Sequencing Strategies to Enhance the Next Generation of Genetic Evaluations

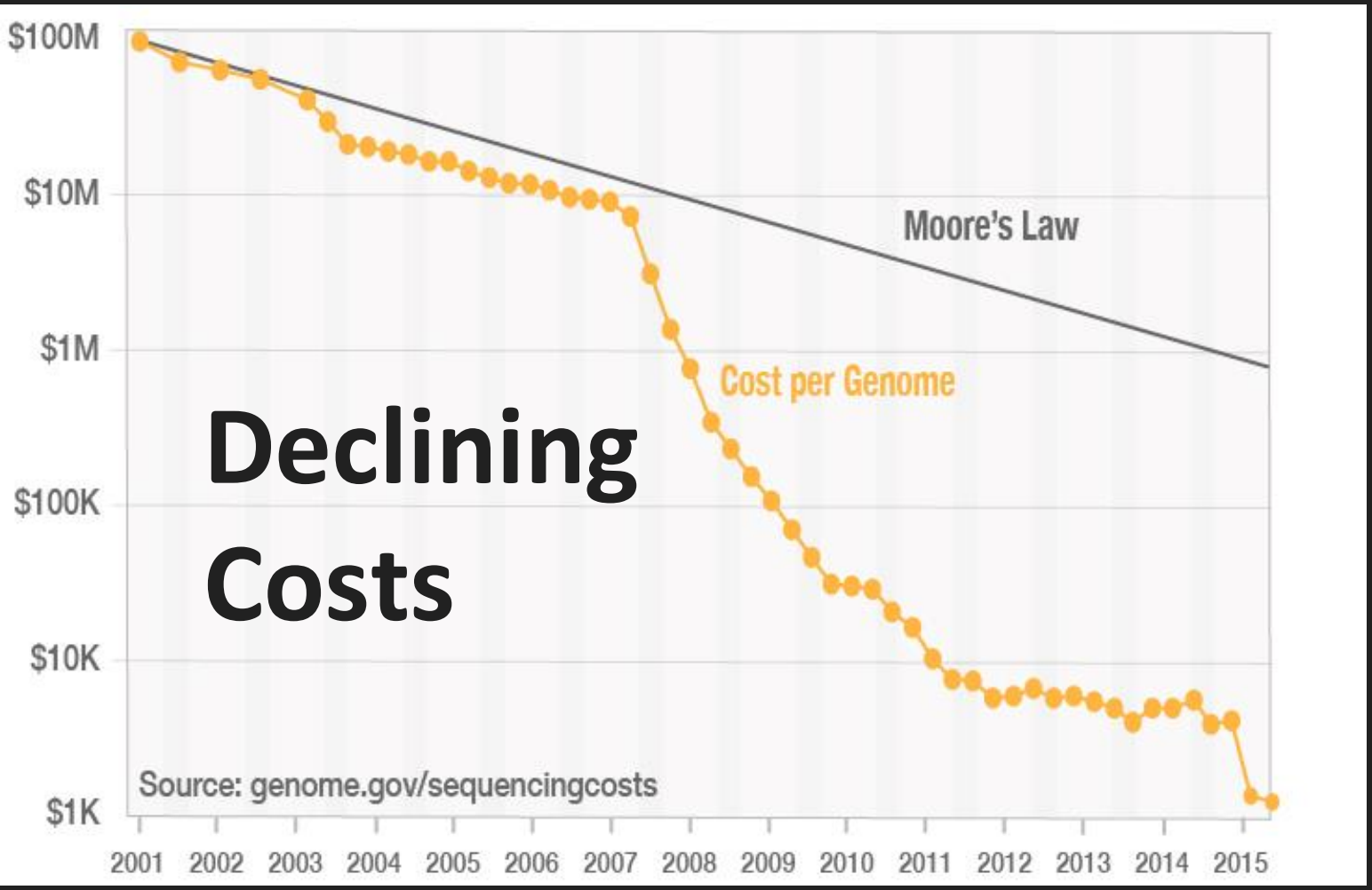
Troy Rowan

BIF Annual Symposium
Advances in Selection Decisions Breakout
July 4, 2023

UTIA INSTITUTE OF AGRICULTURE
THE UNIVERSITY OF TENNESSEE

MacG Jul '16

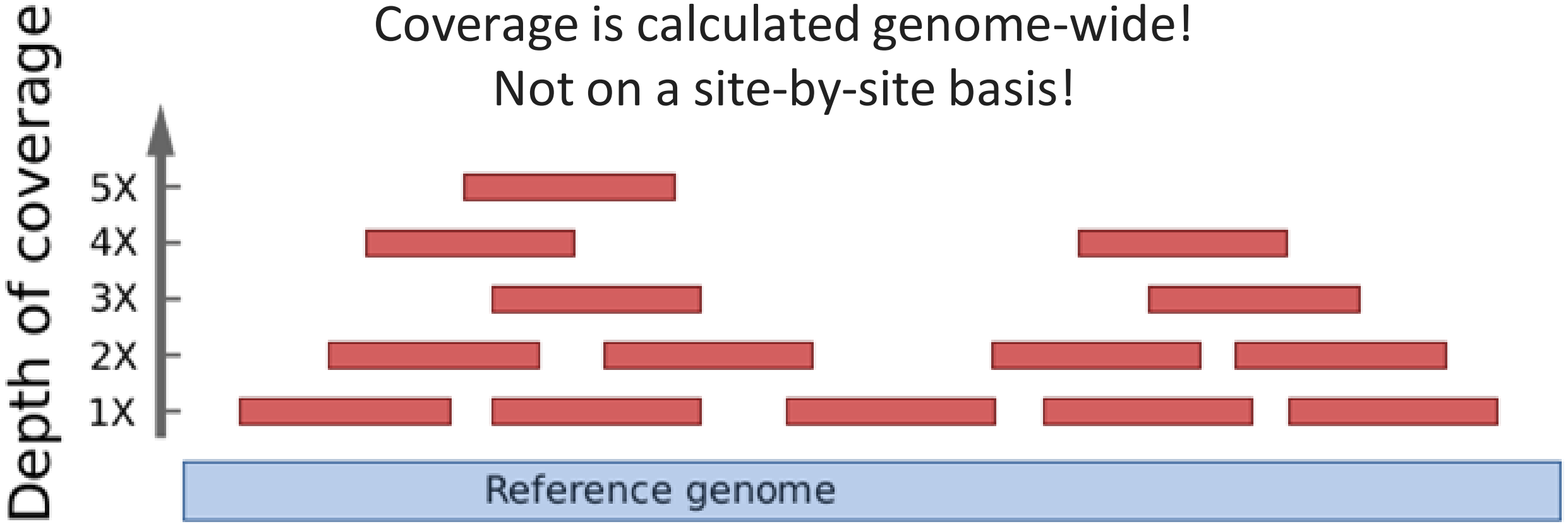
Our low-pass sequencing future



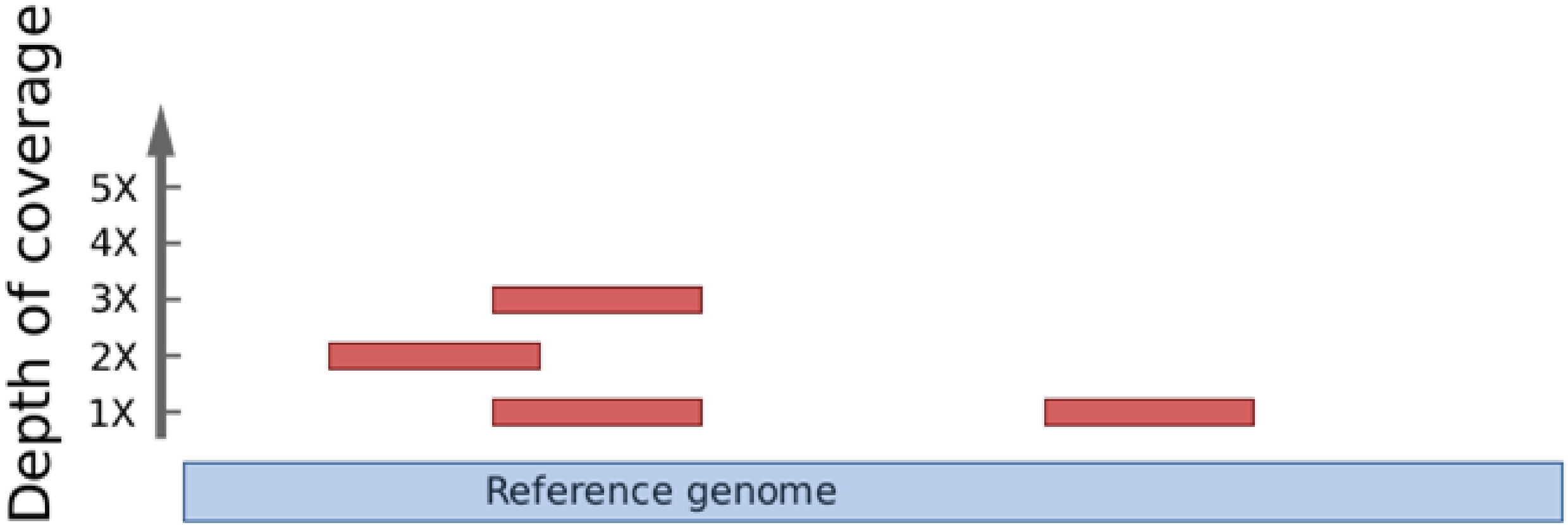
- Potential for further cost reduction
- Rare variation
- No need for chip redesign or updates
- SNP Discovery
- CNV detection

