

Phasing and Haplotype Construction

agresearch
āta mātai, mātai whetū

Andrew Hess
Imputation Workshop
28/9/20

What is a Haplotype?

- Haplotype definition
 - A set of genes inherited together from one parent on one chromosome
 - All genes on a chromosome inherited together from a single parent
 - Inherited together because of genetic linkage
 - i.e. genes that are close to each other on the same chromosome tend to be inherited together.
 - Can also refer to variants inherited together on a chromosome, rather than genes
 - This workshop will focus primarily on SNPs
- Resolving haplotypes i.e. Phasing
 - Sequencing/wet lab approaches
 - Computational approaches
 - This workshop will focus primarily on computational approaches
 - Aims to capture genomic segments that are the same between individuals because they are identical-by-descent (inherited from the same ancestor)

Sequencing Methods

Haplotype-resolved genome sequencing: experimental methods and applications

Matthew W. Snyder¹, Andrew Adey², Jacob O. Kitzman^{3,4} and Jay Shendure¹

- Dense direct methods
 - Extensively resolve local haplotypes
 - Any given heterozygous variant is successfully phased with respect to the other variants in the same region
 - Yield haplotype blocks that are typically hundreds of kb to several Mb
 - Little or no experimental information relates these haplotype blocks to other blocks on the same chromosome.
 - Long-read sequencing, CPT-seq, Dilution pooling
- Sparse direct methods
 - Leave many individual variants unphased but provide phase information for a subset of variants across much longer physical distances, up to entire chromosomes.
 - HaploSeq, Single-chromosome sequencing, Emulsion PCR

Haplotype phasing: existing methods and new developments

Sharon R. Browning and Brian L. Browning†*

Computational Methods

- “Unrelated” individuals
 - Use haplotypes found within the population
 - Very reliant on the number of individuals present
 - Generally: More individuals → more haplotypes observed → more accurate phasing and imputation
 - Do a poor job at imputing rare haplotypes/variants
- Related individuals
 - Utilize pedigree information initially
 - Tracking Identity by descent (IBD)
 - Fill in any gaps/uncertainties using population haplotypes
- Accuracy is influenced by:
 - Sample size
 - Marker density
 - Genotype accuracy
 - Degree of relatedness
 - Genetic background
 - Allele frequency

Comparison

- Computational
 - Low cost – feasible on large number of individuals
 - Challenged by low-frequency variants, private variants and de novo variants
 - Limited by the magnitude and extent of linkage disequilibrium
 - Differs depending on ancestry
 - Pedigree-based haplotype inference requires the genotyping of multiple individuals from the same family
- Sequencing
 - Fully resolve haplotypes for all forms of variation genome-wide using only the sample of interest
 - Expensive and don't scale well
 - Limitations of direct methods can be partially overcome through their combined application with inferential methods.
- Both
 - Properties of the chromosome itself
 - Runs of homozygosity may result in a break in the haplotype assembly
 - Long regions of repetitive sequence also typically result in breaks in haplotype assemblies
 - i.e. duplications, CNVs

Statistical Phasing

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12	SNP13	SNP14	SNP15	SNP16
Ind1	1	1	1	1	2	2	2	0	0	2	1	1	2	2	2	0
Ind2	1	1	2	1	2	1	2	0	0	2	1	1	1	2	1	1
Ind3	1	1	2	0	2	2	1	0	1	2	2	1	2	2	2	0
Ind4	2	2	2	1	1	1	1	0	1	2	1	0	1	2	1	1
Ind5	1	1	1	0	1	2	1	0	1	2	2	1	2	2	2	0

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12	SNP13	SNP14	SNP15	SNP16
Ind1_M	0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind1_P	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
Ind2_M	1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
Ind2_P	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind3_M	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
Ind3_P	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind4_M	1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
Ind4_P	1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
Ind5_M	0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind5_P	1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

	Haplock1				Haploblock2				Haploblock3				Haploblock4			
Ind1_M	0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind1_P	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
Ind2_M	1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
Ind2_P	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind3_M	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
Ind3_P	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind4_M	1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
Ind4_P	1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
Ind5_M	0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind5_P	1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

	HB1	HB2	HB3	HB4
Ind1_M	A	A	A	A
Ind1_P	B	A	B	A
Ind2_M	B	B	B	B
Ind2_P	C	A	A	A
Ind3_M	D	C	C	A
Ind3_P	C	A	A	A
Ind4_M	B	B	B	B
Ind4_P	D	D	C	A
Ind5_M	A	A	A	A
Ind5_P	D	D	C	A

Many programs:

- Beagle
- Fimpute
- PHASE

Downstream Analysis

- Haplotype Diversity
- Haplotype GWAS
- Etc

Based on:

- Number of SNPs
- Location
- LD/Recombination

Imputation

Reference Haplotypes

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12	SNP13	SNP14	SNP15	SNP16
Ind1_M	0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind1_P	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
Ind2_M	1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
Ind2_P	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind3_M	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
Ind3_P	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind4_M	1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
Ind4_P	1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
Ind5_M	0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
Ind5_P	1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

Low-Density Genotypes

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12	SNP13	SNP14	SNP15	SNP16
Ind6	1	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
Ind7	2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
Ind8	1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

Haplotypes as a mosaic of reference

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12	SNP13	SNP14	SNP15	SNP16
Ind6_M	0	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
Ind6_P	1	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
Ind7_M	1	?	?	?	1	?	1	?	0	1	0	?	?	1	?	0
Ind7_P	1	?	?	?	1	?	1	?	1	1	1	?	?	1	?	0
Ind8_M	1	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0
Ind8_P	0	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0

