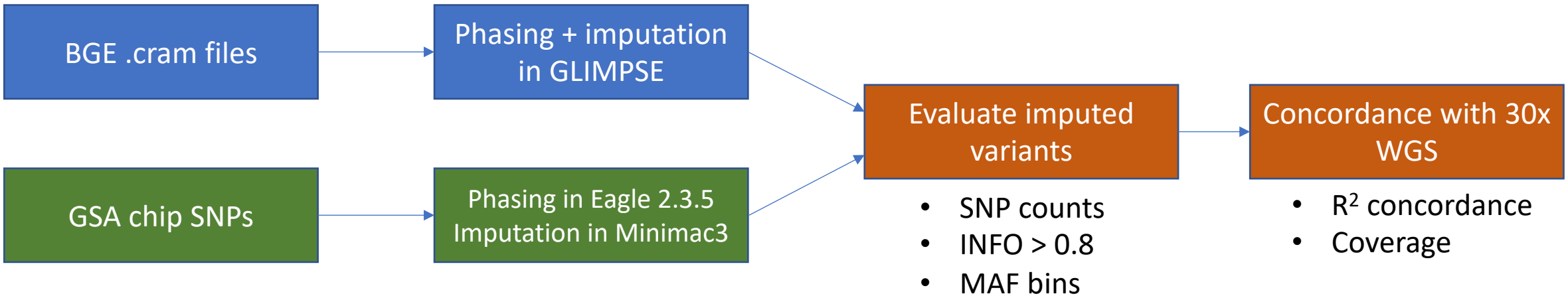


Figure1: GLIMPSE method overview. The input of the method is a matrix of genotype likelihoods defined at all variable positions obtained directly from the sequencing reads (left). GLIMPSE refines the genotype likelihoods using a Gibbs sampler scheme. At each iteration a new pair of haplotypes for each individual is estimated (middle). This involves two main steps: (1.) the haplotype selection using a reference panel and the current estimate of all other target haplotypes (middle, left) and (2.) a linear time sampling algorithm based on the Li and Stephens model (middle, right). As an output, GLIMPSE produces consensus-based haplotype calls and genotype posteriors at every variable position (right).

Testing early products using deep whole genomes

- Using high quality calls from 30x genomes as our “truth” dataset
  - Compare low-pass GLIMPSE imputation against Global Screening Array (GSA) chip

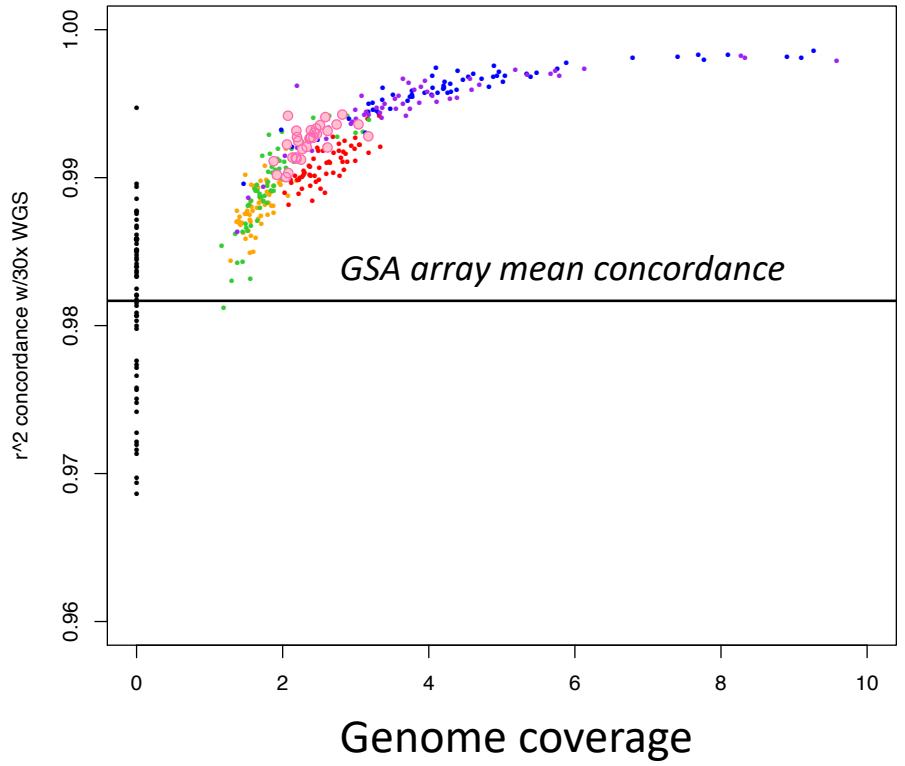
- Pilot sample sets
  - Early rounds: 31 to 62 Hispanic samples
  - Later rounds: 23 African samples from PUMAS (Ethiopia and South Africa)



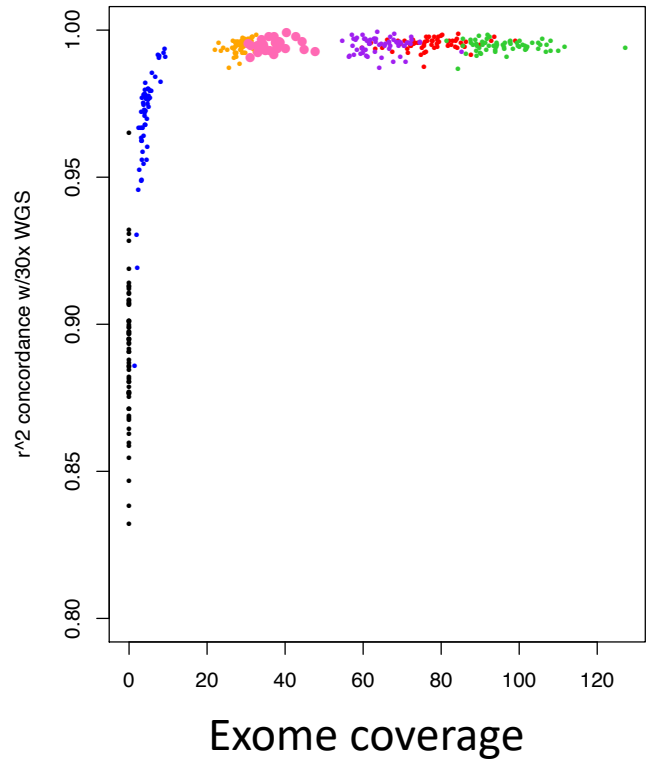
Finding the optimal blend of coverage

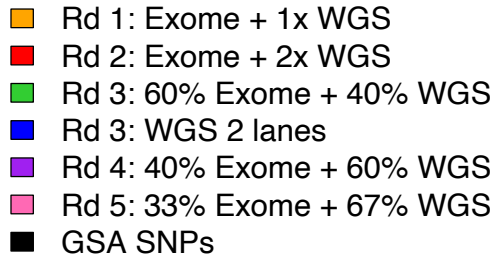
- 31-62 Hispanic samples
- Blood derived DNA
- HRC imputation
- GRCh37/hg19

