OAK RIDGE National Laboratory

Chestnut Chat GWAS 101

David Kainer

Computational Biology,
Biosciences Division, ORNL

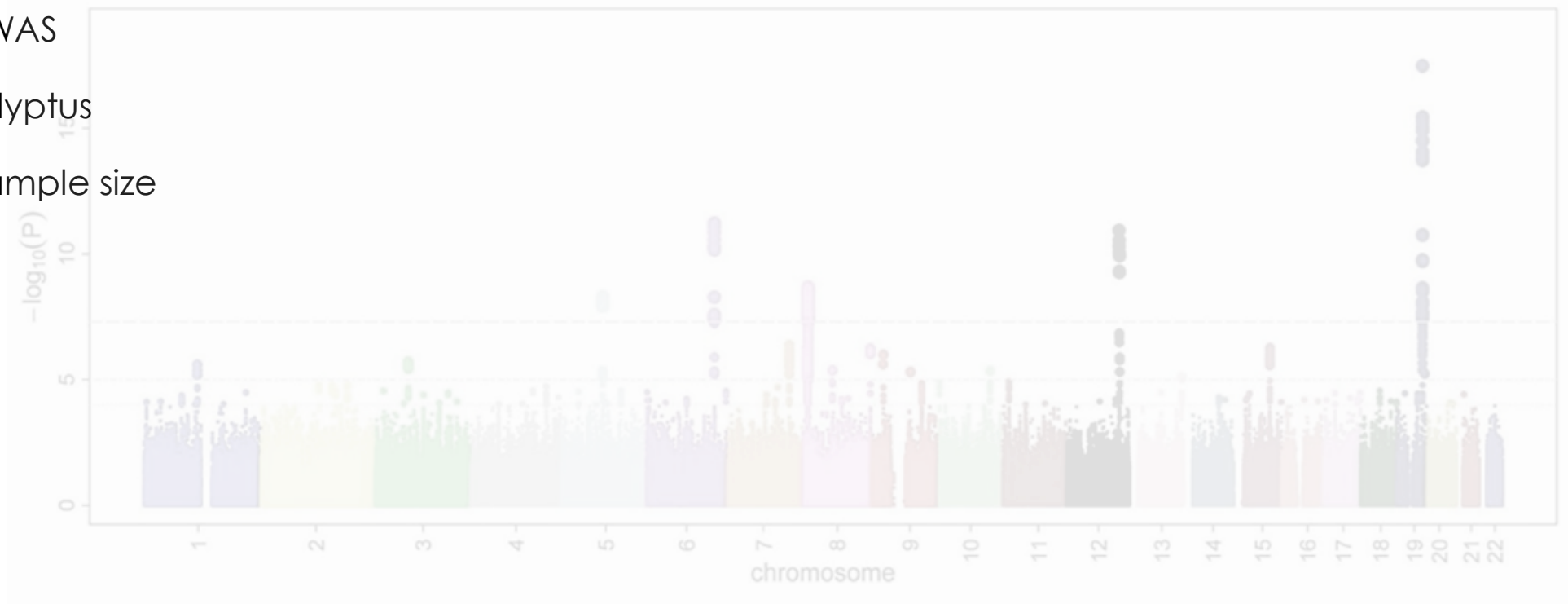
kainerd@ornl.gov

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

GWAS 101

1. Overview of GWAS
2. What kind of traits can be analyzed?
3. Phenotypes and Genotypes
4. The three steps of GWAS
5. A case study in Eucalyptus
6. The importance of sample size
7. Population Structure
8. Interpreting results



OAK RIDGE National Laboratory

GWAS according to the NIH

A genome-wide association study (GWAS) is an approach used in genetics research to associate specific genetic variations with particular diseases.

The method involves scanning the genomes from many different people and looking for genetic markers that can be used to predict the presence of a disease. Once such genetic markers are identified, they can be used to understand how genes contribute to the disease and develop better prevention and treatment strategies.

GWAS according to me

Phenotype = Genotype + Environment

GWAS for pathogen resistance in trees

Poplar

PNAS

Association mapping, transcriptomics, and transient expression identify candidate genes mediating plant–pathogen interactions in a tree

Wellington Muchero[a], Kelsey L. Sondreli[b], Jin-Gui Chen[a], Breeanna R. Urbanowicz[c], Jin Zhang[a], Vasanth Singan[d], Yongil Yang[a], Robert S. Brueggeman[e], Juan Franco-Coronado[e], Nivi Abraham[e], Jeong-Yeh Yang[c], Kelley W. Moremen[c], Alexandra J. Weisberg[b], Jeff H. Chang[b], Erika Lindquist[d], Kerrie Barry[d], Priya Ranjan[a], Sara Jawdy[a], Jeremy Schmutz[d,f], Gerald A. Tuskan[a,d], and Jared M. LeBoldus[b,e,g,1]

[a]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831; [b]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331; [c]Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602; [d]Joint Genome Institute, US Department of Energy, Walnut Creek, CA 94598; [e]Department of Plant Pathology, North Dakota State University, Fargo, ND 58102; [f]HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806; and [g]Forest Engineering, Resources, and Management, Oregon State University, Corvallis, OR 97331

Ash

ARTICLES
https://doi.org/10.1038/s41559-019-1036-6
nature ecology & evolution

Genomic basis of European ash tree resistance to ash dieback fungus

Jonathan J. Stocks[1,2], Carey L. Metheringham[1,2], William J. Plumb[1,2,3], Steve J. Lee[4], Laura J. Kelly[1,2], Richard A. Nichols[1] and Richard J. A. Buggs[1,2*]

Eucalyptus

A Genome-Wide Association Study for Resistance to the Insect Pest *Leptocybe invasa* in *Eucalyptus grandis* Reveals Genomic Regions and Positional Candidate Defense Genes

Lorraine Mhoswa[1], Marja M. O'Neill[1], Makobatjatji M. Mphahlele[1,2], Caryn N. Oates[1], Kitt G. Payn[3], Bernard Slippers[1], Alexander A. Myburg[1] and Sanushka Naidoo [1,*]

Beech

RESEARCH ARTICLE Open Access

Genome-wide association study identifies a major gene for beech bark disease resistance in American beech (*Fagus grandifolia* Ehrh.)