Inferred Genotypes

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12	SNP13	SNP14	SNP15	SNP16
Ind6	1	1	1	1	2	2	2	0	0	2	2	0	2	2	2	0
Ind7	2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	0
Ind8	1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	0

Programs Available: HMM

Beagle

(<http://faculty.washington.edu/browning/beagle/beagle.html>)

- Beagle 4.0
 - Can use genotype likelihoods as input
 - Allows use of pedigree information
- Beagle 5.0
 - Substantially faster and more accurate than 4.0
 - Can accommodate multiallelic markers

SHAPEIT

- SHAPEIT2
(https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html)
 - Integrate sequence and array data
- SHAPEIT4
(<https://github.com/odelaneau/shapeit4>)
 - Integrates multiple sources of information
 - Externally phased reference panels
 - Collections of pre-phased genotypes
 - Long-read sequence data

Programs Available: Rule-Based

Fimpute

- Academic licence available upon request
- Uses pedigree information
- Fills in the remaining missing information using a window-based approach
- Fast and less memory intensive method for phasing

AlphaImpute

(<https://alphagenes.roslin.ed.ac.uk/wp/software-2/alphaimpute/>)

- Uses pedigree
- Heuristic method
 - basic rules of Mendelian inheritance
 - segregation analysis
 - long-range phasing
 - haplotype library imputation

NEW: AlphaImpute2

- Combines pedigree and HMM methods for fast and accurate algorithm



A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes

J. E. Pryce,^{*1} S. Bolormaa,^{*} A. J. Chamberlain,^{*} P. J. Bowman,^{*} K. Savin,^{*} M. E. Goddard,^{*†} and B. J. Hayes^{*}

^{*}Biosciences Research Division, Department of Primary Industries Victoria, Bundoora 3083, Australia

[†]Faculty of Land and Food Resources, University of Melbourne, Parkville 3010, Australia

- The haplotype method improves the power to detect QTL as demonstrated by
 - reduced FDR in the discovery data set
 - greater proportion of validated associations
 - shorter genomic regions containing QTL
 - detection of a putative QTL for fertility that was not detected by single SNP associations

RESEARCH ARTICLE

Open Access



Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population

Melanie Hess^{1,2*}, Tom Druet³, Andrew Hess¹ and Dorian Garrick^{1,4}

- Fitting haplotype alleles rather than SNPs can increase prediction accuracy
- Improved genomic prediction accuracy with comparable computation time to fitting SNPs
- Increased accuracy is likely to increase genetic gain by changing the ranking of selection candidates

Applications: GWAS


Use of Ancestral Haplotypes in Genome-Wide Association Studies

Tom Druet and Frédéric Farnir

- Overview of HMM for phasing
- Discusses use of ancestral haplotype states for GWAS
 - Multiple examples, one being (<https://www.genetics.org/content/184/3/789>)
- PHASEBOOK
 - Efficient heuristic approach based on hidden Markov models (HMM)
 - It simultaneously phases and sorts haplotypes in clusters that can be used directly for mapping or other purposes
 - Exploits familial as well as population information



Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle

Majid Khansefid^{1,2*} , Jennie E. Pryce^{2,3}, Sunduimijid Bolormaa², Yizhou Chen⁴, Catriona A. Millen⁵, Amanda J. Chamberlain², Christy J. Vander Jagt² and Michael E. Goddard^{1,2}

Application: Genetical Genomics

- Expression Data (<https://link.springer.com/article/10.1186/s12864-018-5181-0>)
 - Association of particular variant with expression differences (eQTL) – tracking alleles allows detection of alleles allows association of specific alleles with gene expression
 - Identifying paternally/maternally inherited haplotypes allows for detection of whether there is a SNP that alters gene expression a Parent-of-Origin manner (i.e. whether the locus is expected to be imprinted).
- Can also be used with phenotypes (<https://www.sciencedirect.com/science/article/pii/S0002929715003213>)
 - Detect loci where, e.g. disease outcome, is dependent on which parent the genetic variant is inherited from.

Review

Using Haplotype Information for Conservation Genomics

Maeva Leitwein,^{1,3,*} Maud Duranton,^{2,3} Quentin Rougemont,^{1,3} Pierre-Alexandre Gagnaire,^{2,4} and Louis Bernatchez^{1,4}

Application: Conservation Genomics

- Using haplotype information can improve the inference of population demographic parameters including effective population size and migration
 - Tracked by IBD segments
- Can be used for local ancestry inference/admixture
- More accurate identification of selection sweeps

Considerations for Experimental Design

How many individuals?

What is the minimum number of individuals needed to capture the diversity in the population?

- How many are needed for a “core” set of individuals



Which individuals?

Which animals should be selected to maximize the information captured?

- Individuals to capture a higher density/depth

Value of Resources Already Available?

How representative are the resources already available?

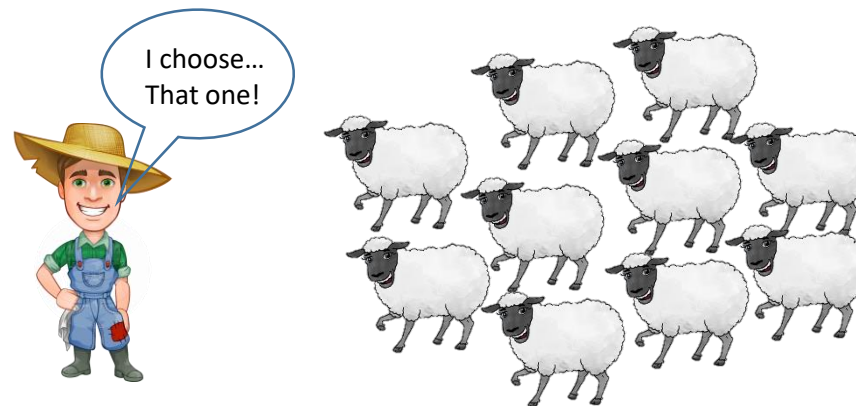
- Is there value in capturing more than what is already available?