Genome concordance



Exome concordance

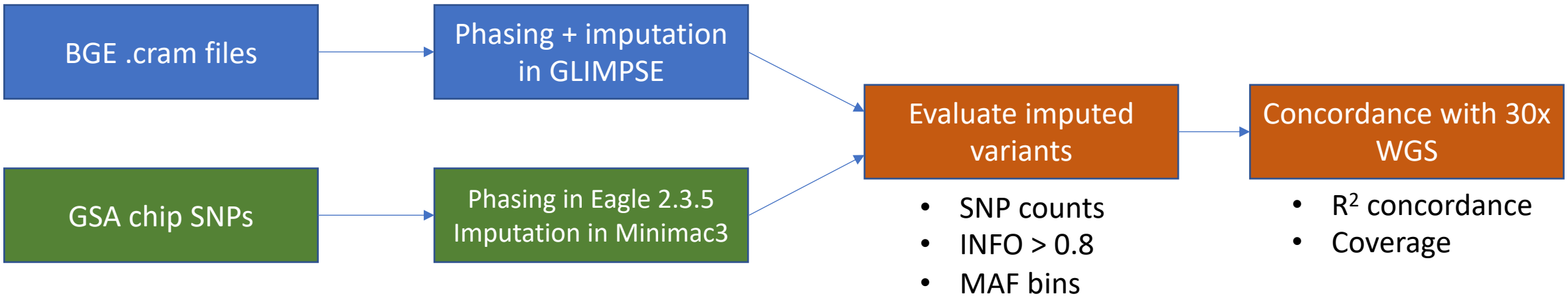




Results from the BGE pilot dataset

- 764 participants
  - 64 NeuroGAP South Africa (UCT)
  - 317 NeuroGAP Ethiopian (AAU)
  - 381 China (BioX)

- 23 participants also have deep WGS (30x coverage) for concordance comparison
  - 13 NeuroGAP South Africa (UCT)
  - 10 NeuroGAP Ethiopian (AAU)



GLIMPSE runtime is robust to sample size

- 764 pilot samples
- All calls

Full Sample run time
Jobs capped at 8GB max RAM

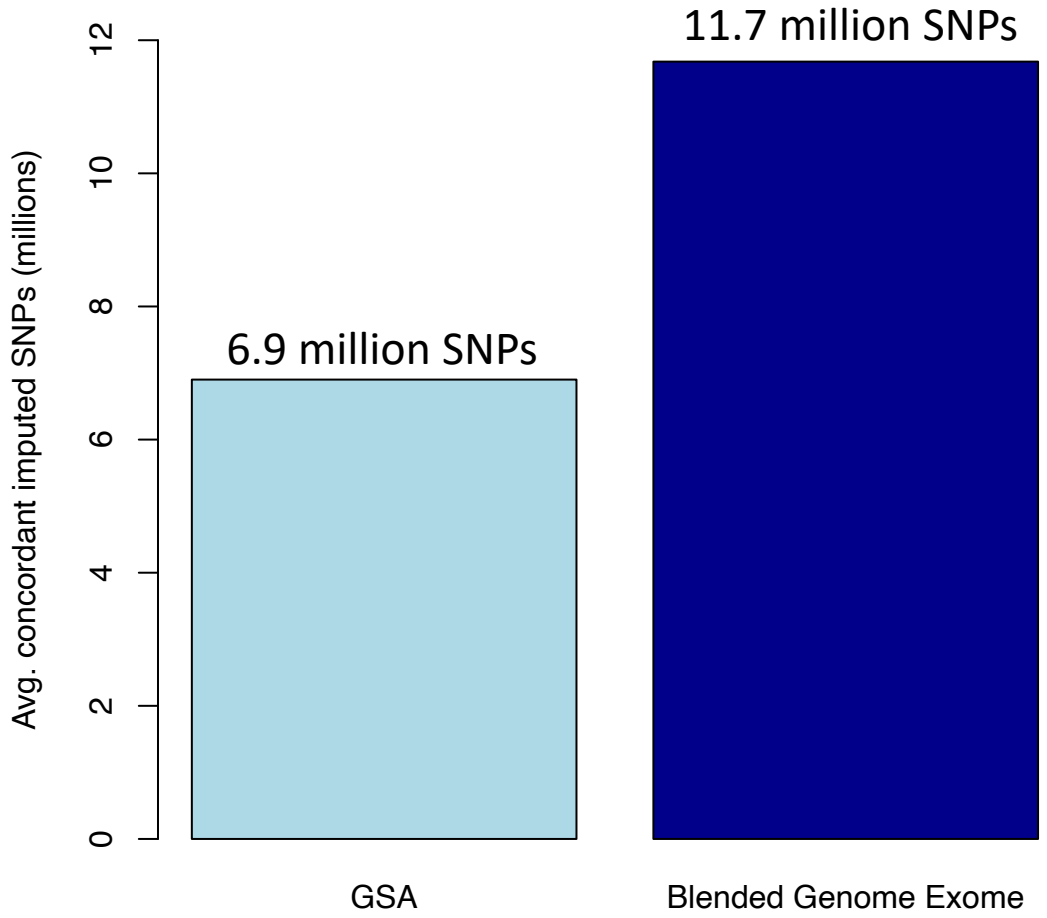
GLIMPSE step	Jobs submitted	Run Time (CPU hrs)
cram2GL	16808	**1370.36**
VcfCombine	22	184.75
GetSiteInfo	22	3.20
GenomeChunked	22	0.06
ChunkImpute	952	**8732.44**
ChunkLigate	22	7.40
Phase	22	2.01

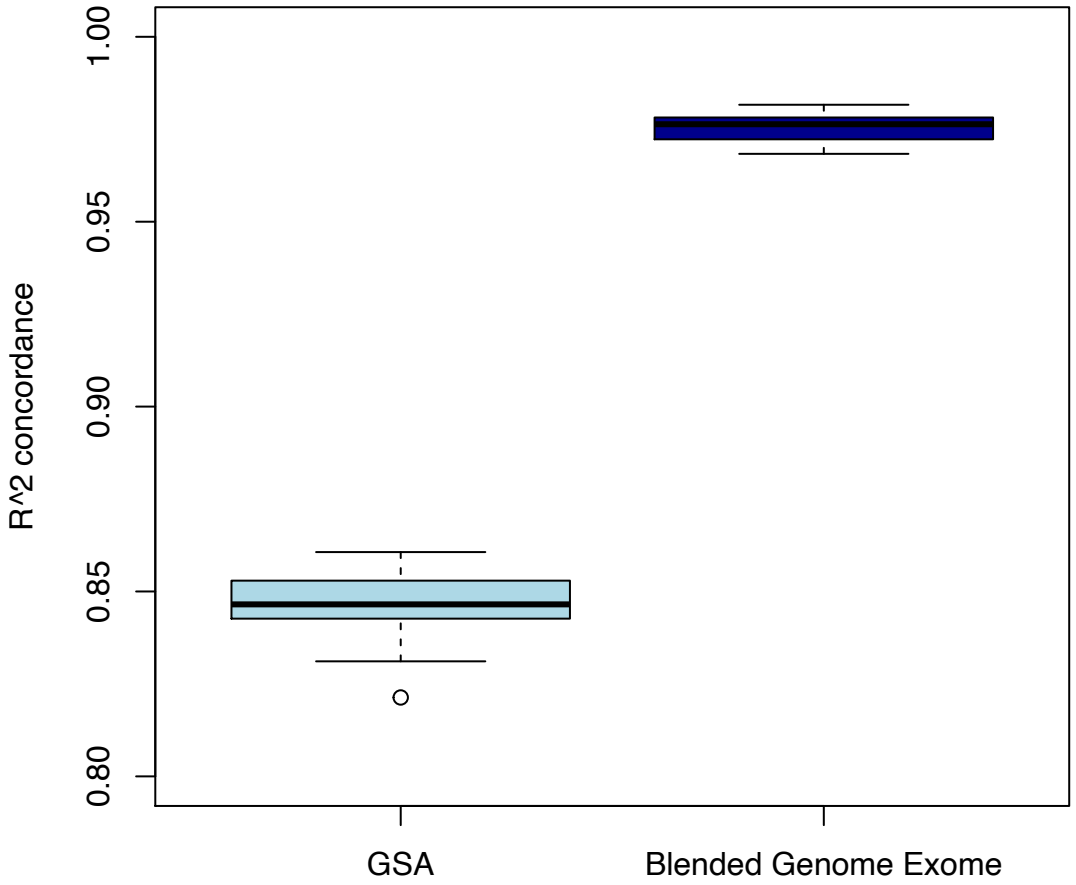
Cohort-specific run time
Jobs capped at 4GB max RAM