Irina Ćalić[1], Jennifer Koch[2], David Carey[2], Charles Addo-Quaye[3,4], John E. Carlson[5] and David B. Neale[1*]

Eucalyptus

New Phytologist Research

Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in *Eucalyptus*

Rafael Tassinari Resende[1], Marcos Deon Vilela Resende[1,2], Fabyano Fonseca Silva[3], Camila Ferreira Azevedo[1], Elizabete Keiko Takahashi[4], Orzenil Bonfim Silva-Junior[5,6] and Dario Grattapaglia[5,6]

GWAS for wood properties

Phenotype	Species	Population	Sample size	No. of markers	Method	Reference
Growth and wood properties	*Eucalyptus globulus*	Families and bulk collections	303	7,680 [Diversity Array Technology markers (DArT)]	General linear model (GLM) and unified mixed model (UMM)	Cappa et al., 2013
Wood density, stiffness, microfibril angle, and ring width	*Picea glauca*	Open-pollinated families	1694	7434 (SNPs)	Mixed linear model (MLM)	Lamara et al., 2016
16 wood chemistry/ ultrastructure traits	*Populus trichocarpa*	Unrelated individuals	334	29,233 (SNPs)	GLM	Porth et al., 2013
Lignin percentage, Lignin S:G ratio, 5-carbon sugars, and 6-carbon sugars	*Populus deltoides*	Unrelated individuals	391	334,679 (consensus SNPs), 185,526 (Common SNPs), 76,804 (functional SNPs)	Single-variant and multiple-variant associations on GLM	Fahrenkrog et al., 2017
Basic wood density (BWD), bleached pulp, pulp yield (SPY), and pulp bleaching content	*Eucalyptus grandis* × *Eucalyptus urophylla*	Hybrid breeding population	768	24 806 (SNPs)	GWAS and regional heritability mapping	Resende et al., 2017
17 wood-quality traits	Norway spruce	Mother trees	517	178101 (SNPs)	Multilocus LASSO penalized regression	Baison et al., 2018
Seven wood properties	*Populus tomentosa*	Unrelated individuals	435	5,482 (InDels)	MLM and Kempthorne model	Gong et al., 2017

Diversity Array Technology (DArT) markers.

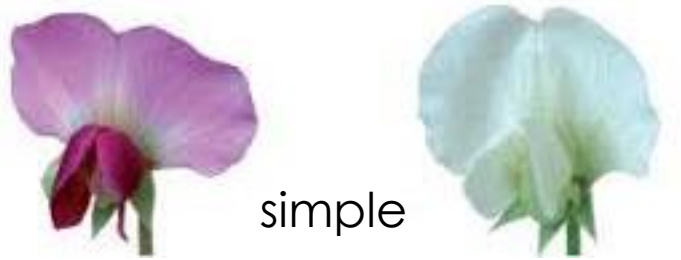
Du, Qingzhang, et al. "Genome-wide association studies to improve wood properties: challenges and prospects." *Frontiers in Plant Science* 9 (2018): 1912.

You can GWAS almost anything you can measure

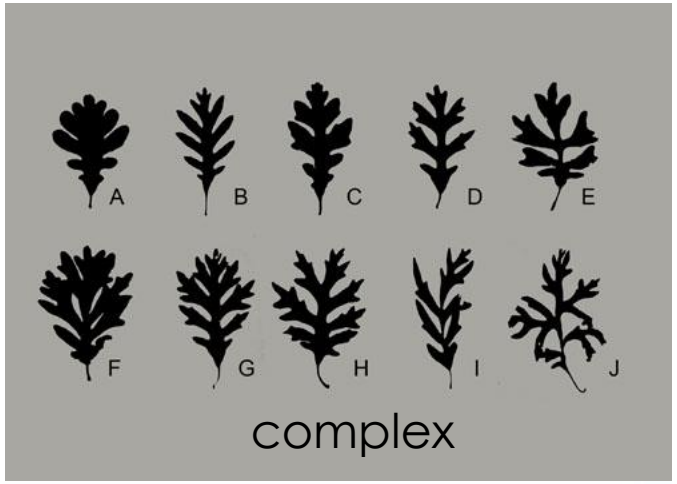
Typical GWAS traits in trees

- Growth: Height, DBH, crown architecture, branching, leaf shape
- Wood properties: lignin, sugars, microfibril angle
- Metabolic: primary or secondary metabolite abundances
- Pathogens: Fungal/bacterial abundances
- Resistance: inoculation outcomes
- Adaptive: flowering time, leaf senescence
- Sustainability: water use efficiency, nitrogen use
- Gene Expression: Gene Transcript (RNAseq) abundances

Phenotypic variation


simple

- Without variation there is little to explore
- Variation is the source material
- Is the phenotypic variation due to:
  - Genetic variation?
  - Environmental variation?
  - Both?