Optimizing Sequencing Resources in Genotyped Livestock Populations Using Linear Programming

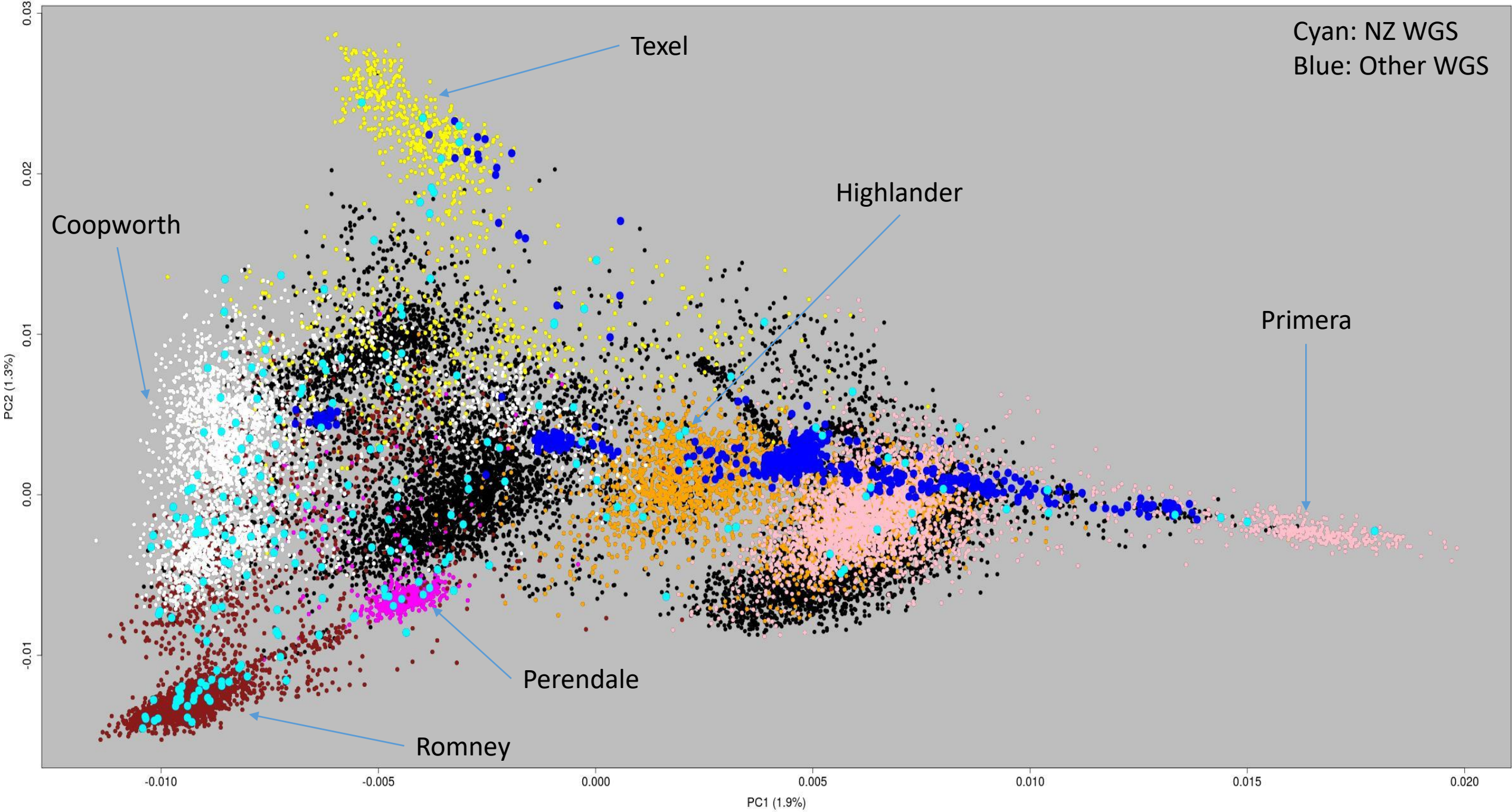
Hao Cheng^{1*}, Keyu Xu¹ and Kuruville Joseph Abraham²

LPChoose

- Linear programming algorithm for allocation of sequencing resources across animals in a population
- Sequencing is costly; therefore it is usually not feasible to sequence all animals
 - Which animals do we need to sequence to:
 - Capture all haplotypes in the population with the least animals sequenced?
 - Maximise the number of haplotypes captured for a given number of animals sequenced?