$$Coverage = \frac{nReads \times len(read)}{Genome \quad Size}$$

Coverage is calculated genome-wide!
Not on a site-by-site basis!

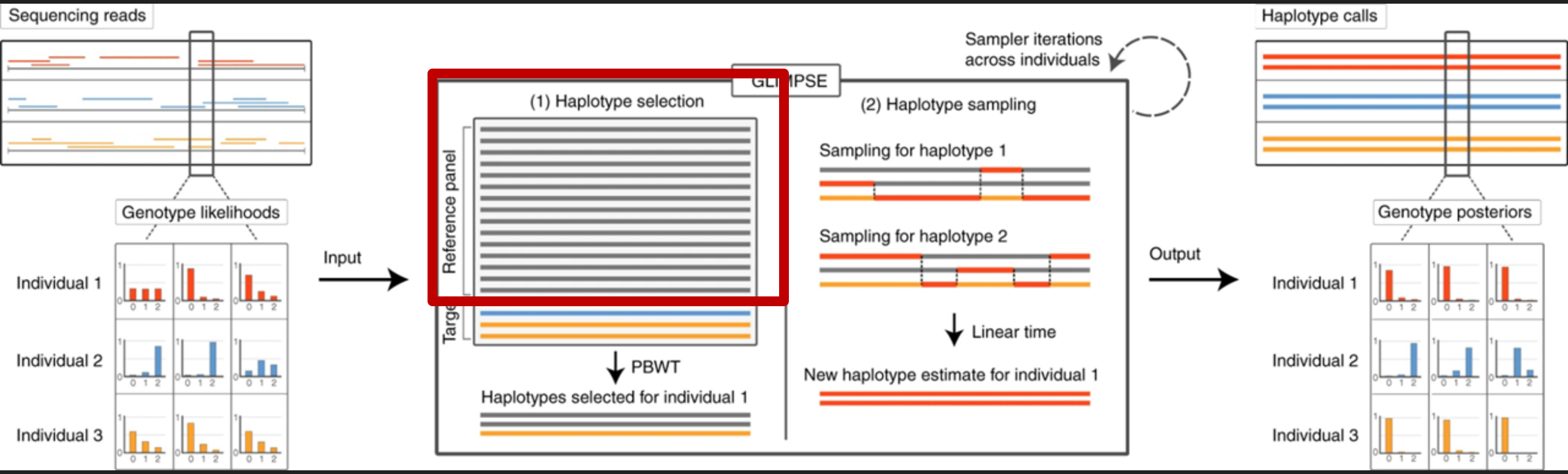


Low-Pass sequencing



www.metagenomics.wiki

Low-Pass Imputation



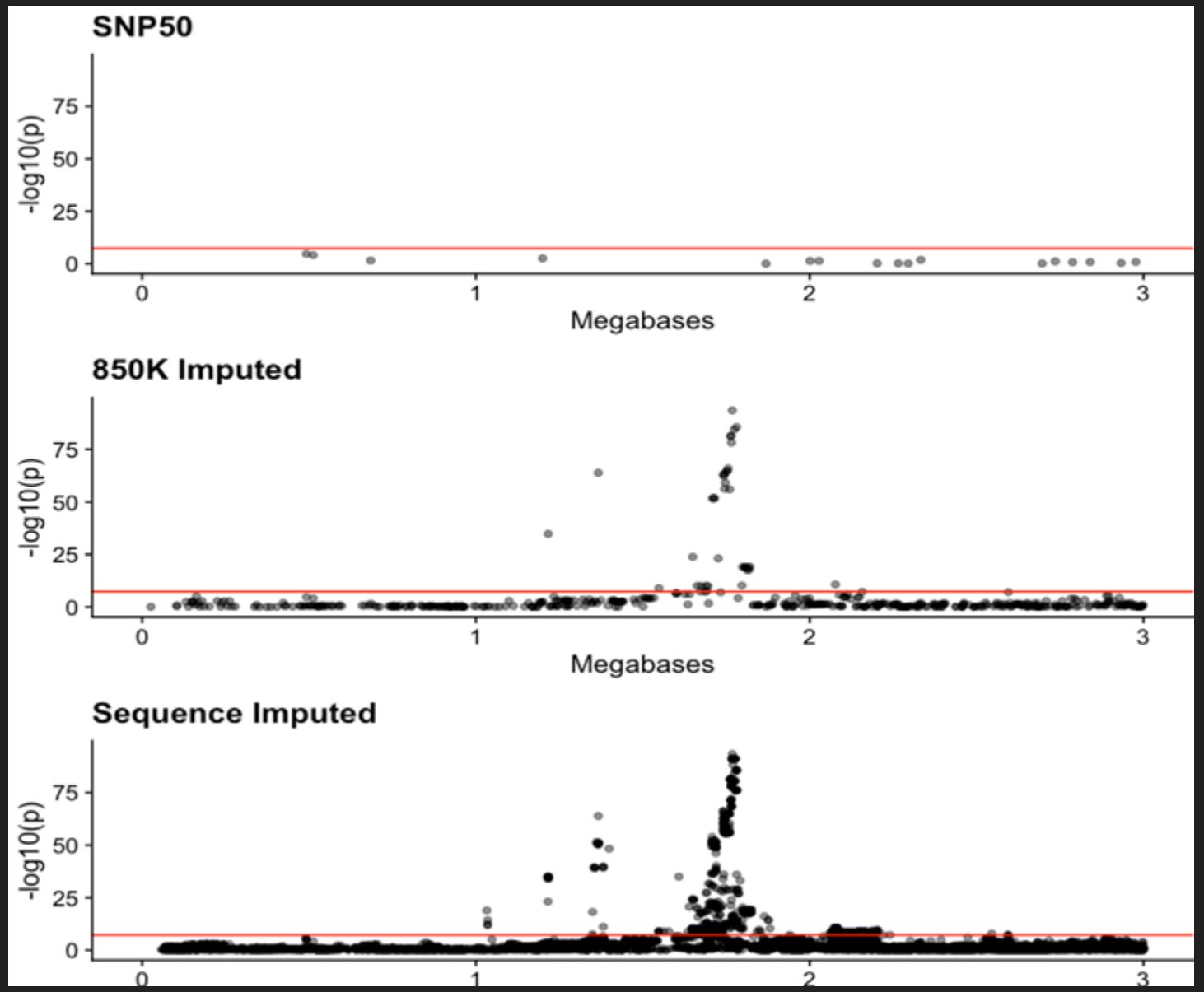
Rubinacci et al 2021

GLIMPSE2

UTAgRESEARCH INSTITUTE OF AGRICULTURE THE UNIVERSITY OF TENNESSEE

UTCVM INSTITUTE OF AGRICULTURE THE UNIVERSITY OF TENNESSEE

The Power of Imputation

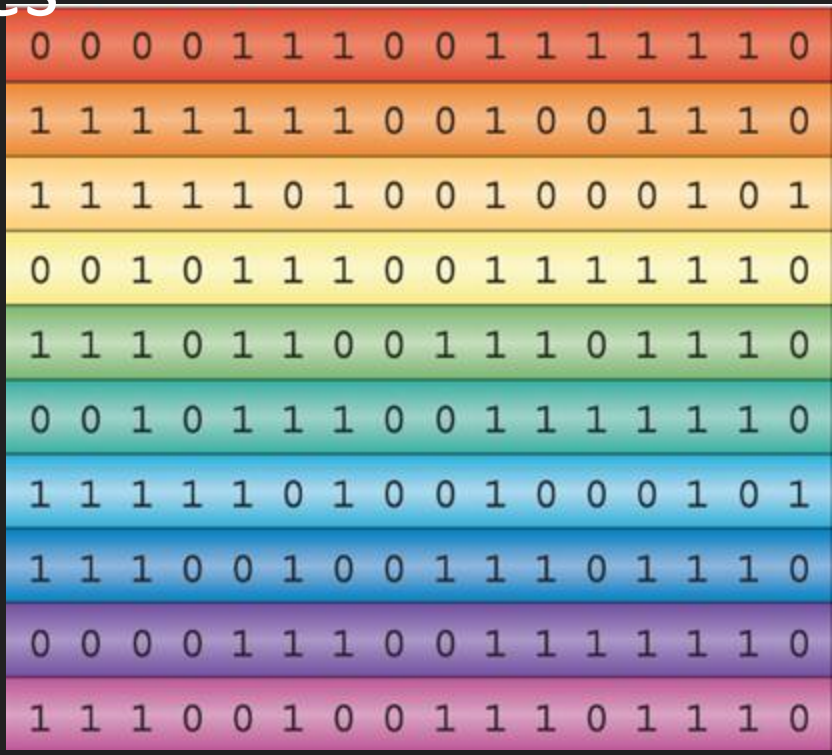


Rowan et al. 2021

UTCVM INSTITUTE OF AGRICULTURE THE UNIVERSITY OF TENNESSEE