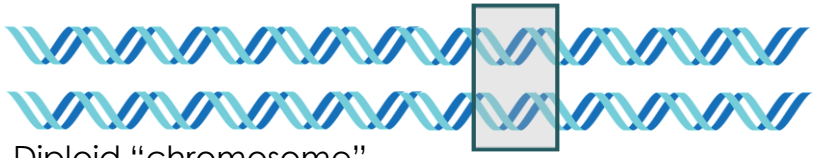


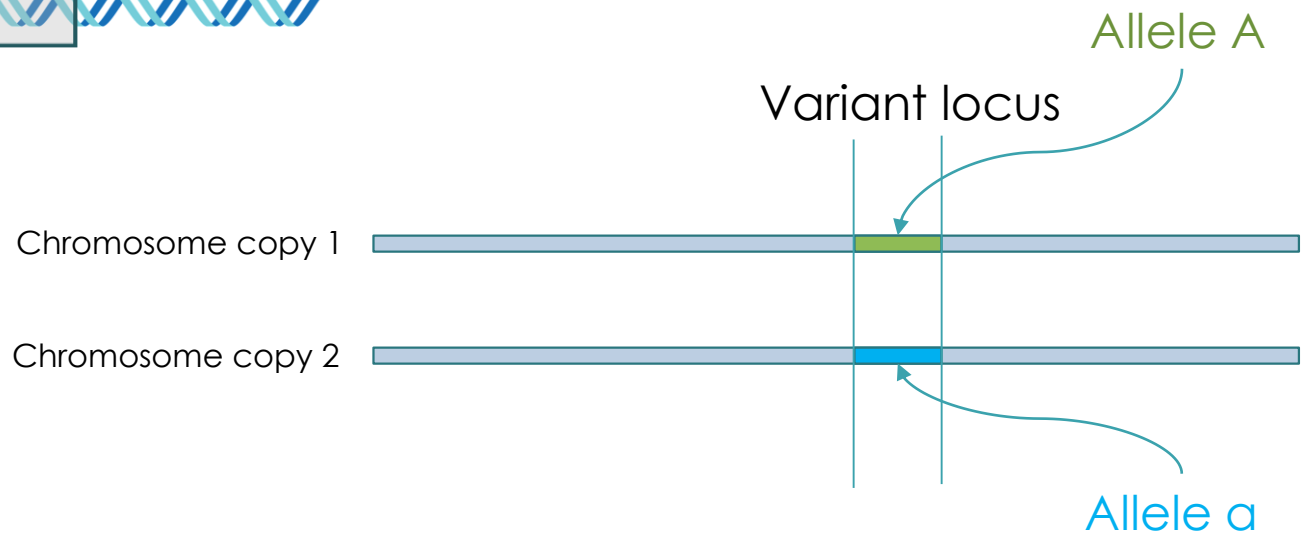
complex

Phenotype = Genotype + Environment

minimal

genetic variation

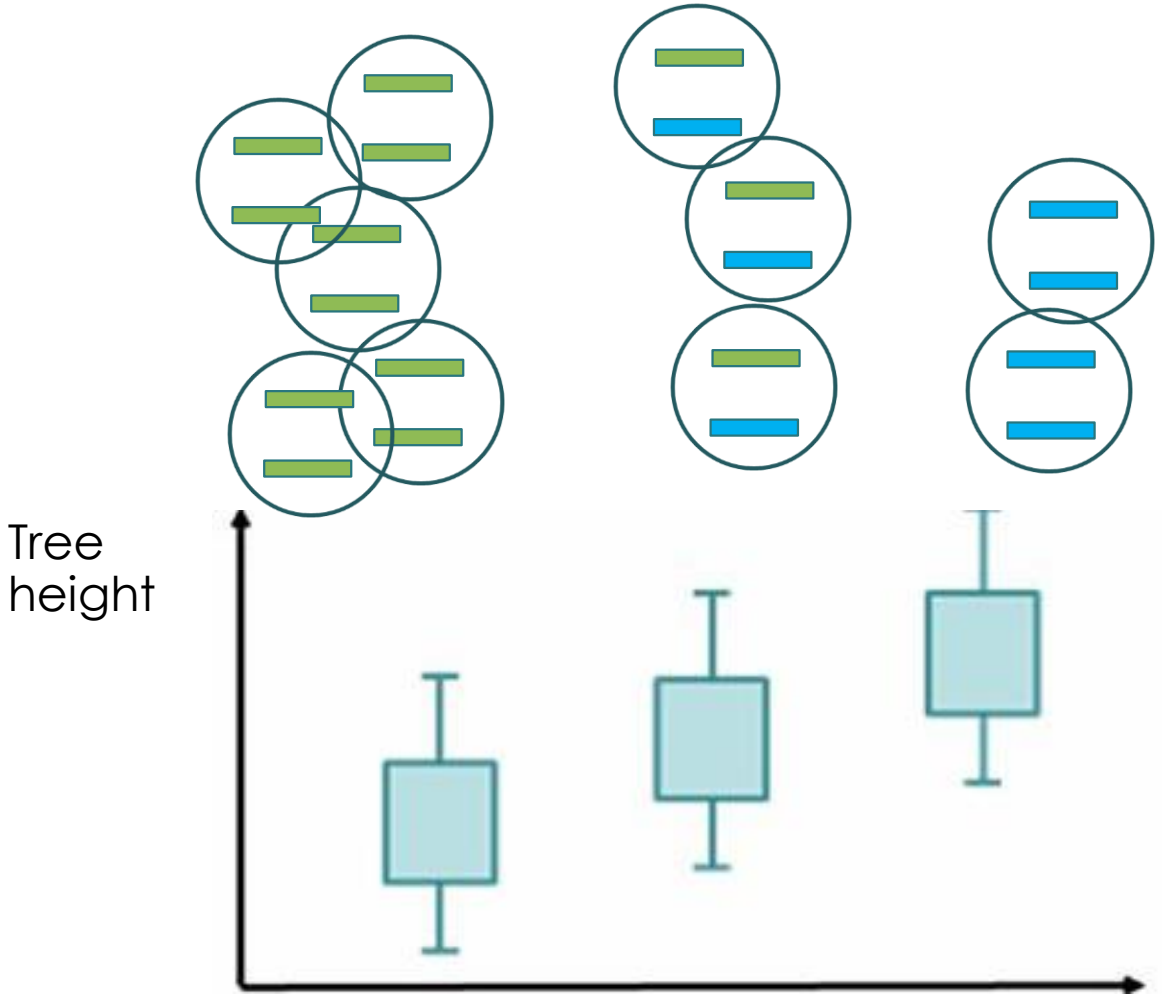




An individual will have:

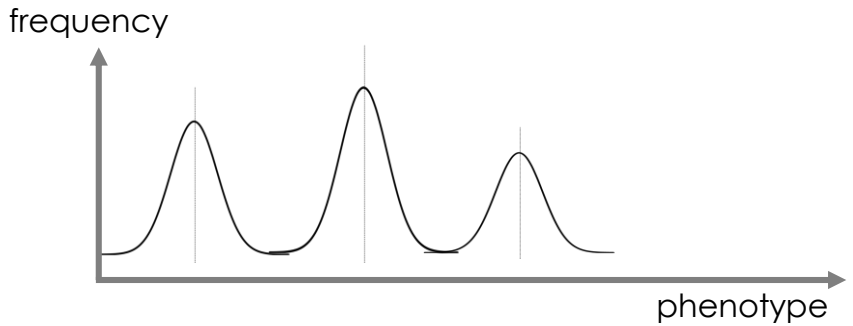
Which **may or may not affect their phenotype** (e.g. +/- for quantitative trait, or increased resistance to a disease)

genetic variation in a population



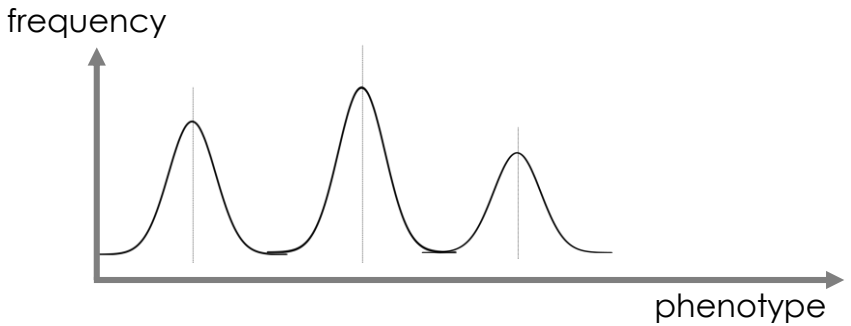
Population Genetic and Phenotypic variation