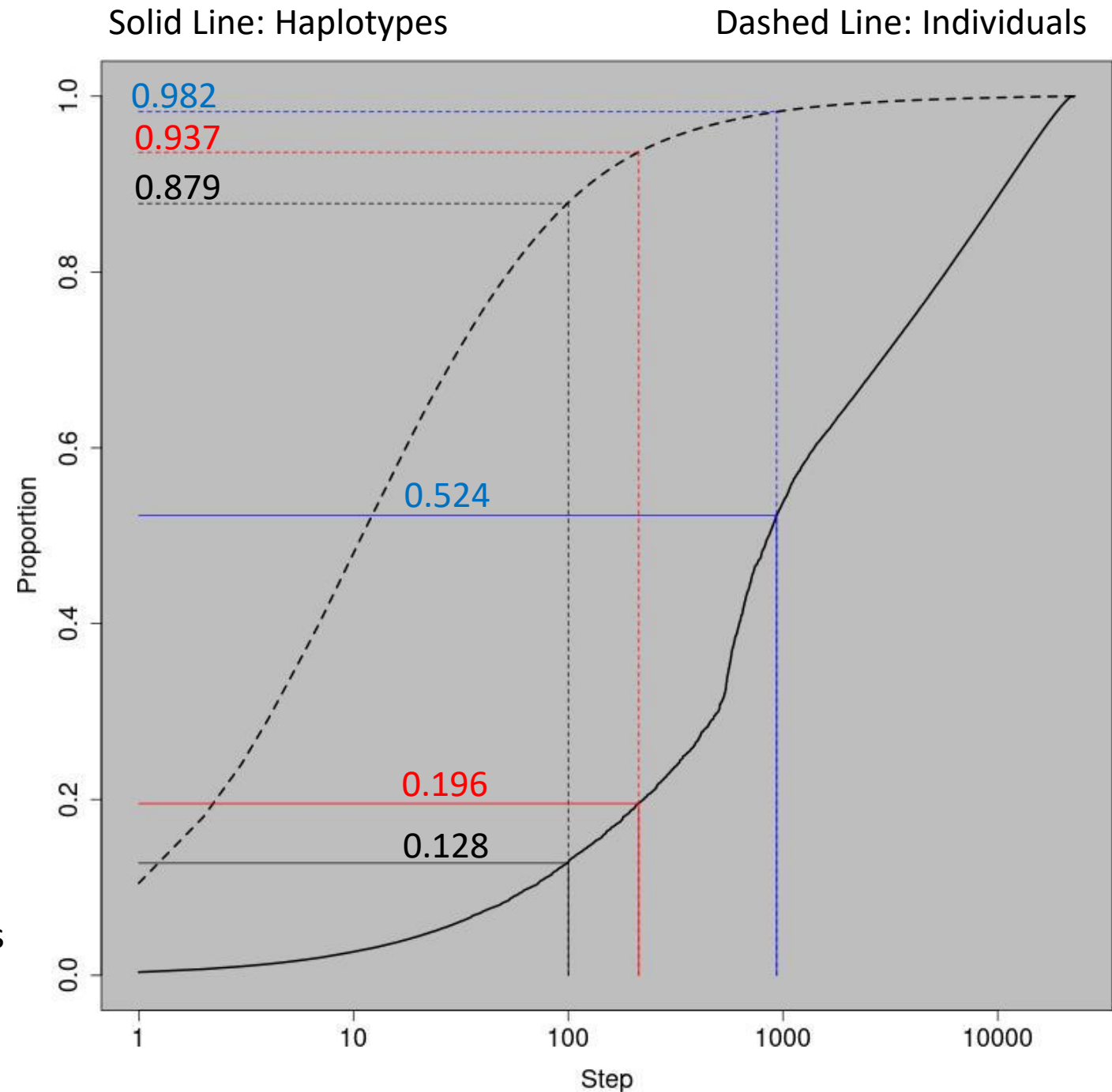
Based on HD Genotypes



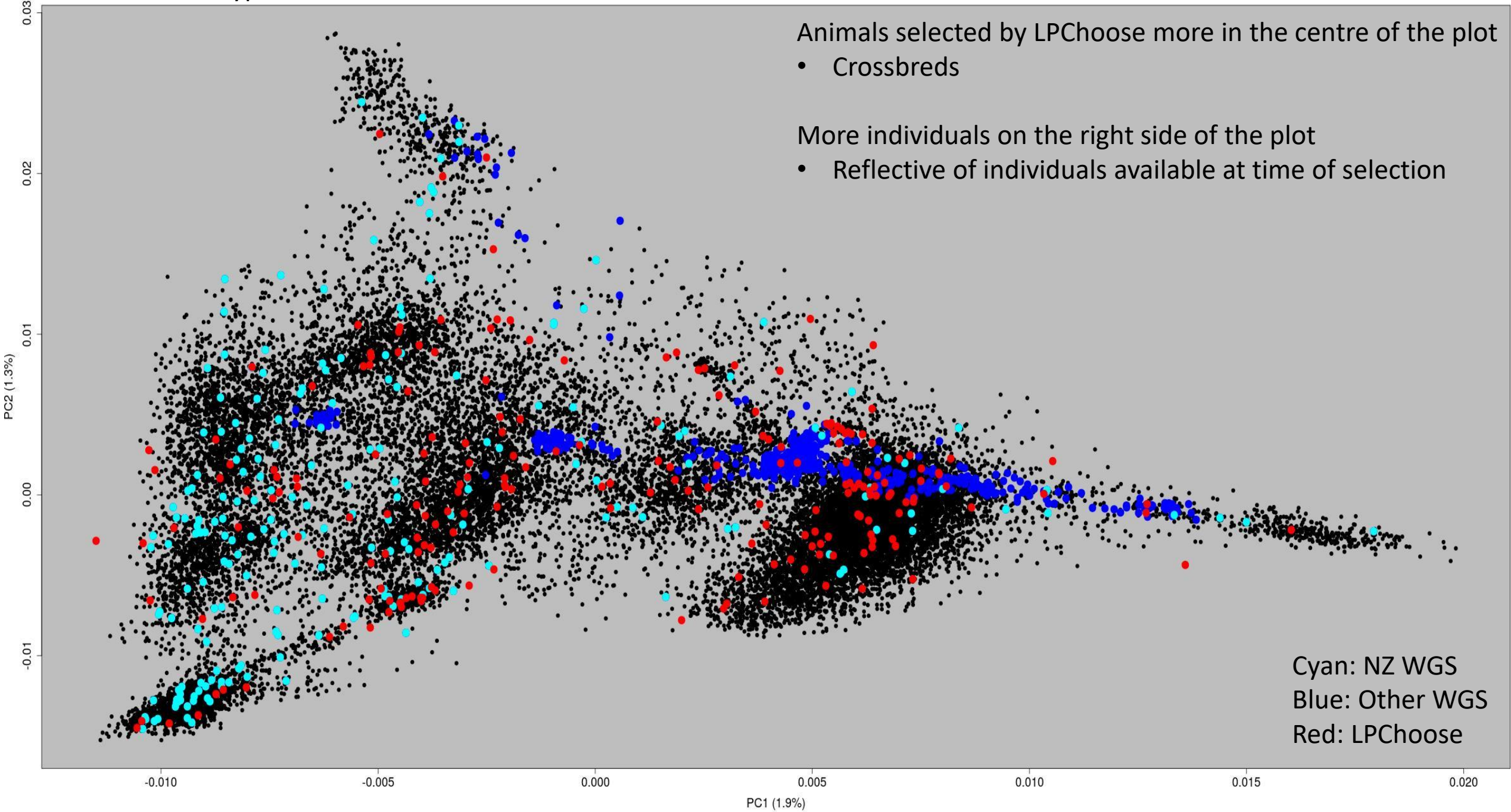
How many animals to choose?