Location	Sample size	Run Time per sample
South Africa	61	**12.5**
Ethiopia	317	**12.1**
East Asia	381	**11.3**

BGE has more SNPs and higher concordance than current GSA platform

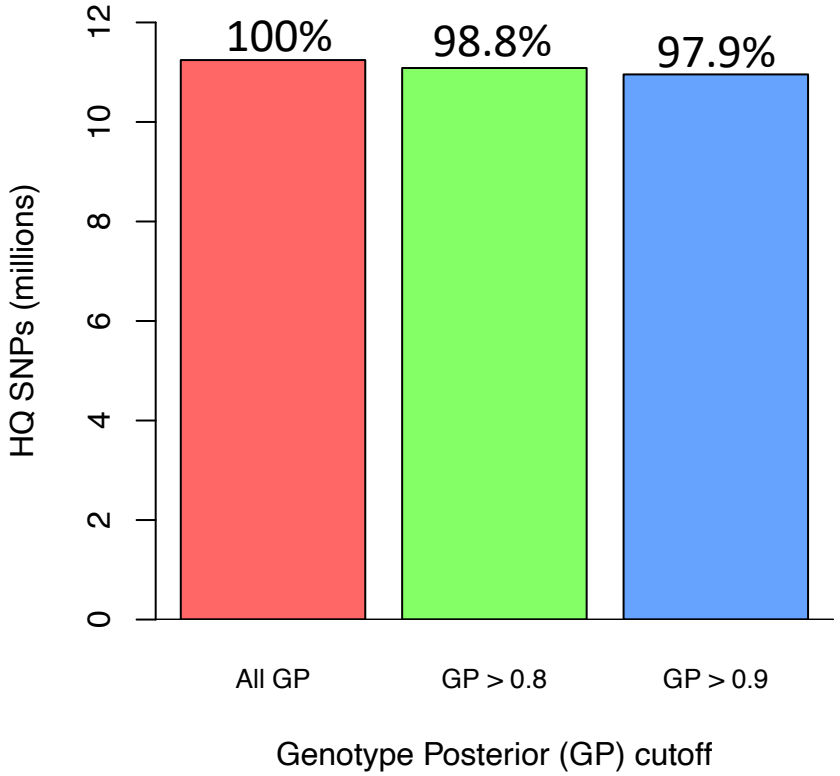
- 23 African samples
- Saliva derived DNA
- HRC imputation
- GRCh38/hg38

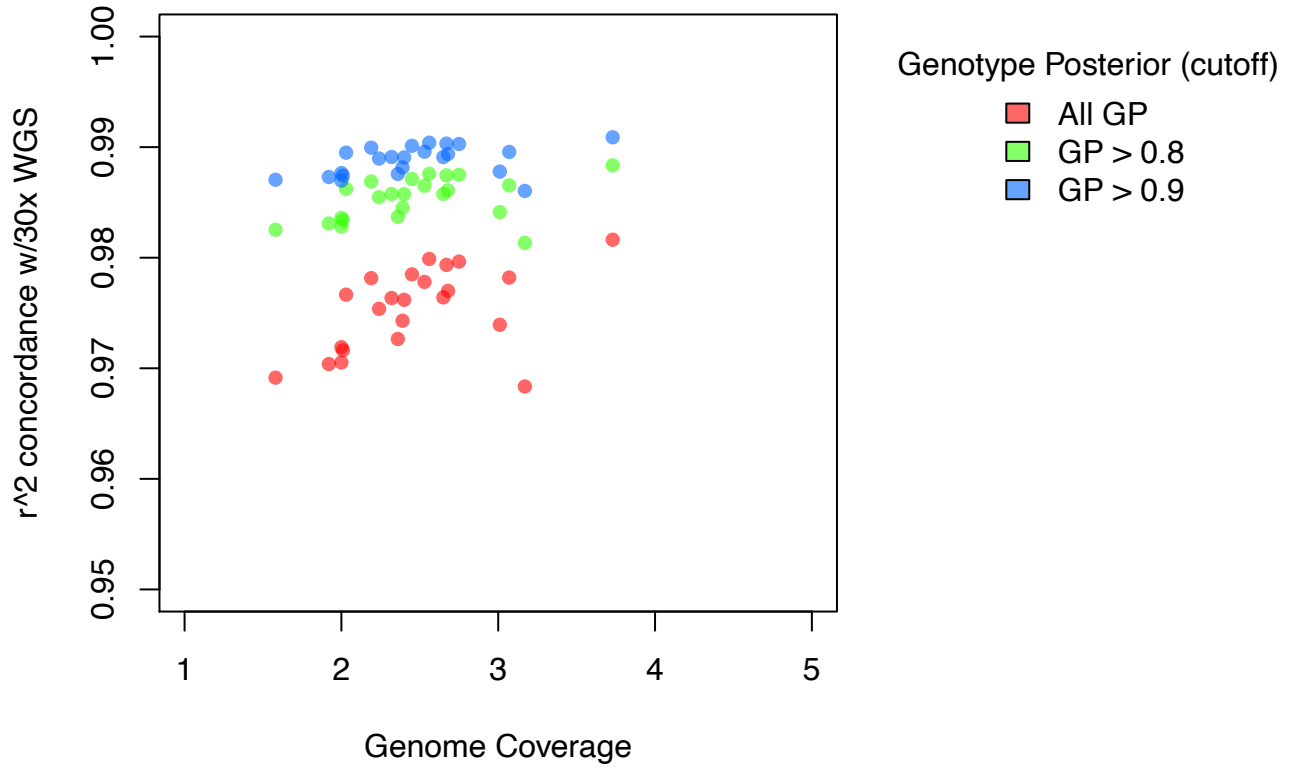




Restricting to higher genotype posterior cutoffs improves concordance with 30x genomes

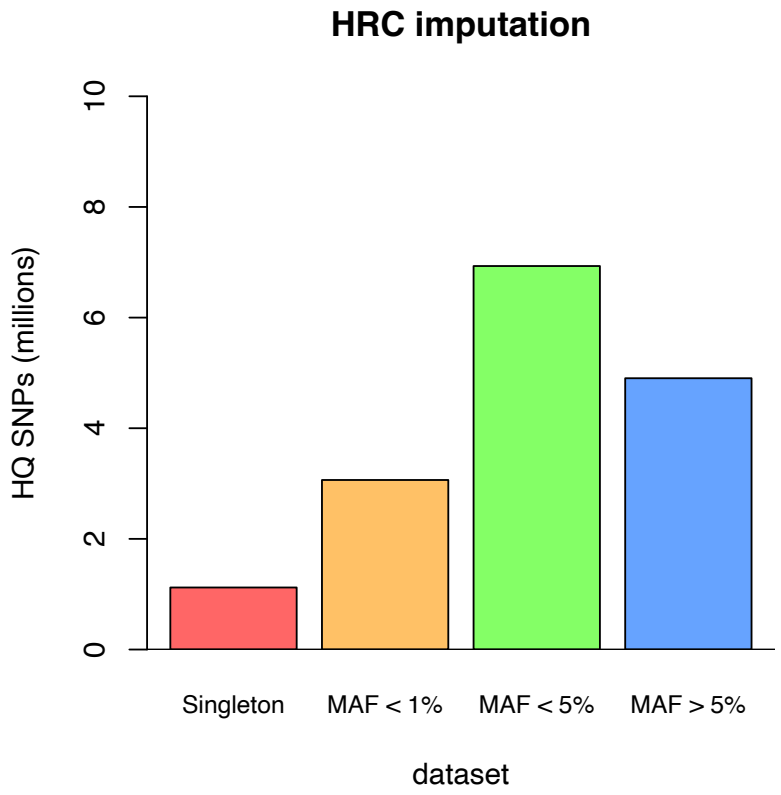
- 23 African samples
- Saliva derived DNA
- HRC imputation
- GRCh38/hg38