1 effective variant



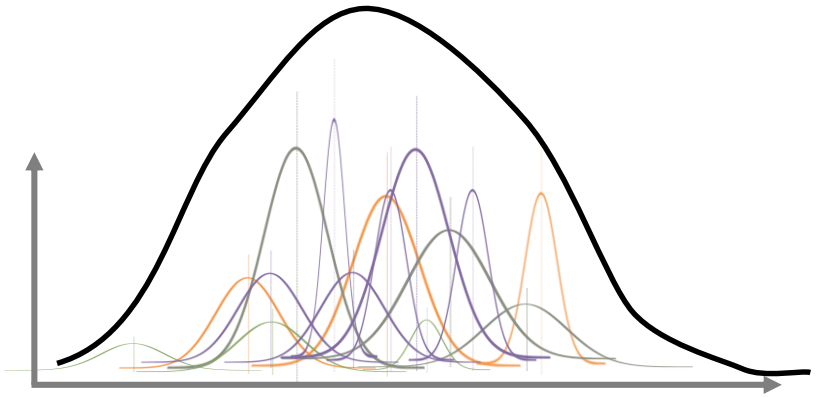
simple

Population Genetic and Phenotypic variation



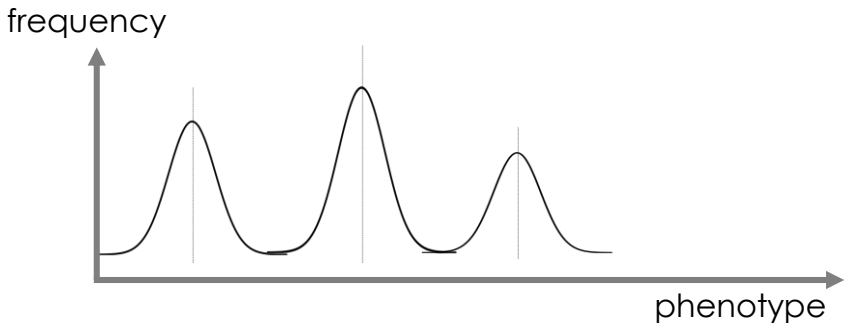
simple

Many effective variants



complex

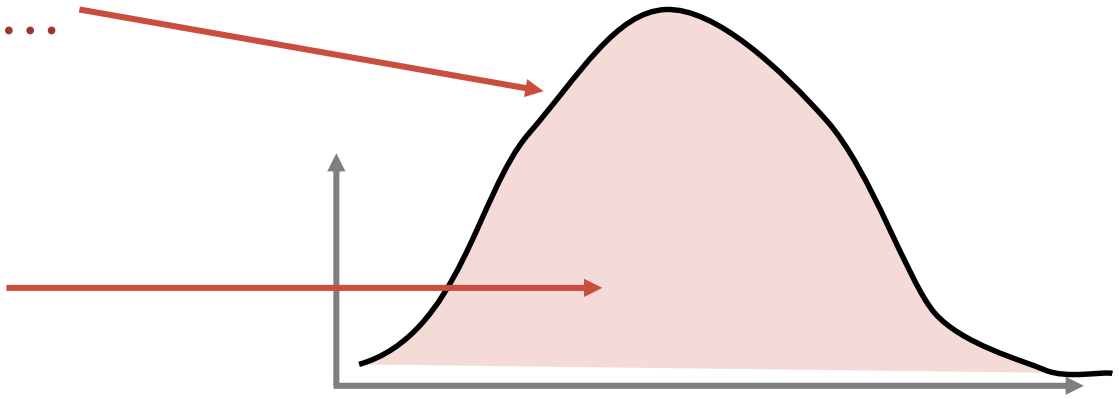
Population Genetic and Phenotypic variation



simple

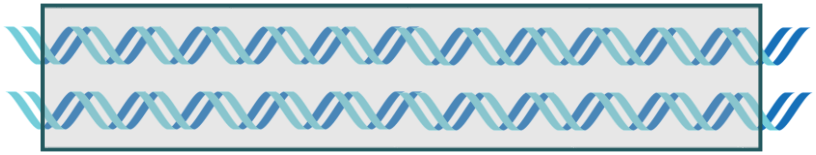
We observe this…

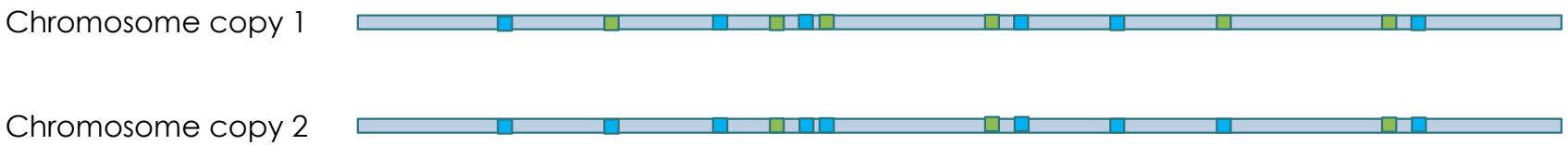
We don't know what goes in here!!!



complex

Genome-wide variation

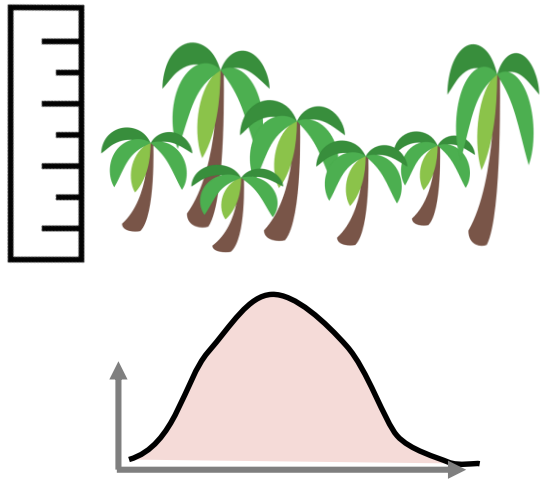




Across the entire genome, there may be millions of variant locations (e.g. SNPs) that have alleles in a population.

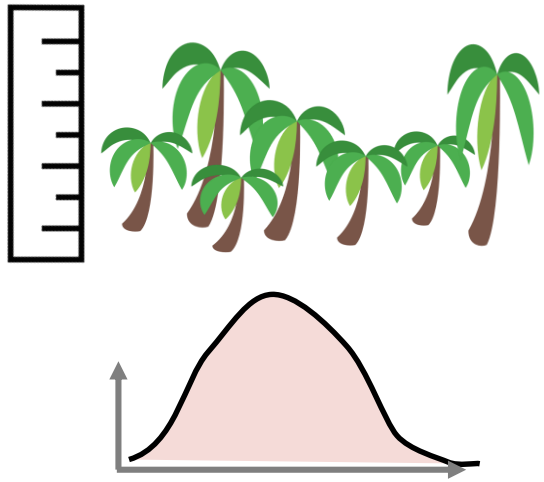
Q: Which ones have an effect on a trait ?
A: GWAS