- Frequent haplotypes get captured quickly
- Many animals have at least one low-frequency haplotype
 - 22,765 of 22,887 to capture all present
- How quickly haplotypes are captured depends on definition

Based on HD Genotypes
Haploblock: 250 Kb



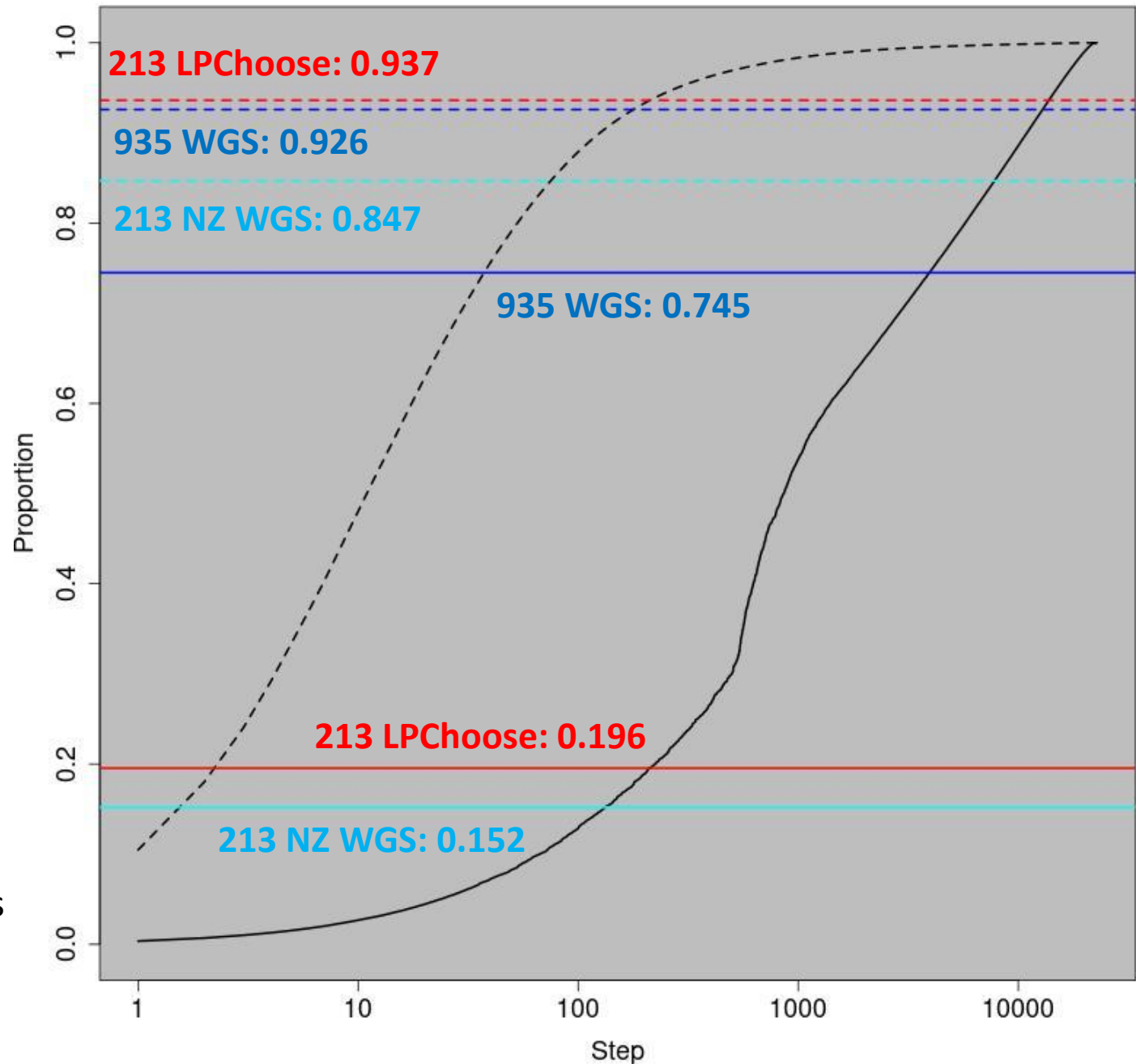
Based on HD Genotypes




How well do the animals with WGS data do?