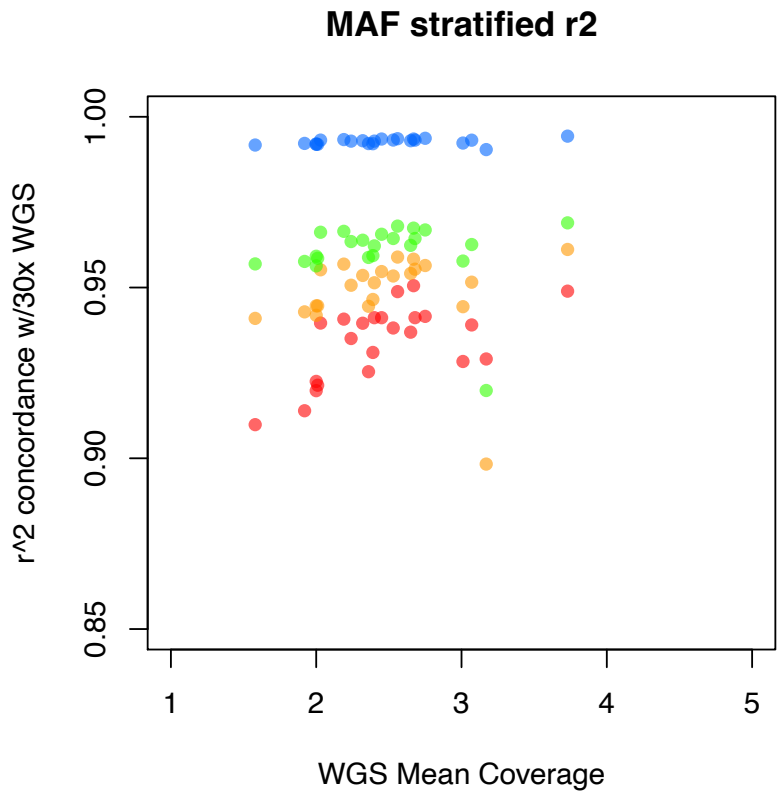




Are we capturing lower frequency variants well?

- 23 African samples
- Saliva derived DNA
- GP > 0.9 calls





Singleton
MAF < 1%
MAF < 5%
MAF > 5%

MAF estimates from 371 NeuroGAP African samples

BGE Samples currently done/underway

Project Cohort	#Samples	Primary Disease	Sequencing status
PUMAS/Paisa, Colombia	9,008	BP/SCZ/Ctrls	complete
NeuroMex, Mexico	2,924	BP/SCZ/Ctrls	complete
QIMR, Australia	7,636	BP/Ctrls	complete
Kenya Psychosis	1,849	Psychosis	complete
Taiwan Bipolar	1,094	BP/Ctrls	underway
BioX, China	18,000	SCZ/Ctrls	underway
PUMAS/NeuroGAP/Ethiopia	6,000	BP/SCZ/Ctrls	underway
PUMAS/NeuroGAP/KEMRI_Kenya	3,118	BP/SCZ/Ctrls	underway
PUMAS/NeuroGAP/Moi_Kenya	5,121	BP/SCZ/Ctrls	underway
PUMAS/NeuroGAP/Uganda	6,000	BP/SCZ/Ctrls	underway
PUMAS/NeuroGAP/SouthAfrica	6,000	BP/SCZ/Ctrls	underway
PUMAS/GCP	4,584	BP/SCZ/Ctrls	underway
GEN-SCRIP, Pakistan	8,231	SCZ/Ctrls	underway
IBD	128	IBD	underway
Totals	**79,565**		

BGE batched in the next 6-18 months

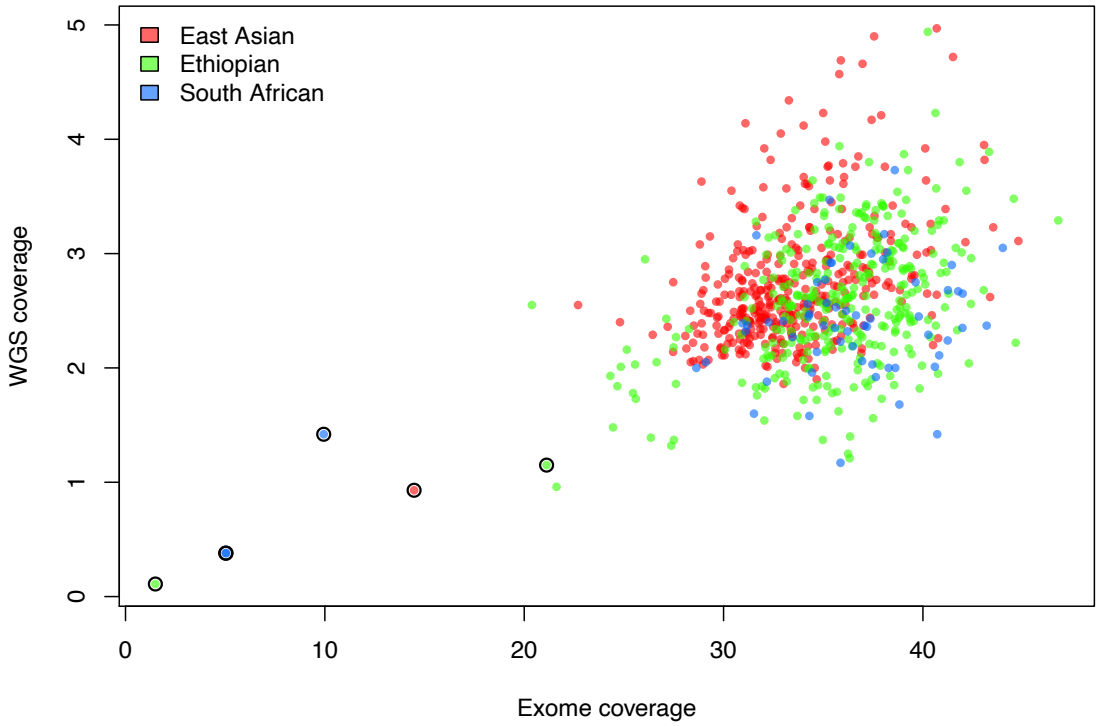
Project	#Samples	Primary Disease
PUMAS/NeuroGAP/Ethiopia	3,450	BP/SCZ/Ctrls
PUMAS/NeuroGAP/Uganda	3,229	BP/SCZ/Ctrls
PUMAS/NeuroGAP/SouthAfrica	1,613	BP/SCZ/Ctrls
PUMAS/Colombia	20,000	BP/SCZ/Ctrls
PUMAS/Brazil	2,000	BP/SCZ/Ctrls
PUMAS/GPC/PAARTENRS/MGS	1,600	BP/SCZ/Ctrls
PUMAS/GPC/Intrepid	600	BP/SCZ/Ctrls
BioX, China	12,000	SCZ/Ctrls
GEN-SCRIP, Pakistan	4,000	SCZ/Ctrls
GEN-BLIP, Pakistan	12,000	BP/Ctrls
NeuroMex, Mexico	3,000	BP/SCZ/Ctrls
Anorexia	20,000	AN/Ctrls
NDD	2,000	NDD
PUMAS/Mozambique	4,000	BP/SCZ/Ctrls
A-BIG-NET cohorts	5,000-10,000	BP/Ctrls
Totals	**89,492**	