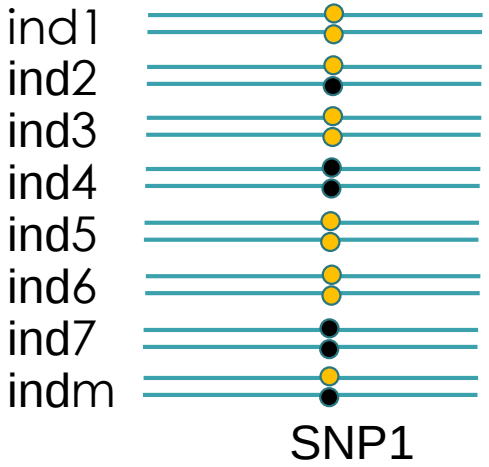
How a GWAS works: quantitative trait



1. Take a population of mostly un-related individuals. Measure a phenotype that varies (e.g. Height)

How a GWAS works: quantitative trait

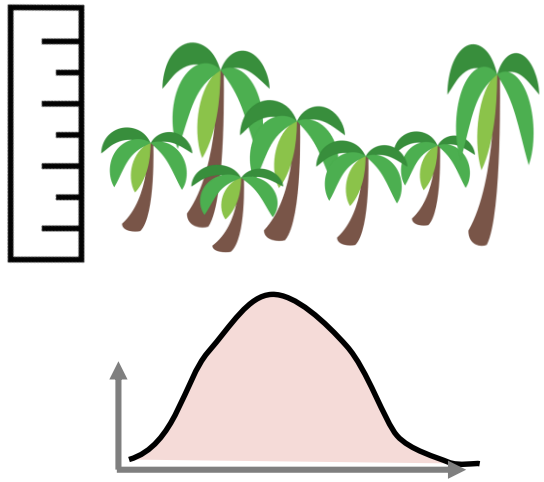


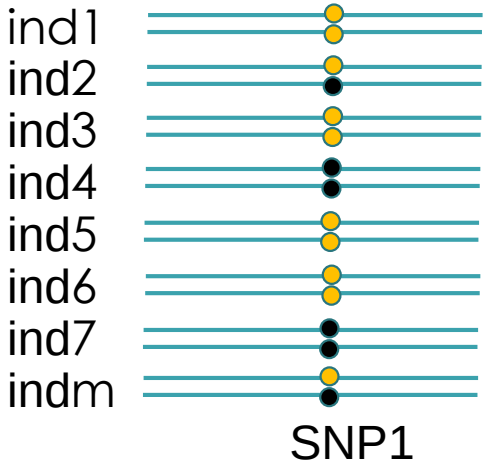


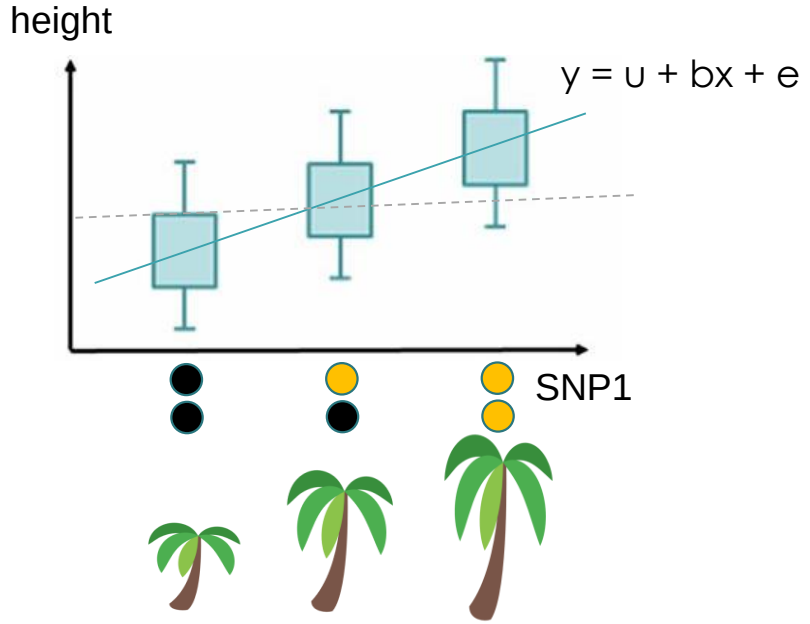
1. Take a population of mostly un-related individuals. Measure a phenotype that varies (e.g. Height)

2. Sequence the DNA of each of them. Find positions in the genome where the individuals vary (e.g. SNPs)

How a GWAS works: quantitative trait





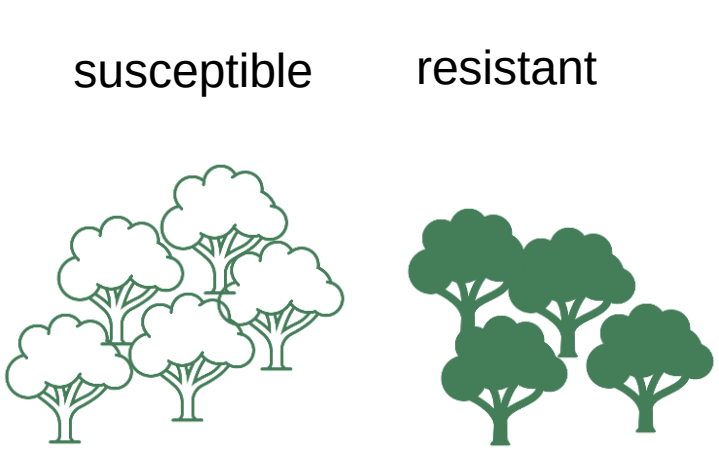


1. Take a population of mostly un-related individuals. Measure a phenotype that varies (e.g. Height)

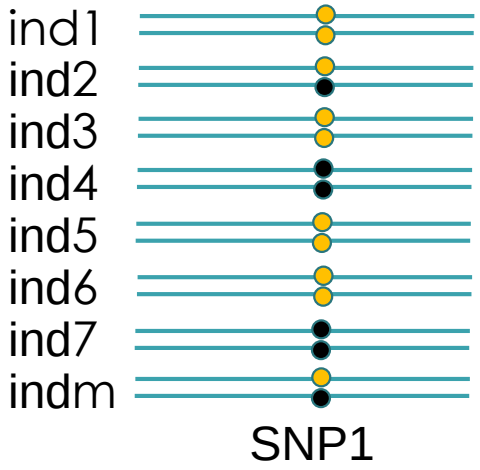
2. Sequence the DNA of each of them. Find positions in the genome where the individuals vary (e.g. SNPs)

3. Test each SNP to see if the alleles significantly correlate with the phenotypic variation. E.g. Does having more copies of the ● allele equate to a significant linear increase in height?

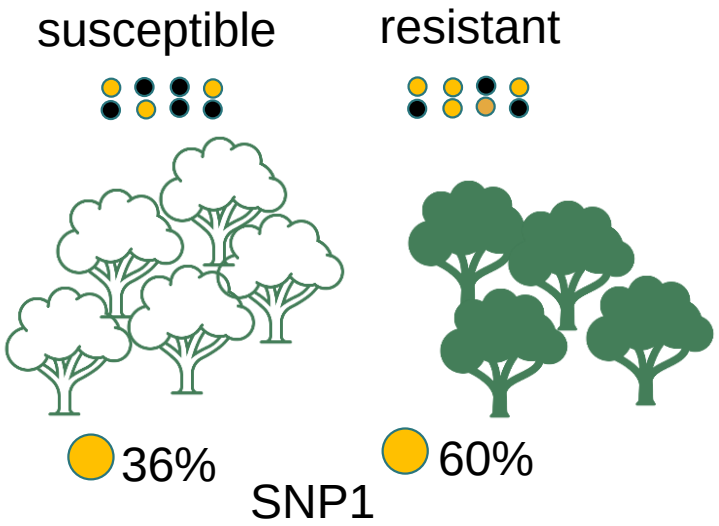
How a GWAS works: binary trait



Take a population of mostly un-related individuals. Split them into two groups (e.g. resistant / susceptible)