- Most of the haplotypes are covered in the WGS set
- NZ animals captured more unique haplotypes than 213 chosen by LPChoose
- Inclusion of non-NZ animals results in more NZ haplotypes being covered

Based on HD Genotypes
Haploblock: 250 Kb



Imputation from HD to WGS

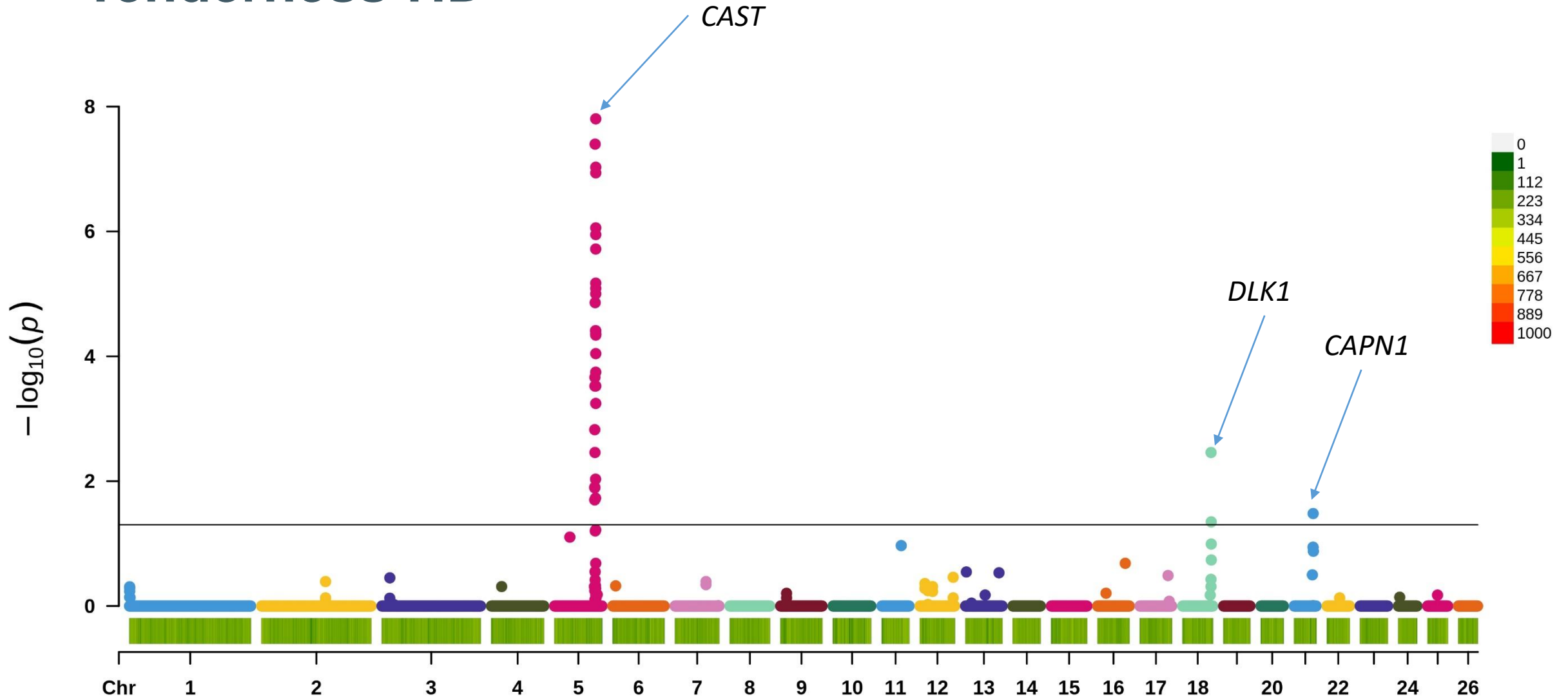


	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12	SNP13	SNP14	SNP15	SNP16
Ind6	1	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
Ind7	2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
Ind8	1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