Creating a QC pipeline for BGE datasets

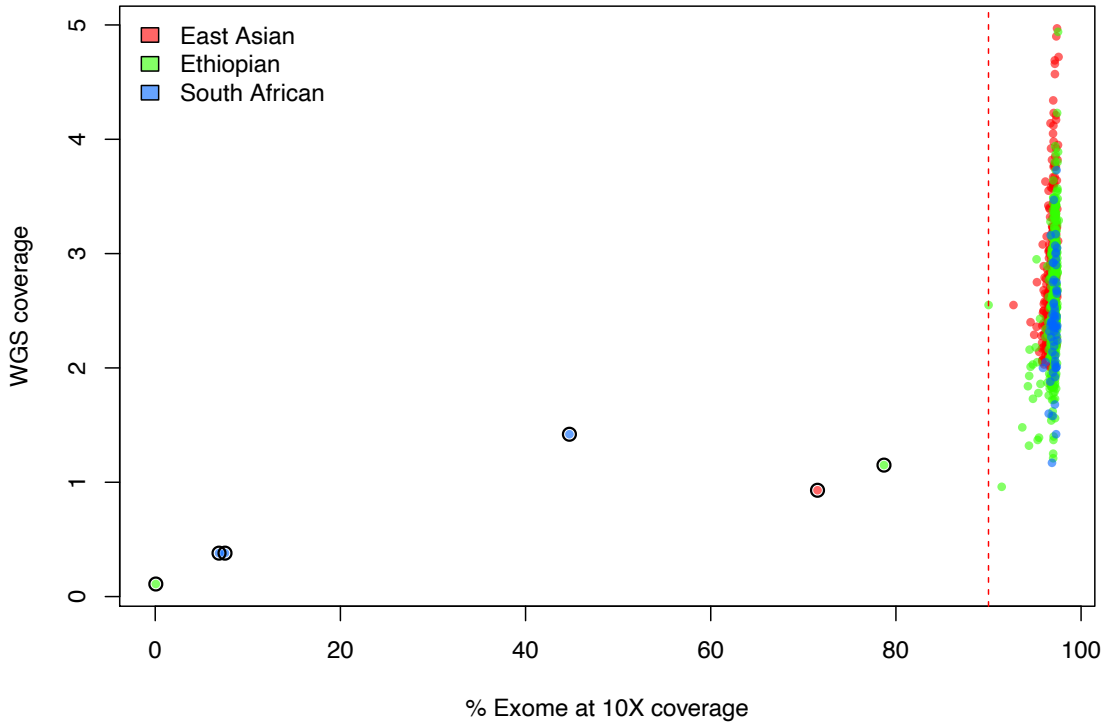
- Imputation happening prior to sample QC
  - SNP/sample QC used to be done first!
- Exome and imputation QC run in parallel
  - Where can one inform the other?
- Delivering data to collaborators
  - Imputation as a service
  - When should one re-impute these data?