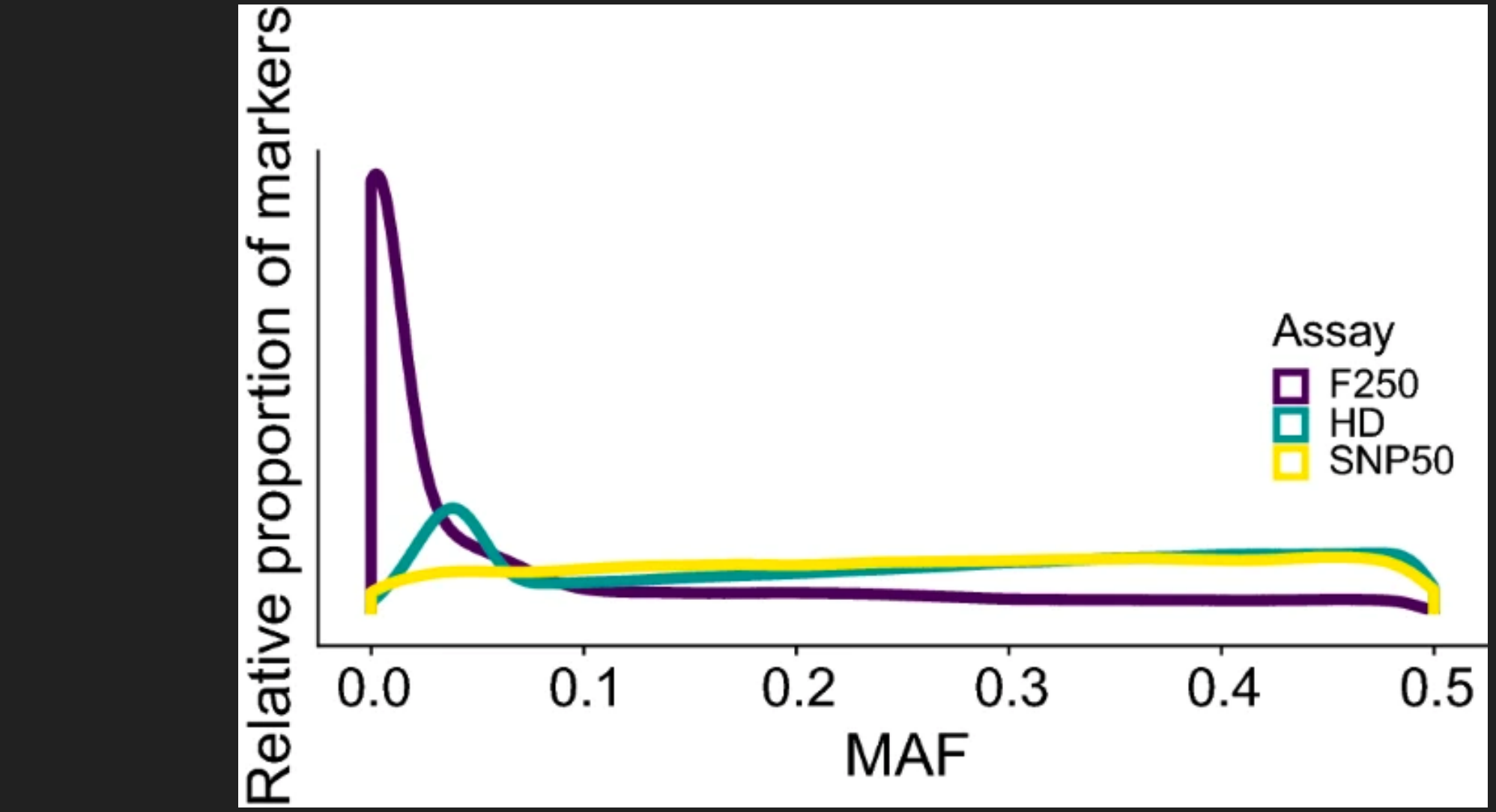
What does accurate imputation need?

- **A large reference set of haplotypes**
  - High-coverage re-sequenced haplotypes
  - Representative of target population haplotypes
- High-quality reference genome
  - Physical positions matter
- Recombination map

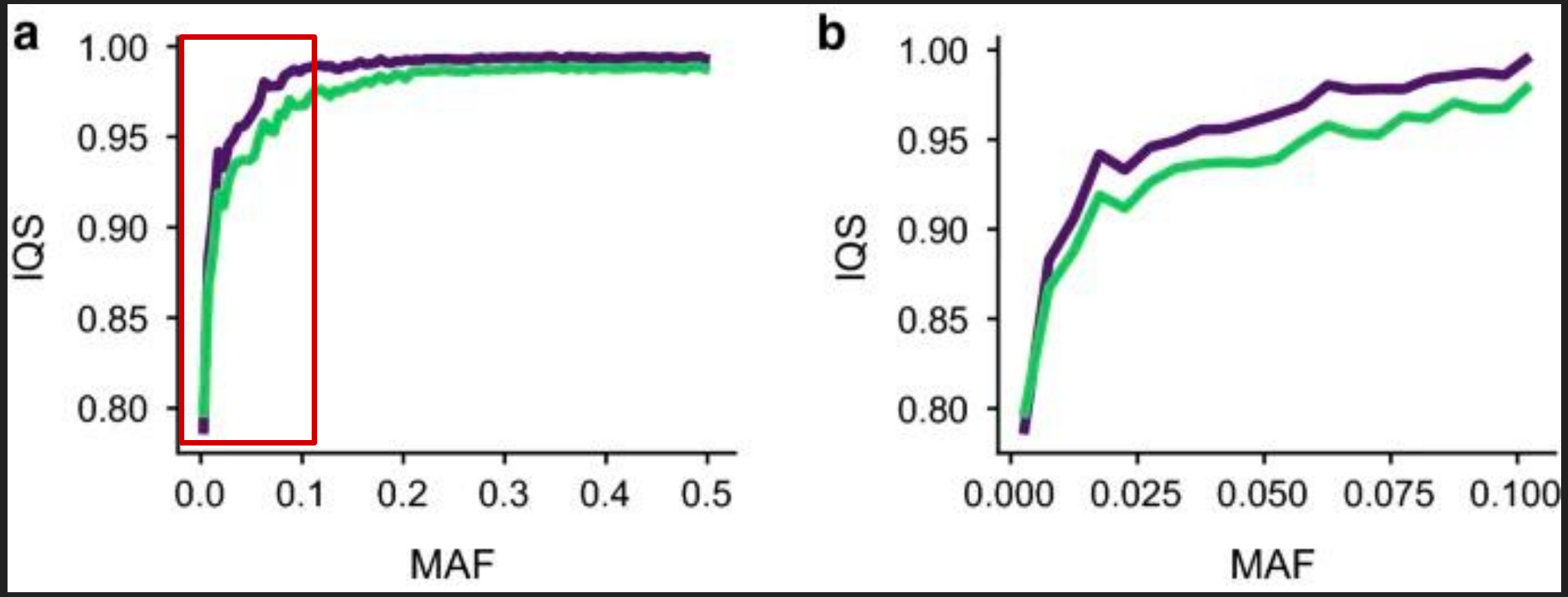


Marchini et al. 2010

Imputation opportunity & challenge: Rare variation



We can’t impute what we don’t observe



As such, rare variation is a challenge

Rowan et al., 2019

The big questions:

- Who do we sequence?
- How deep do we sequence?
- How often do we update?
  - Reference
  - Imputed samples in evaluation

The big questions:

- Who do we sequence?
- How deep do we sequence?
- How often do we update?
  - Reference
  - Imputed samples in evaluation

Before we talk about sequencing for imputation…

We should be sequencing ALL sires with even moderate levels of AI usage!

Why sequence all AI sires?

- “Insurance Policy”: Accelerate abnormality mapping and management
- Proactive monitoring of *de novo* mutations
- Lethal haplotype mapping at sequence resolution
- Enable haplotype-aware analyses
- Increased imputation qualities
- Current costs make this tenable!

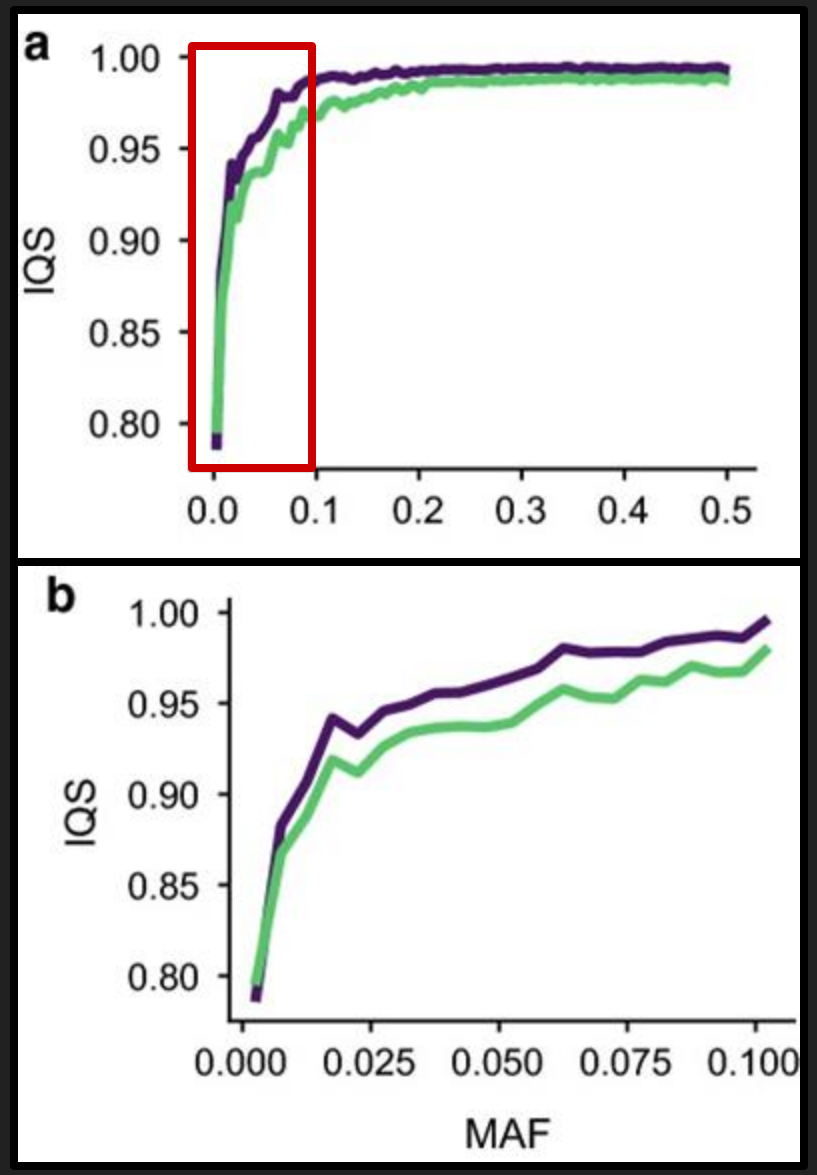
How to build a reference panel?

Breed-specific

Multi-breed

Admixed populations will benefit from a multi-breed reference

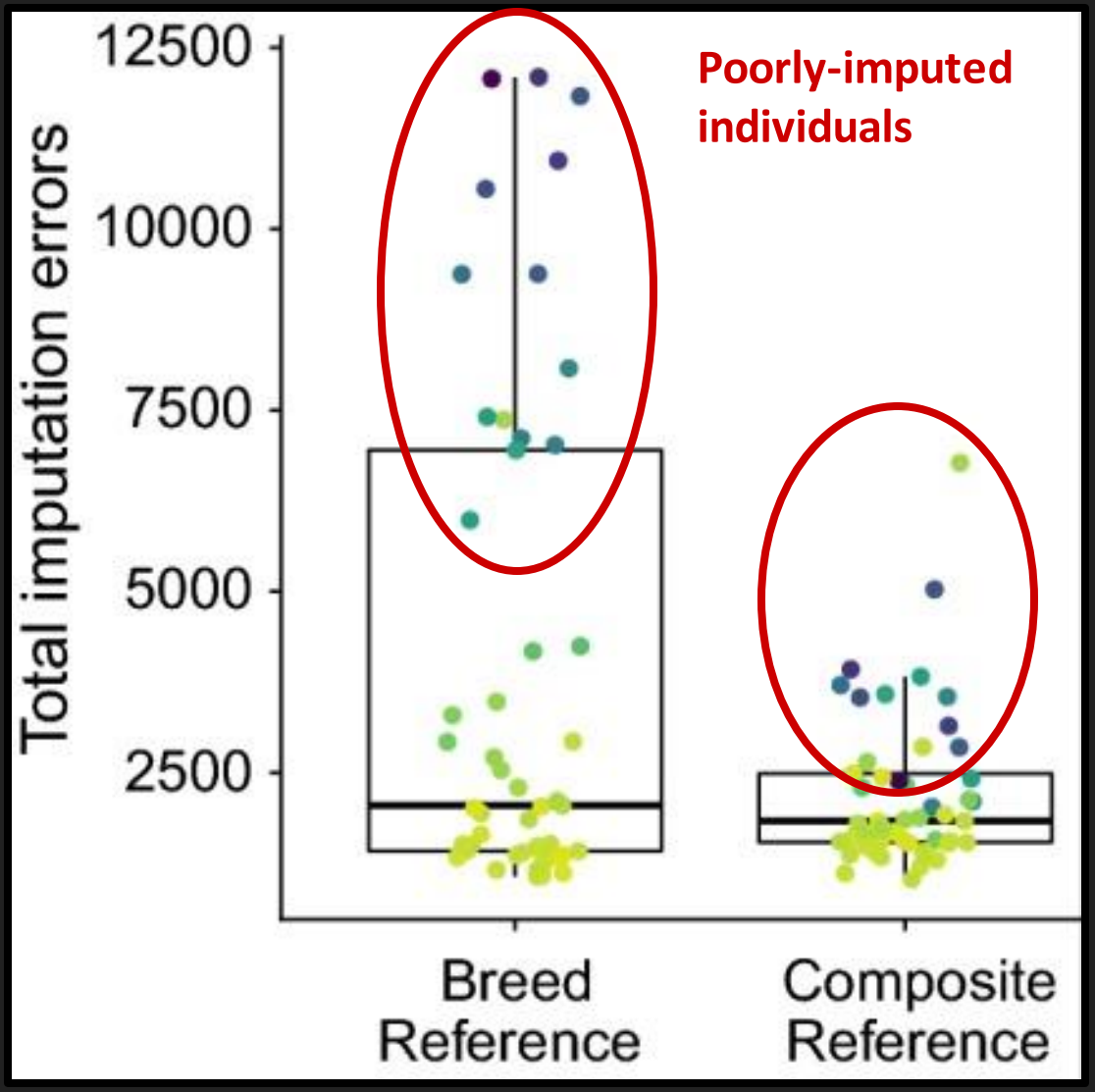
- Admixed populations need representation across diversity of individuals
- Labelled population ≠ Actual population
- Draw on haplotype diversity from other population in imputation reference
- **Using multi-population reference significantly improves per-SNP and per-individual imputation accuracy across samples!**