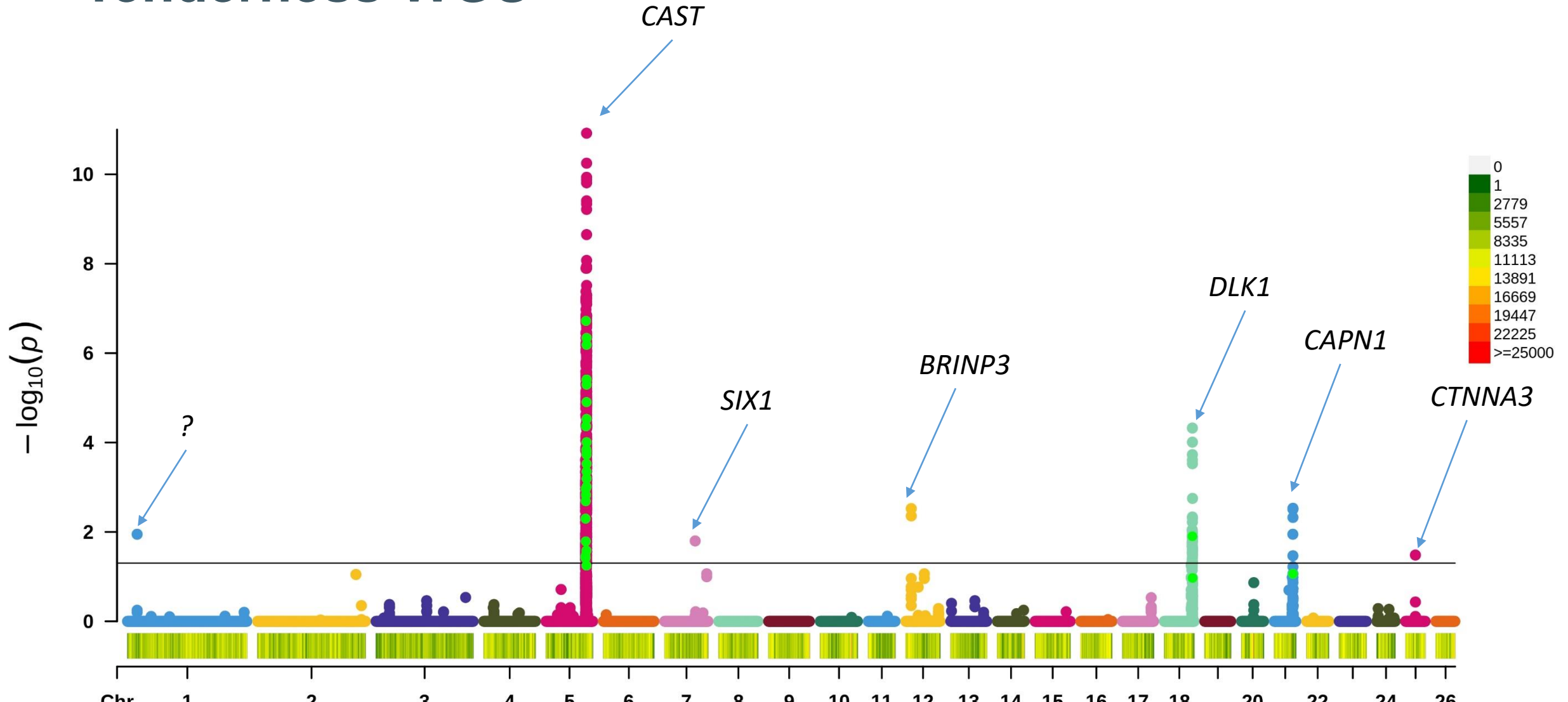
	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP11	SNP12	SNP13	SNP14	SNP15	SNP16
Ind6	1	1	1	1	2	2	2	0	0	2	2	0	2	2	2	0
Ind7	2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	0
Ind8	1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	0

- Phasing/Imputation in Beagle 5.1
 - Default parameters except $N_e=500$
- Masked 759 Chromosome 26 SNPs in HD animals
 - Imputed to 505,190 WGS SNPs
 - Individual accuracy as concordance between true and imputed genotype across all masked SNPs
- Imputation Accuracy (HD→WGS): 0.965
- Filtered based on MAF (0.01)/Callrate (0.99) in NZ WGS animals (280,750 SNPs)
 - NZ WGS Animals as reference: 0.958
 - NZ and AUS WGS Animals as reference: 0.969
 - All WGS Animals as reference: 0.972