Evaluating the full pilot cohort – sequence coverage

Mean coverage

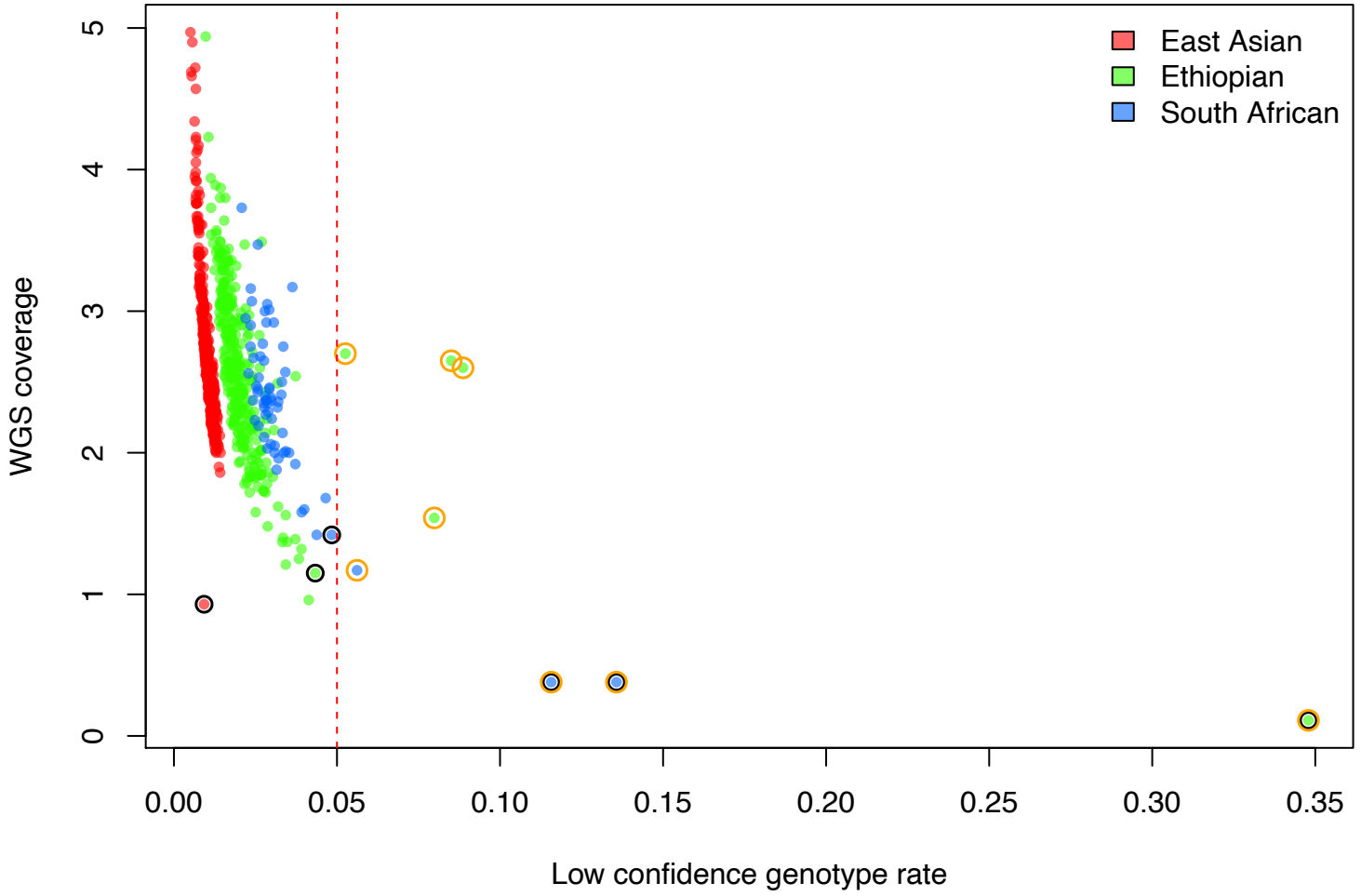


% of exome target reaching 10x coverage



6 samples in pilot data not meeting exome target coverage goals

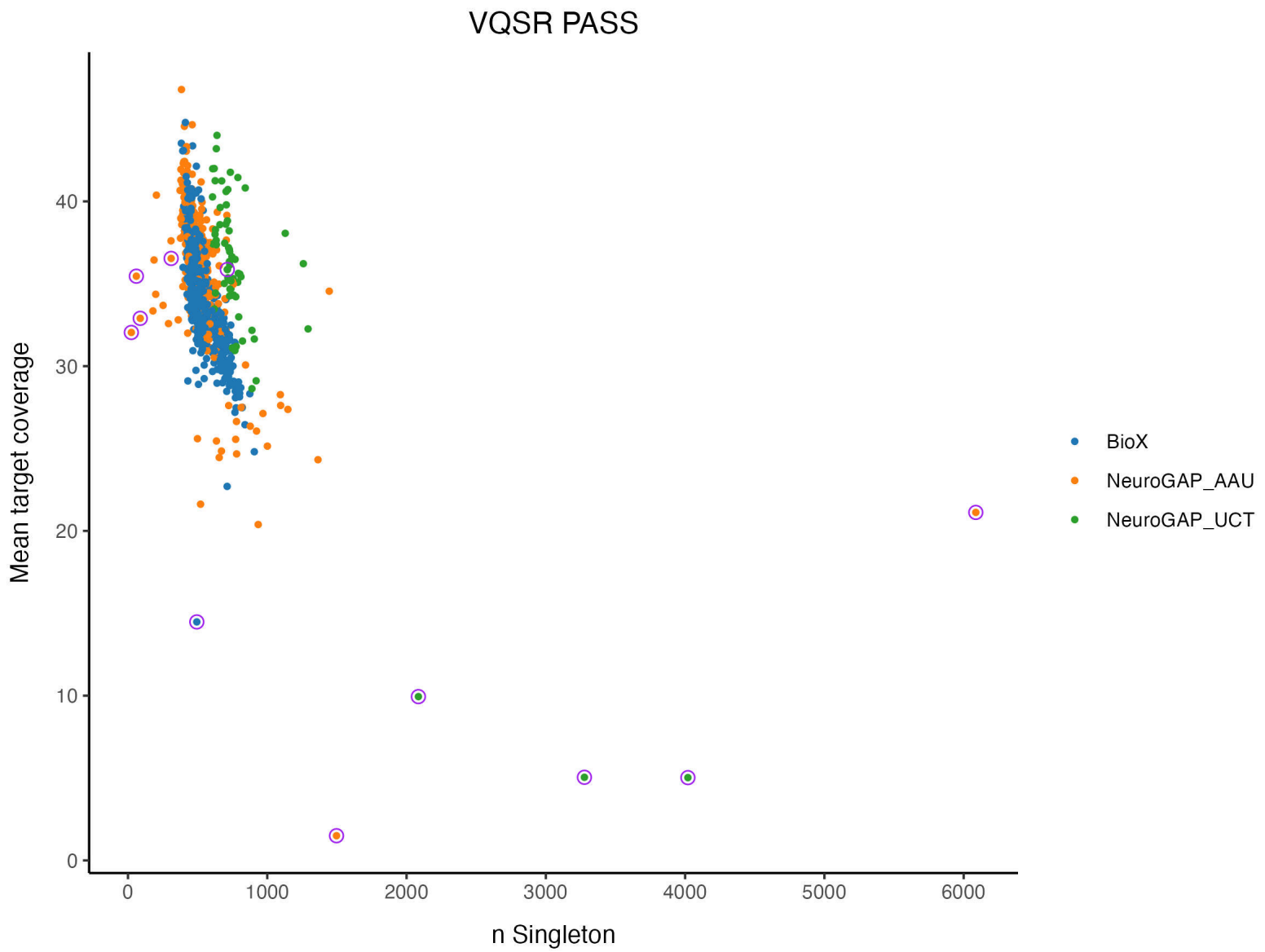
Lower coverage leads to lower high quality imputed calls



Low confidence genotypes
Genotype Posterior (GP) < 0.9

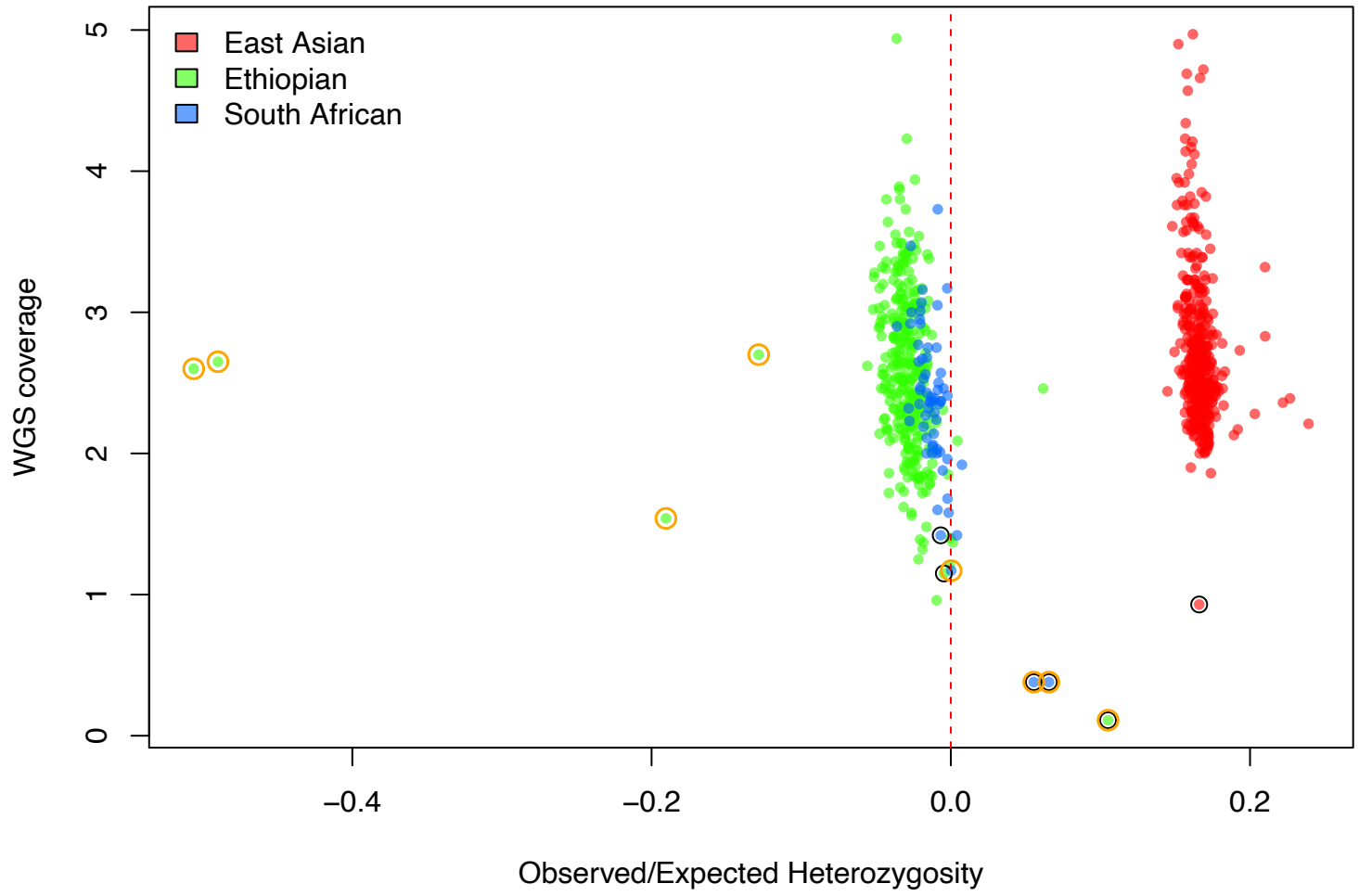
11 samples flagged by coverage and imputation quality (circled)