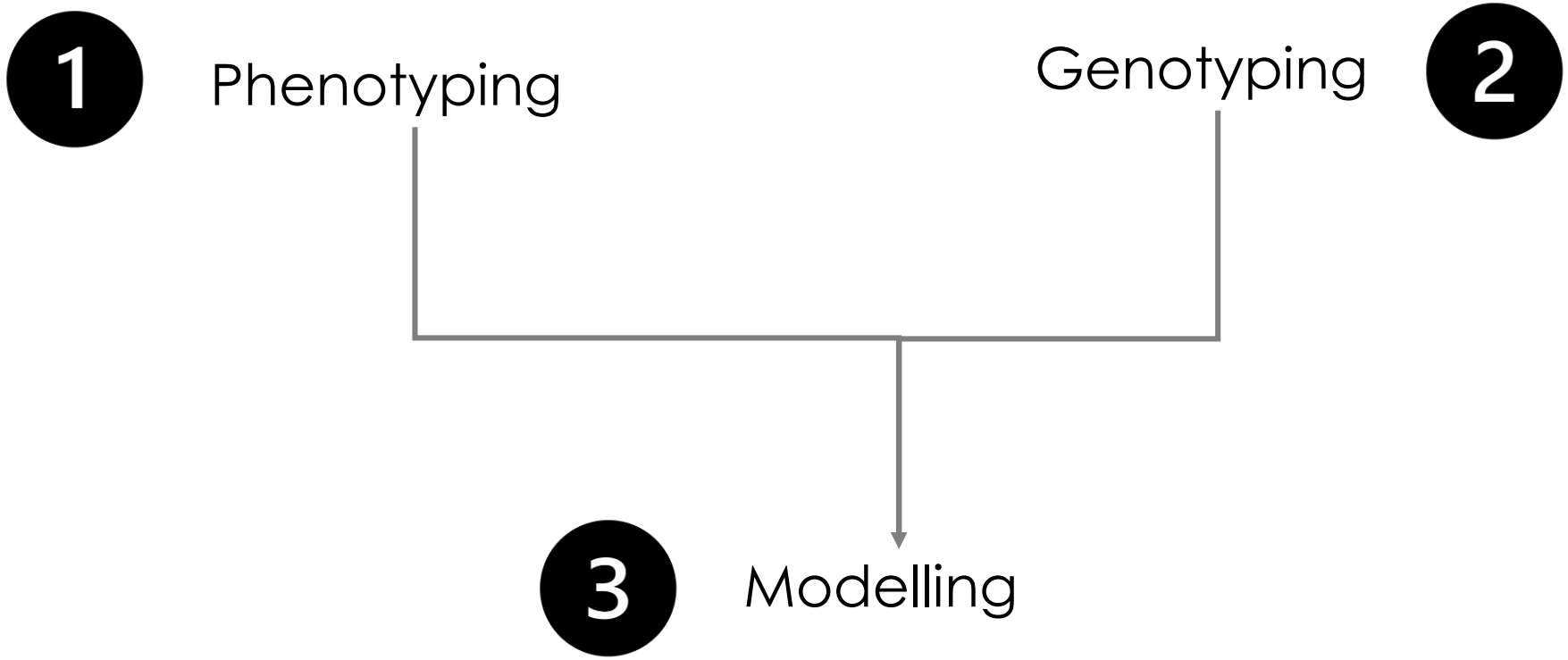


Sequence the DNA of each of them.
Find positions in the genome where the individuals vary (e.g. SNPs)

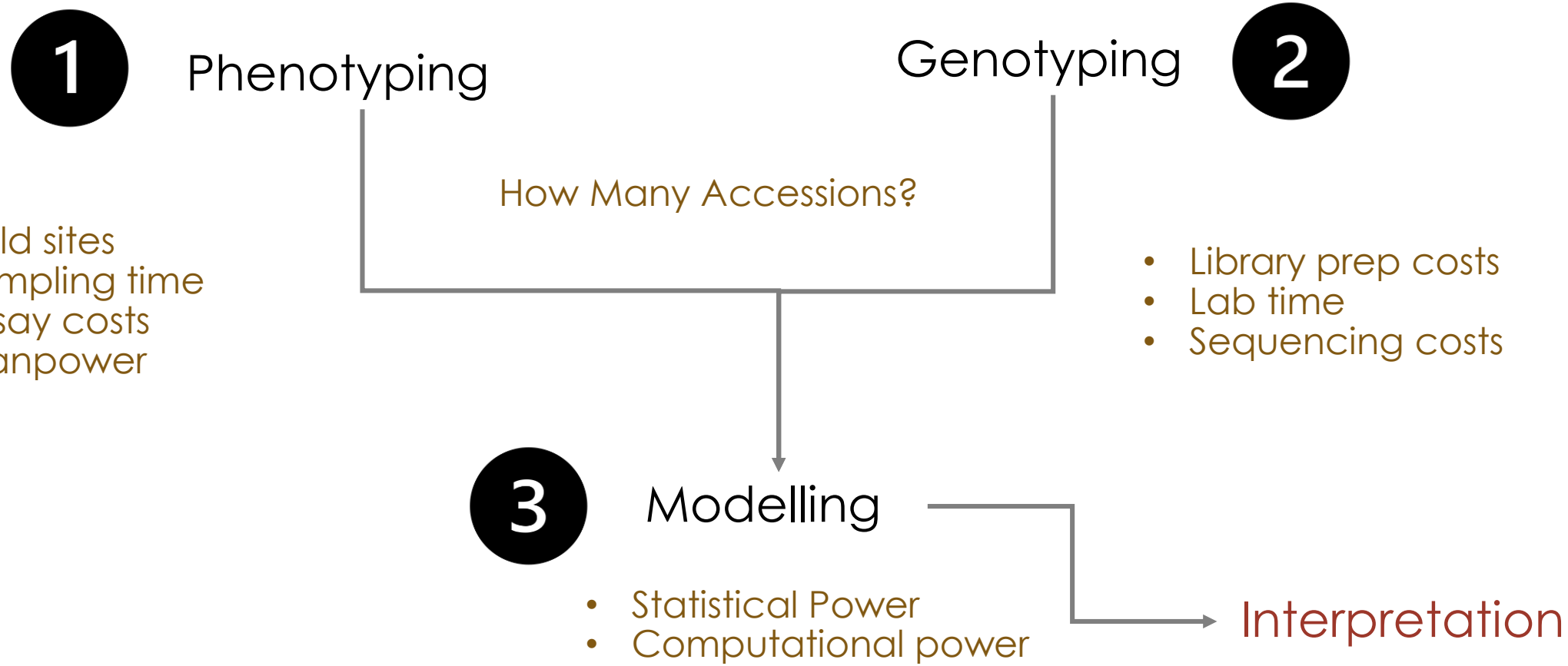


Test each SNP to see if its alleles have significantly different frequency in one group compared to the other