Rowan et al. 2019

Imputation is just pattern matching!

n =50 Gelbvieh

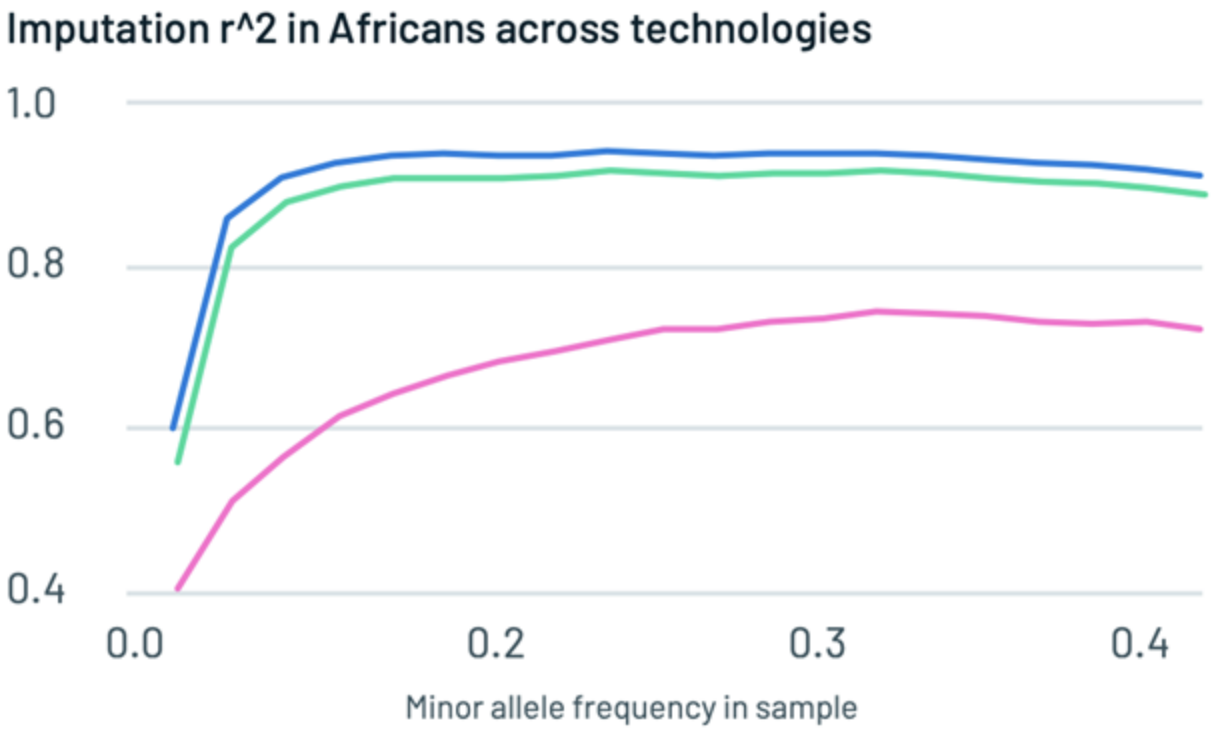


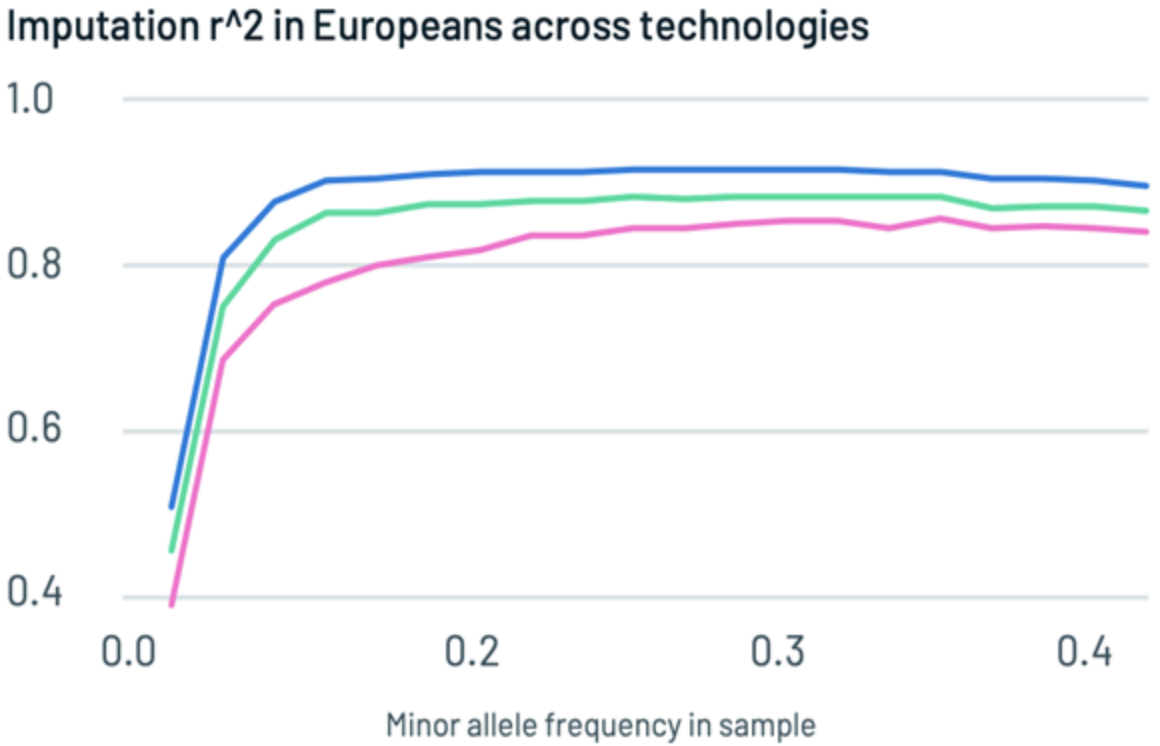
850K Chip Imputation

Breed	Mean	Min	Max
Gelbvieh	0.998	0.994	0.999
Hereford	0.997	0.991	0.999
Holstein	0.997	0.995	0.998
Simmental	0.996	0.984	0.999
Angus	0.995	0.959	0.999
Jersey	0.995	0.991	0.997
Limousin	0.989	0.930	0.996
Nelore	0.981	0.977	0.984
Brahman	0.941	0.932	0.961
Gir	0.903	0.869	0.948
Romagnola	0.874	0.855	0.896
N'Dama	0.763	0.747	0.803