Singleton count in the **Whole Exome Sequence** shows similar pattern

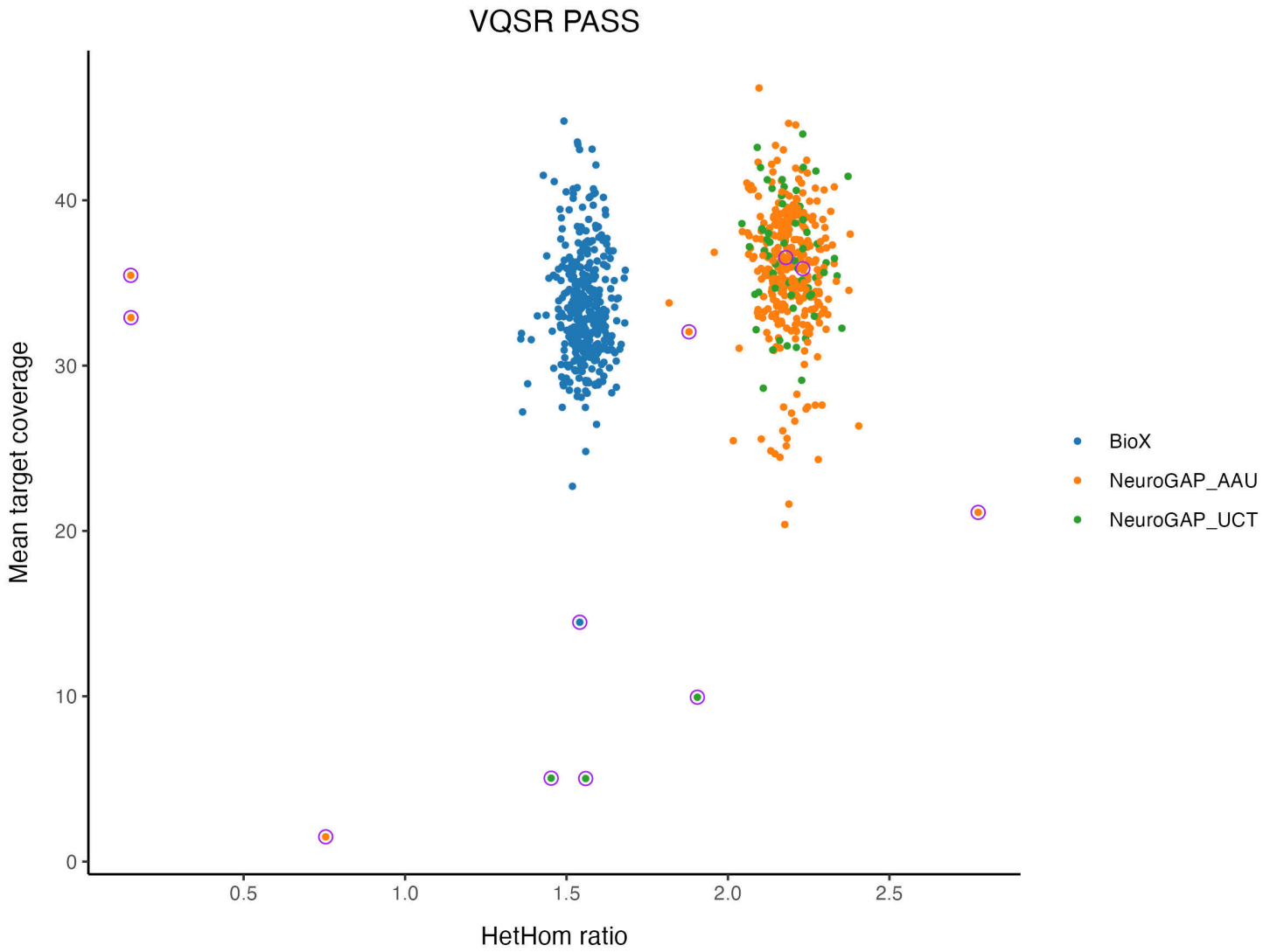


Exome results courtesy of Bob Ye

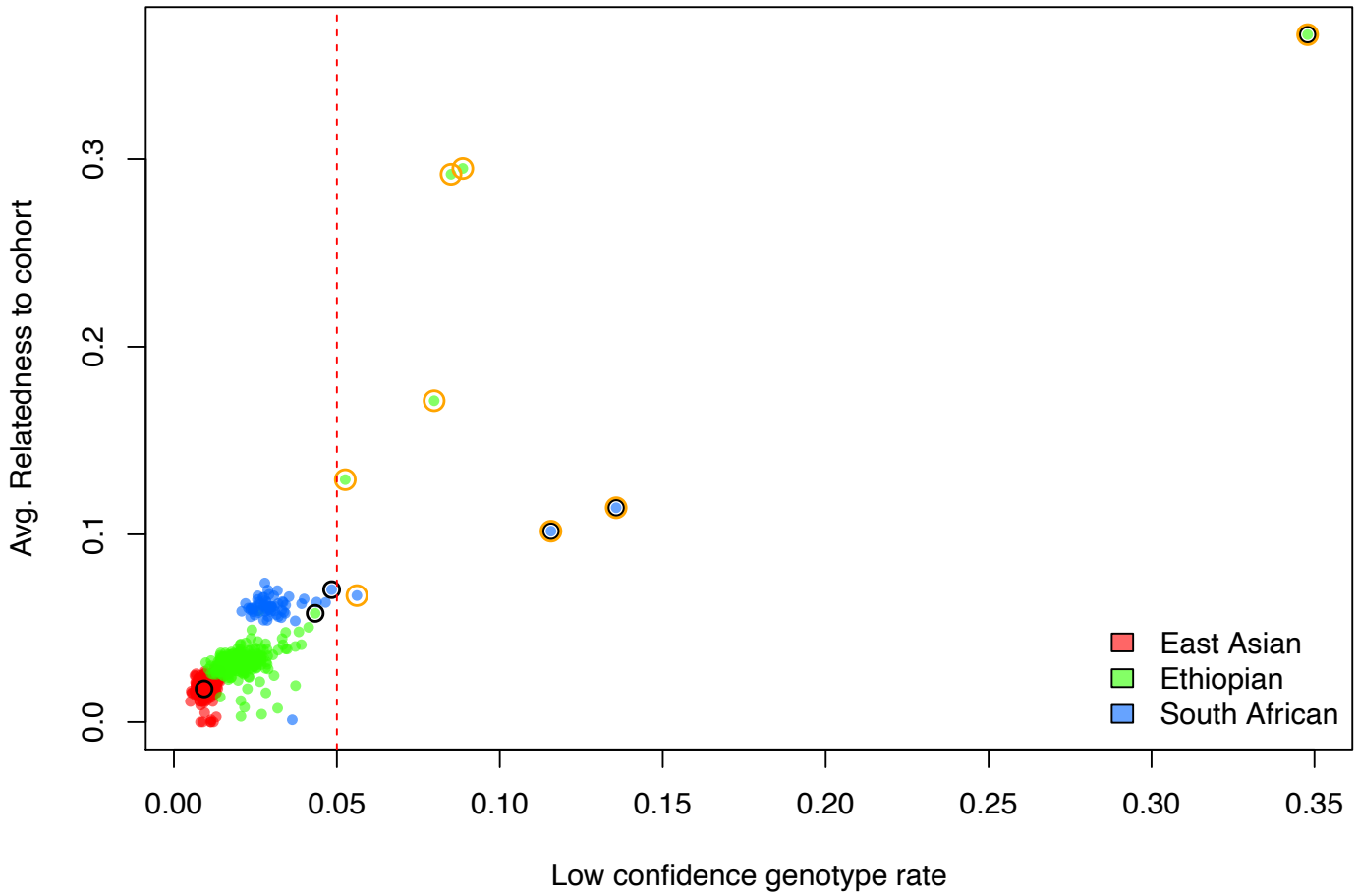
Heterozygosity levels as an indicator of sample quality