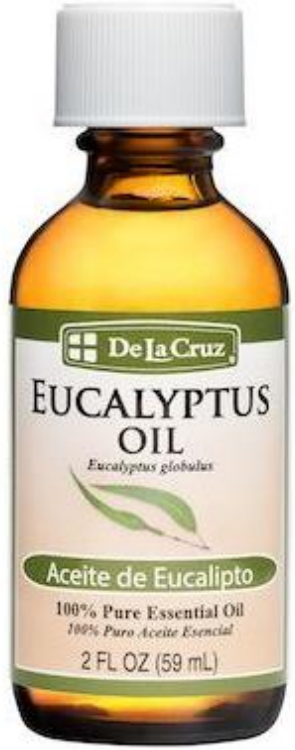
Three main steps for GWAS



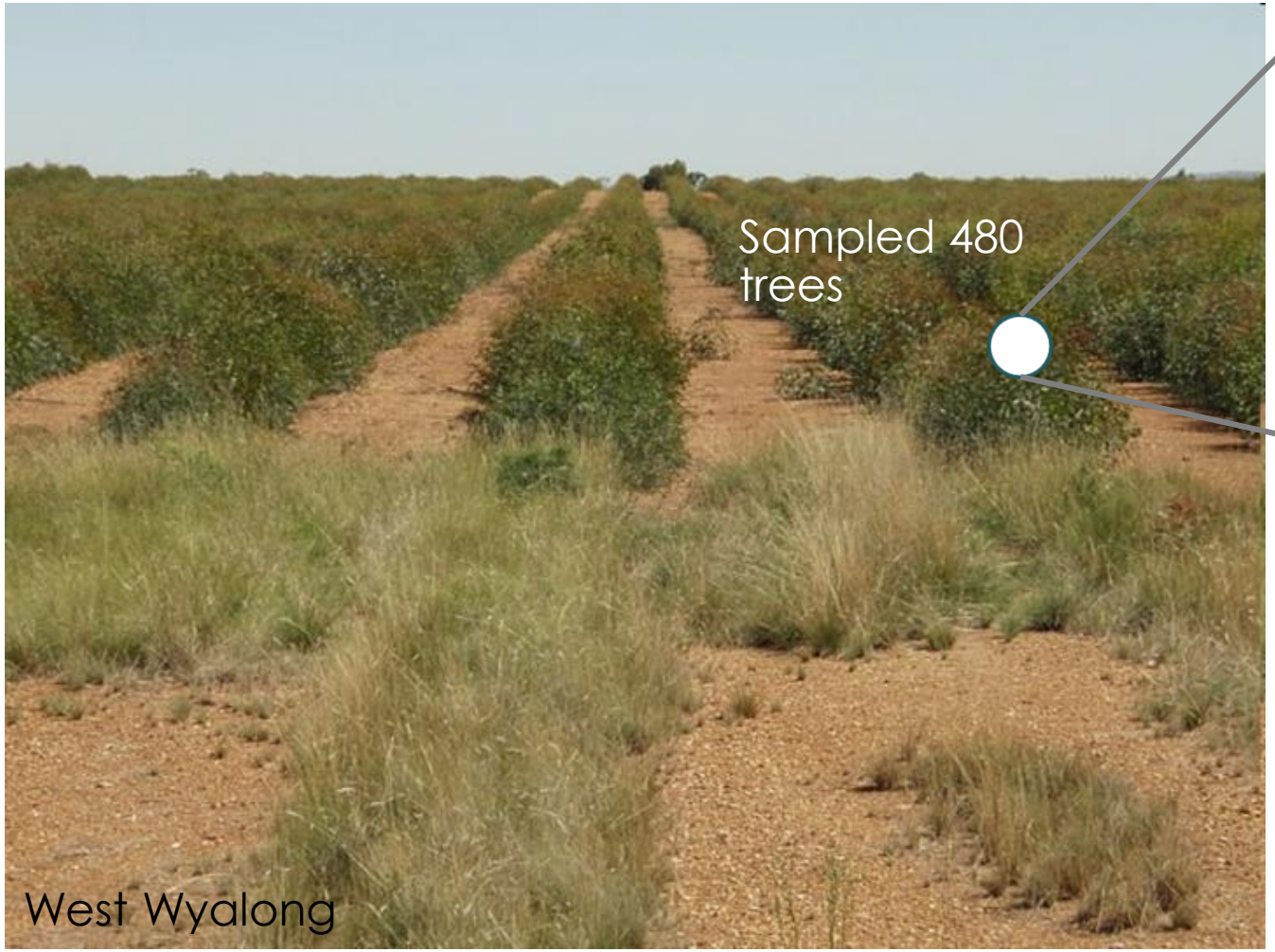
Three main steps for GWAS



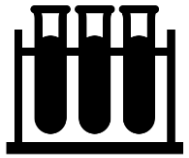
GWAS for oil traits in *Eucalyptus polybractea*