Rowan et al. 2019

Ascertainment Bias: Human Example





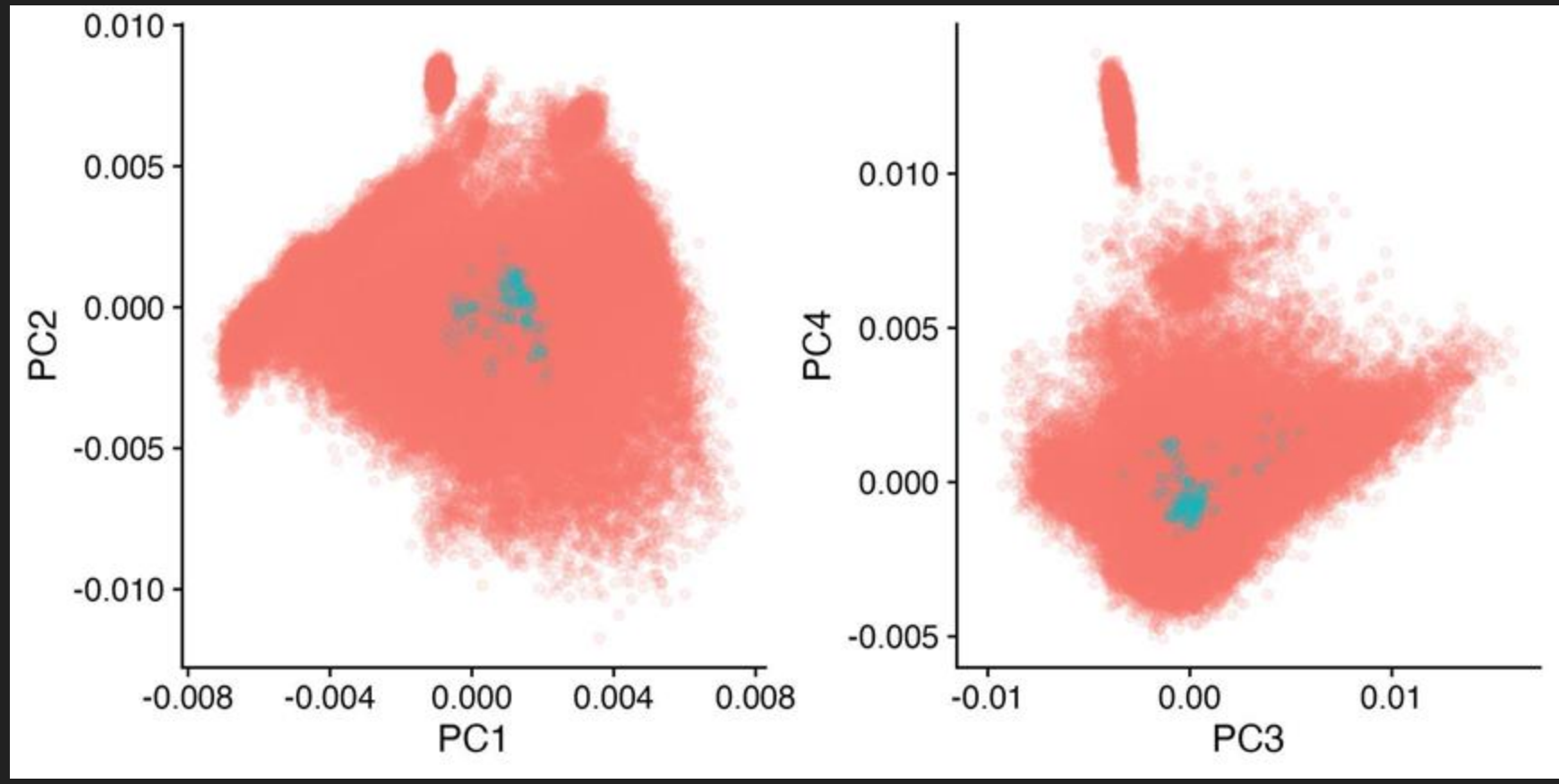
Gencove 0.5x Gencove 1x Illumina GSA

We have to move past sequencing only most common animals

Case Study from ASA:

- Top 150 sires cluster very closely together in PCA of full genomic dataset

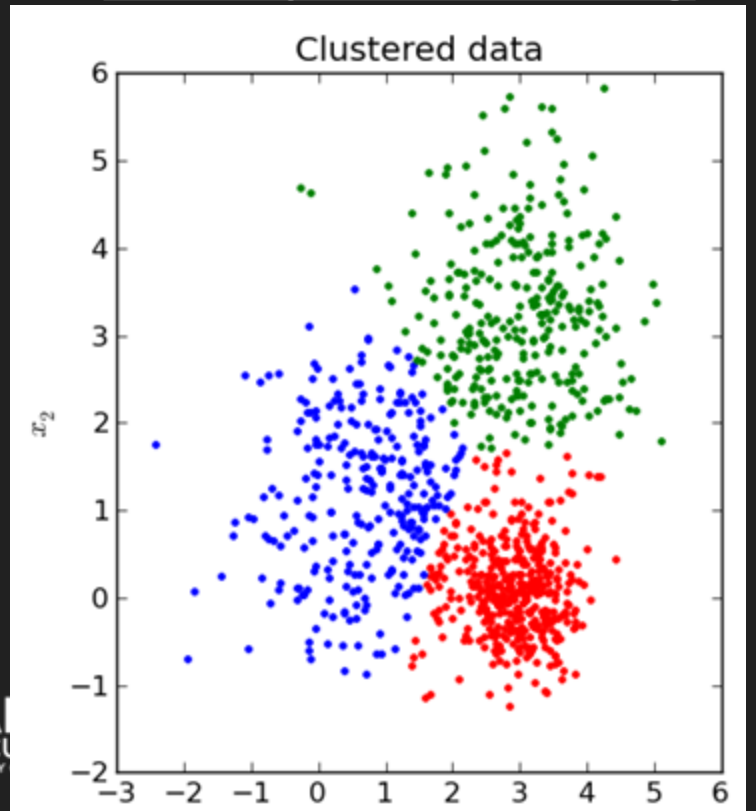
- Sequencing only heavily-used bulls will sample only a small portion of haplotype-space



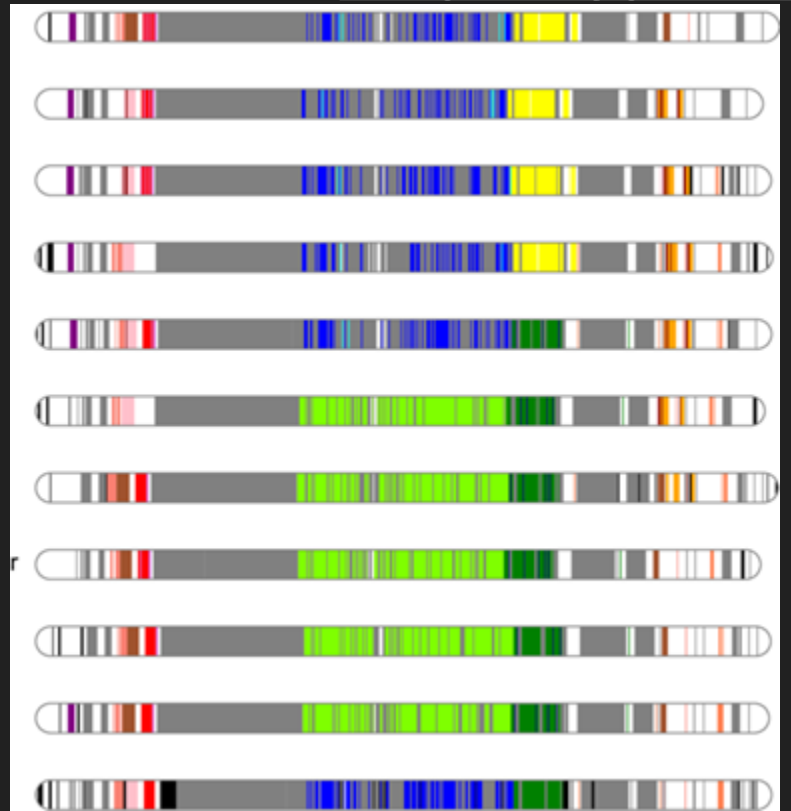
So how might we select sires to sequence?