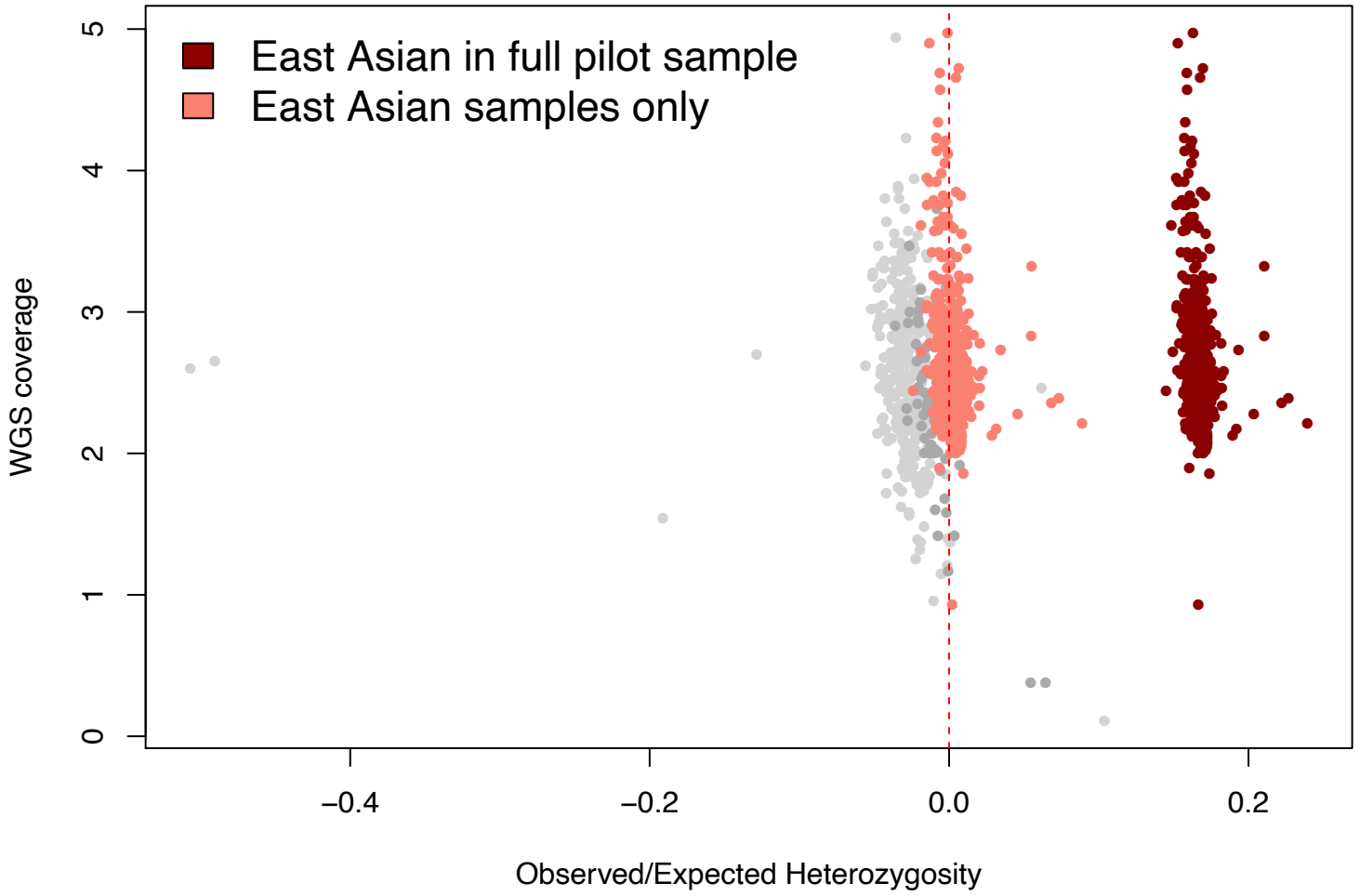
Het / hom ratio in the **Whole Exome Sequence** as an analogous metric



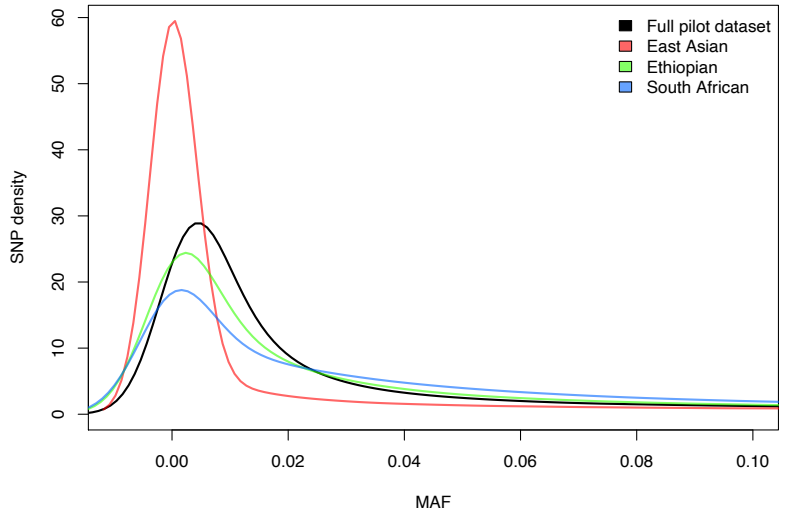
Relatedness checking confirms the presence of contaminated samples



Multi-ancestry cohorts affect how we QC



Allele frequency spectrum



Ongoing projects to evaluate BGE

- Transition of BGE sequencing to the Illumina NovaSeq X / DRAGEN pipeline
- Benchmarking for SNV calling
- BGE as an imputation reference panel
  - Many current reference panels are 4-8x WGS coverage

BLENDED GENOME EXOME

- Unbiased common variant capture
- Deep whole exome
- More SNPs and better accuracy than standard array
- Cost effective

BGE core dev team

Matthew Defelice
Jonna Grimsby
Brendan Blumenstiel

BGE core analysis/feedback

Sinead Chapman
Kai Yuan
Benjamin Neale
Hailiang Huang
Alicia Martin

ATGU

- Mark Daly
- Nik Baya
- Hail Team
- Raymond Walters
- TJ Singh
- Laura Gauthier
- ATGU/DSP group

Stanley Center

- Caroline Cusick
- Christine Stevens
- Sam Bryant
- Karesten Koenen
- Rocky Stroud
- Anne Stevenson
- NeuroGAP participants
- BioX participants

External collaborators

- Joseph Buxbaum
- Alexander Kolevzon
- Irva Hertz-Picciotto
- Margaret Pericak-Vance

Broad Genomics / Data Sciences Platform contributors

- Laurie Holmes
- Steven Ferriera
- Tera Bowers
- Michelle Cipicchio
- Greg Nakashian
- Matthew Lee
- Scott Anderson
- David Zdeb
- John Walsh
- Jon Thompson
- Samuel DeLuca
- Megan Giles
- Marissa Gildea
- Faye Reagan
- Jacquelyn Schneider
- Jessie Tang
- Erin LaRoche
- Andrew Bernier
- Jordan Callahan
- Matthew Coole
- Kimberly Sisley
- Mariela Mihaleva
- Tom Howd
- Nasko (Atanas) Mihalev
- Laurie Doe
- Justin Abreu
- Junko Tsuji
- Niall Lennon

Ti/Tv and Insertion/Deletion ratio as Exome QC metrics