Tenderness-HD



Tenderness-WGS

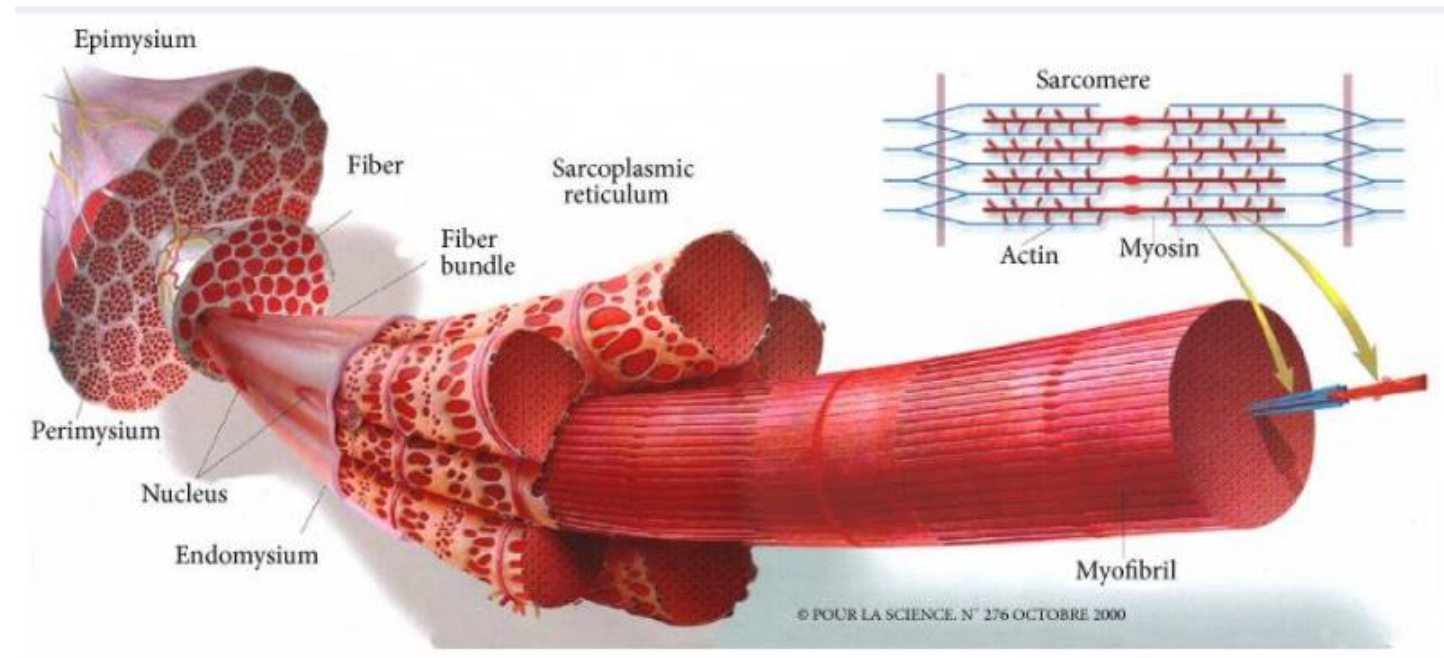


Link between Muscle Fibre Types and Meat Tenderness

Type I	Type IIA	Type IIX	Type IIB
a. Very low mATPase activity	a. Medium mATPase activity	a. High mATPase activity	a. Very high mATPase activity
b. Very high oxidative and very low glycolytic metabolism	b. High oxidative and glycolytic metabolism	b. Low oxidative and high glycolytic metabolism	b. Very low oxidative and very high glycolytic metabolism
c. Very slow contraction speed	c. Medium contraction speed	c. Rapid contraction speed	c. Very rapid contraction speed
d. Small size	d. Small size	d. Large size	d. Very large size
e. Very dense capillary network	e. Dense capillary network	e. Sparse capillary network	e. Sparser capillary network
f. Very high levels of intracellular myoglobin	f. High levels of intracellular myoglobin	f. Low levels of intracellular myoglobin	f. Very low levels of intracellular myoglobin
g. Very high resistance to fatigue	g. High resistance to fatigue	g. Low resistance to fatigue	g. Very low resistance to fatigue
h. Very high mitochondrial density	h. High mitochondrial density	h. Low mitochondrial density	h. Very low mitochondrial density
i. Red color	i. Intermediate color	i. White color	i. White color

† Source: Gerrard and Grant, 2003; Lefaucheur, 2010.

doi:10.2527/af.2017.0437



<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4789028/>

Use in Genomic Prediction

- Forward prediction:
 - Train 2010-2014 (n=9,361)
 - Validate 2015 (n=785)
- Single GRM
 - HD
 - HD+WGS
 - WGS: All SNPs surpassed FDR adjusted p-value (0.05)
- Two GRM
 - Relationship matrix for HD and WGS SNPs separately
 - Allows large effects to remain large

- Breeder's Equation:

$$\Delta G = \frac{i * r_{(EBV, TBV)} * \sigma_G}{L}$$

- Pleiotropic Effects
 - Selection on an index
 - Develop strategies to account for antagonisms between traits
- Epistasis
 - Capture interactions between loci

	HD	HD+WGS	HD+WGS (2GRM)
$r_{(EBV, Phenotype)}$	0.179	0.197	0.243
$r_{(EBV, TBV)}$	0.345	0.373	0.424
		↑ 8%	↑ 23%

Summary

- Haplotypes can be resolved by wet-lab/sequencing or computational methodologies
 - Each approach has its benefits
 - A combined approach can lend more confidence to establishing haplotype phase that can be applied on a population level
- There are a variety of programs that utilise different sources of information for phasing and imputation
- Haplotypes have a variety of uses
 - Increase power for GWAS
 - Population studies for diversity, admixture, migration
 - Imputation for identification of causal variants
 - Self-improvement (reference population design)

Expanding the Definition

Computational pan-genomics: status, promises and challenges

The Computational Pan-Genomics Consortium*

Corresponding author: Tobias Marschall, Center for Bioinformatics at Saarland University and Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany. Tel.: +49 681 302 70880; E-mail: t.marschall@mpi-inf.mpg.de

*The Computational Pan-Genomics Consortium formed at a workshop held from 8 to 12 June 2015, at the Lorentz Center in Leiden, the Netherlands, with the purpose of providing a cross-disciplinary overview of the emerging discipline of Computational Pan-Genomics. The workshop was organized by Victor Guryev, Tobias Marschall, Alexander Schönhuth (chair), Fabio Vandin, and Kai Ye. Consortium members are listed at the end of this article.

- Haploid genome assemblies on a single individual effective for the detection of SNVs and indels
 - Limited in ability to capture structural variation
 - Imputation of structural variation from SNP data has poor performance (<https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-6627-8>)
- Pan-genomes can capture more complex (structural) variation
 - Data structures (i.e. genome graphs) have the potential to store haplotypes
 - Potential to be combined with statistical phasing methods for a more holistic view of haplotype diversity at the population level

Enabling Platforms

- Binning
 - Trio binning (<https://www.nature.com/articles/nbt.4277>) – uses sequencing of trios to producing two reference-quality haplotypes from a single individual
 - Gamete binning (<https://www.biorxiv.org/content/10.1101/2020.04.24.060046v4>) – single-cell sequencing of hundreds of haploid gamete genomes to separate conventional long sequencing reads into two haplotype-specific read sets.
- Population scale sequencing methods (<https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13192>)
 - Using the same sequencing efforts as used in standard resequencing studies, linked-read sequencing can provide valuable phasing information