- Use chip genotype data!
- Iteratively search for sequencing candidates

Group clustering



Haplotype search



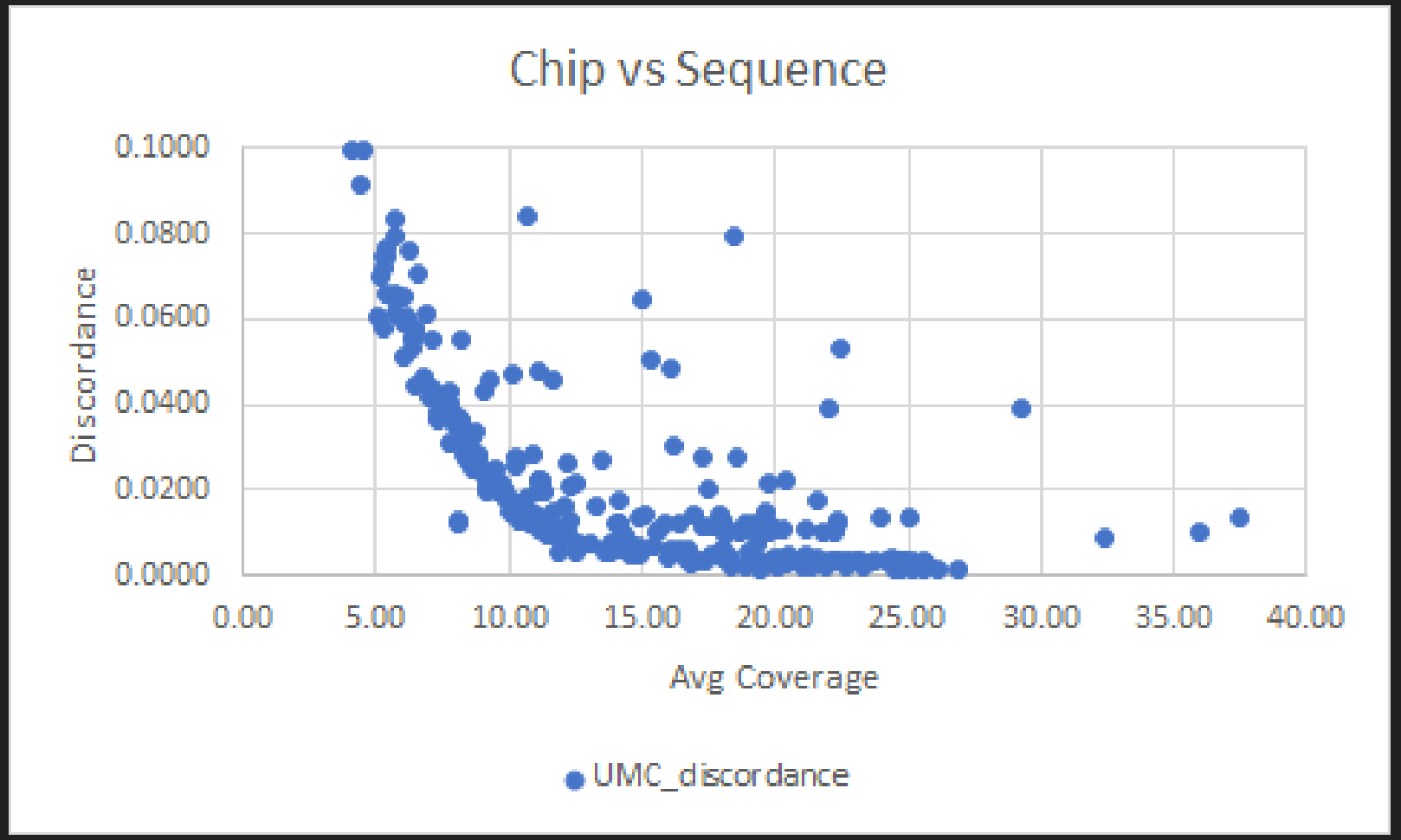
The big questions:

- Who do we sequence?
- How deep do we sequence?
- How often do we update?
  - Reference
  - Imputed samples in evaluation

Dr. Bob Schnabel
Mizzou

Sequencing depth = greater genotype confidence

>10X coverage resulted in substantially more non-reference discordances (i.e. wrongly called genotypes)



% Bob Schnabel

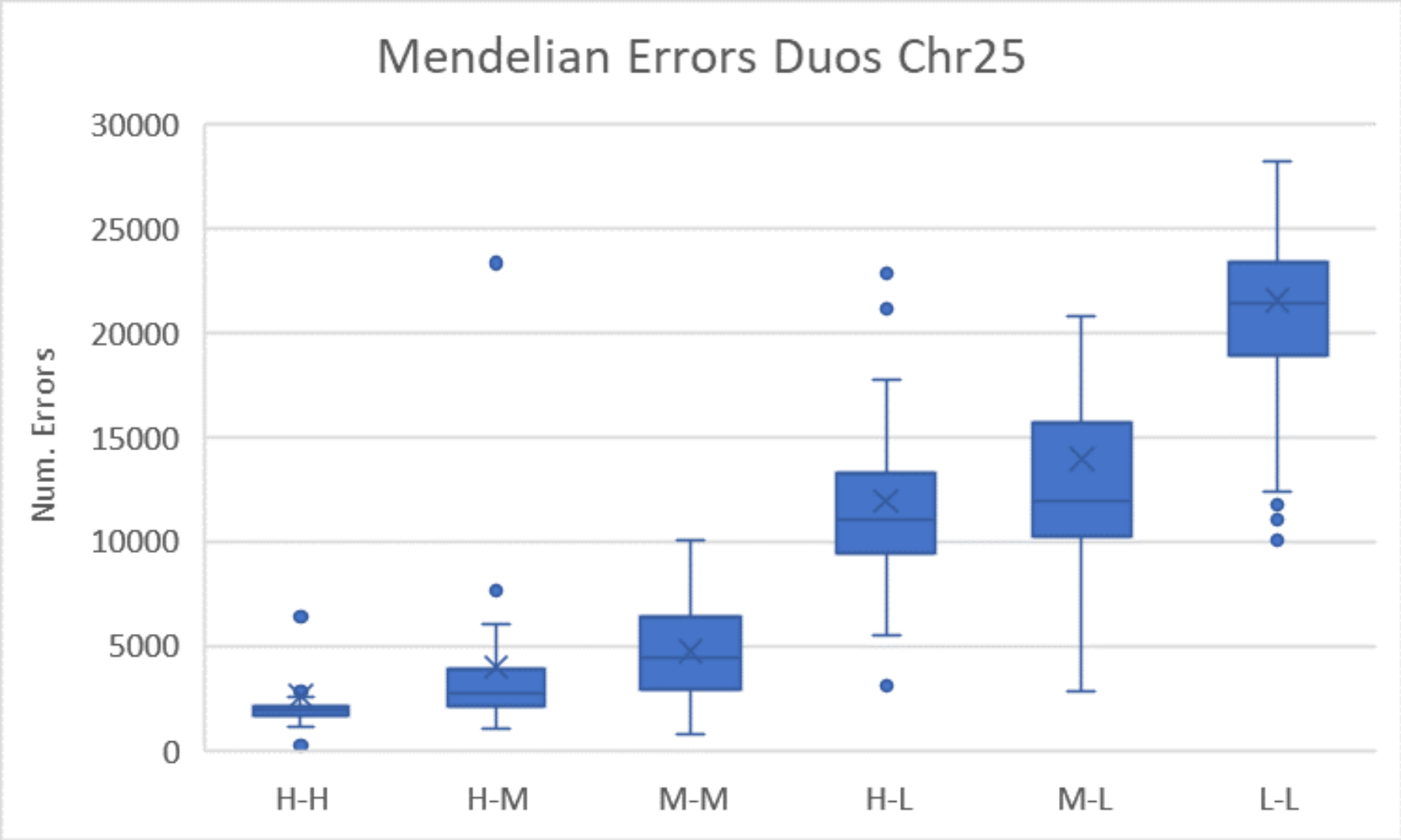
We should generate at least 15X genomes for reference individuals!

Element & Illumina are both generating sequence at $2/GB

Miniaturized library preparations

Marginal cost increase between 5X and 15X is between $60-$150

Coverage Impacts on Genotype Call Quality

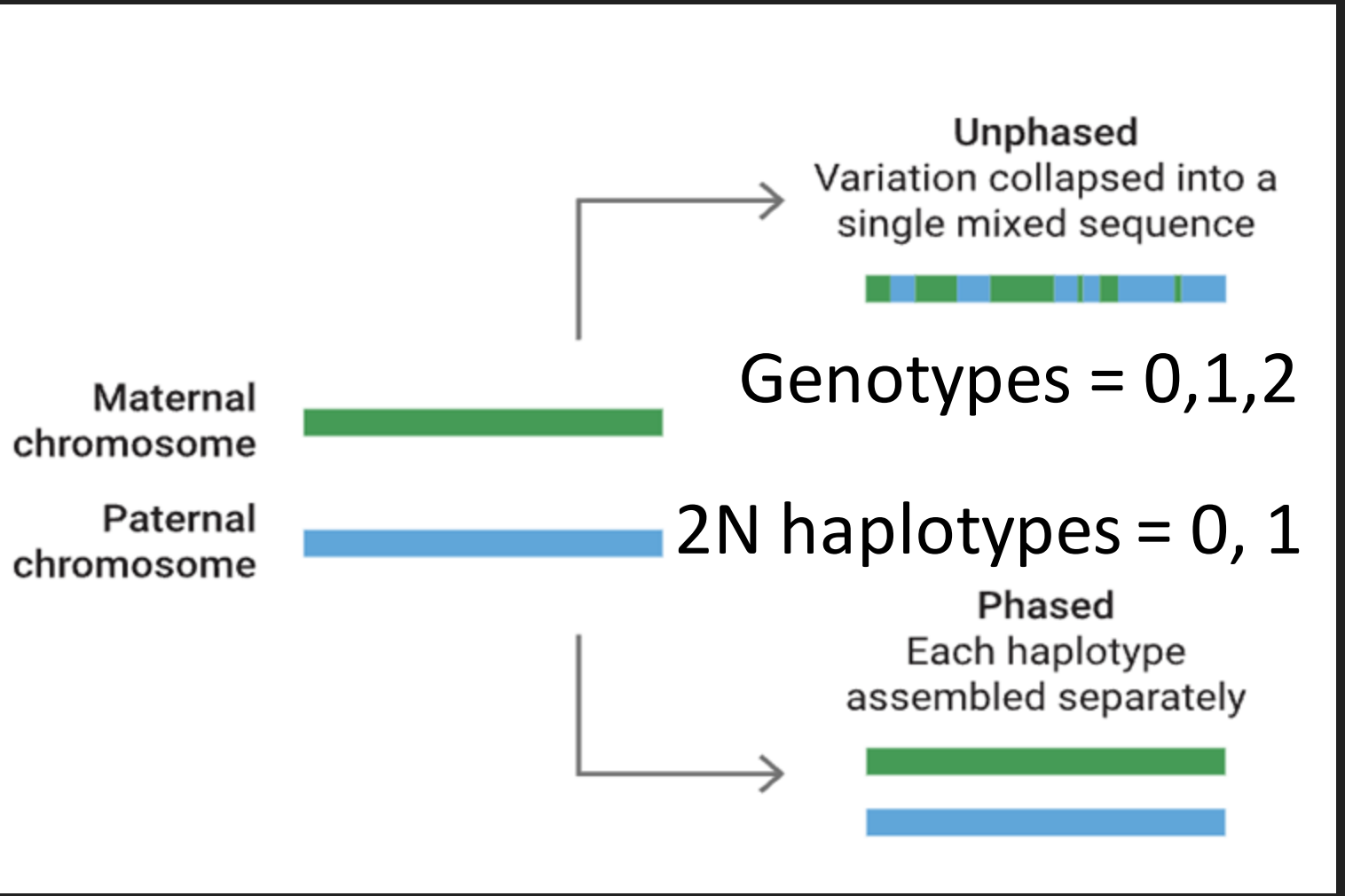


H = > 20X

M = 10-20X

L = < 10X

Sequencing is just the first step… Phasing matters!



Our pedigrees can do much of this phasing for us (>60%)

Read-backed phasing

Various HMM algorithms (SHAPEIT5)

And so does continually empirically evaluating accuracy

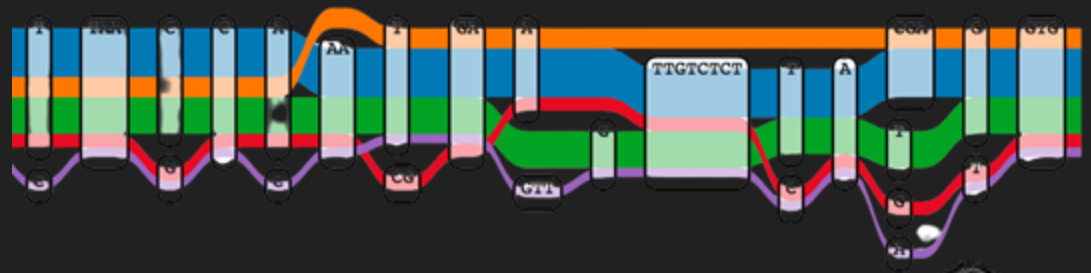
Testing Individuals: True high-coverage calls

Downsampled to low-coverage reads or chip genotypes

Evaluate imputation accuracy: On per-SNP and per-individual basis

Other things to consider:

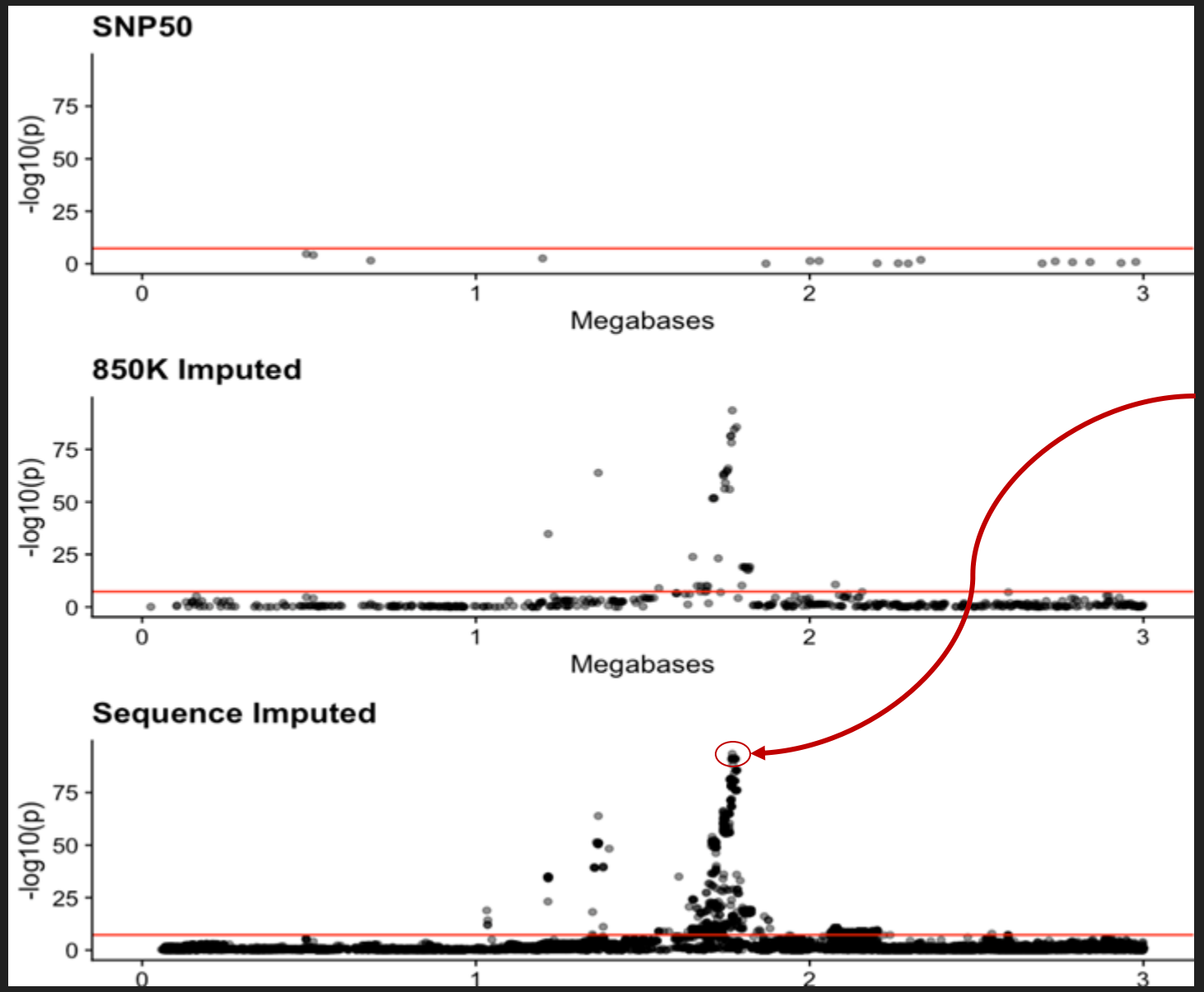
- Pangenomes



- Moderate-coverage sequencing

- Storage of genomes and imputed genotypes

The Million Dollar Question:



How do we use this information to improve predictions?

Rowan et al. 2021

UTCVM INSTITUTE OF AGRICULTURE THE UNIVERSITY OF TENNESSEE

Haplotype reference panels must be representative of the target populations.

Multi-breed > Within-breed

High-quality reference sequences should be at least 10X to be optimally useful.

It is important to regularly evaluate imputation accuracy for both individuals and SNPs

Reach out with questions!

trowan@utk.edu

(865) 974-3190


@TroyNRowan