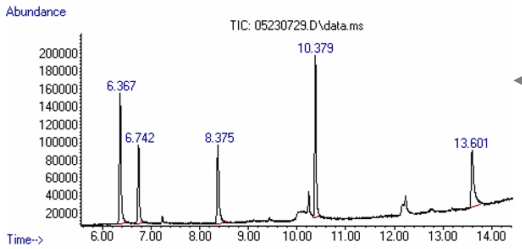
1 Phenotyping



Extract oil in ethanol



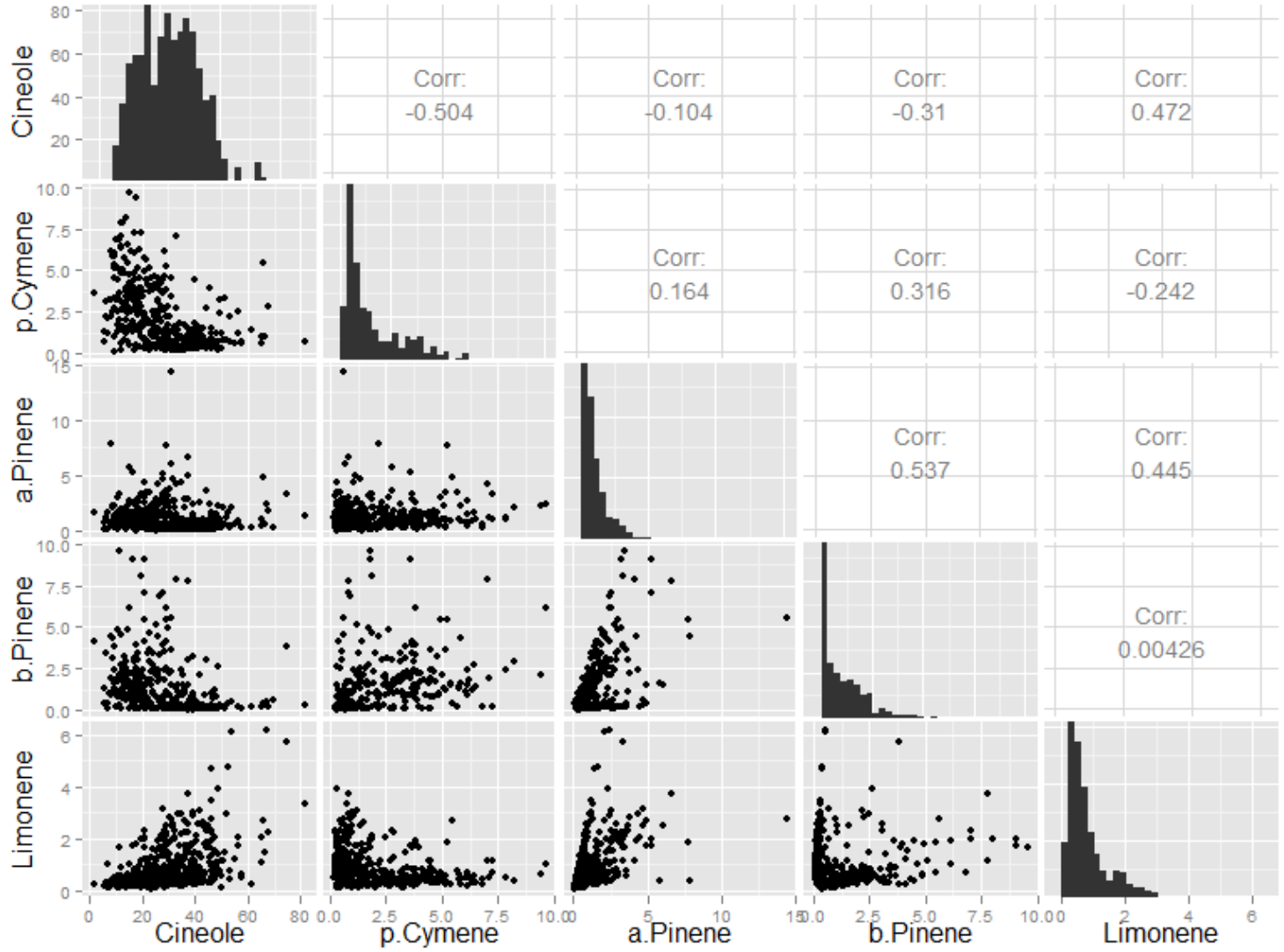
GC-MS



metabolite profiles

	M1	M2	M3	Mn
S1	0	300	12	65
S2	20	500	90	0
S3	2	50	250	0
S480	120	100	15	0

1 Phenotyping



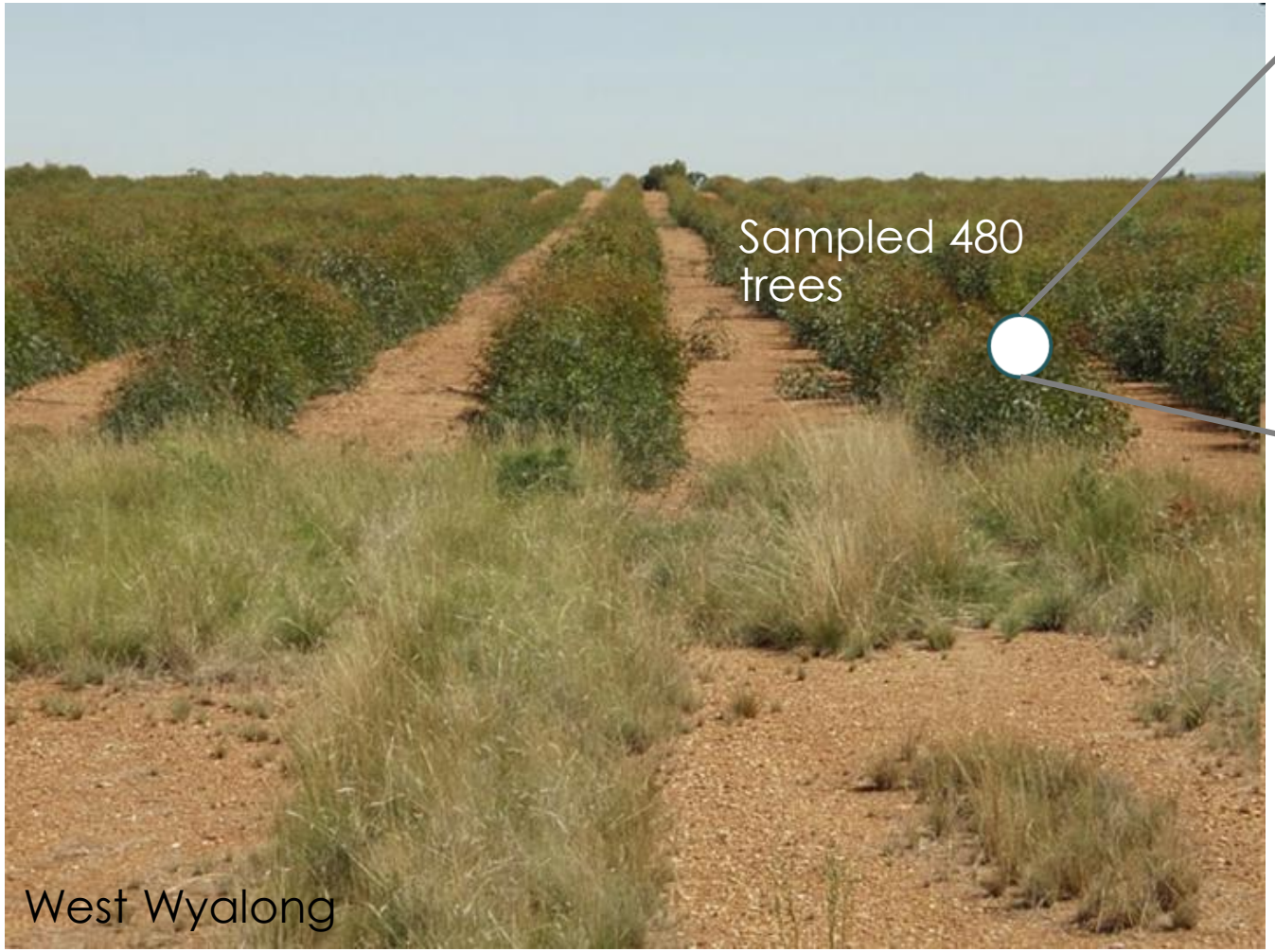
OAK RIDGE National Laboratory

Phenotyping pitfall – environmental variance

1 Phenotyping

- Minimize environmental variance
- Randomization is necessary to avoid batch effects or environmental trends
- Accuracy is hugely important
- consistency

2 Genotyping



Snap freeze with liquid Nitrogen

Extract DNA

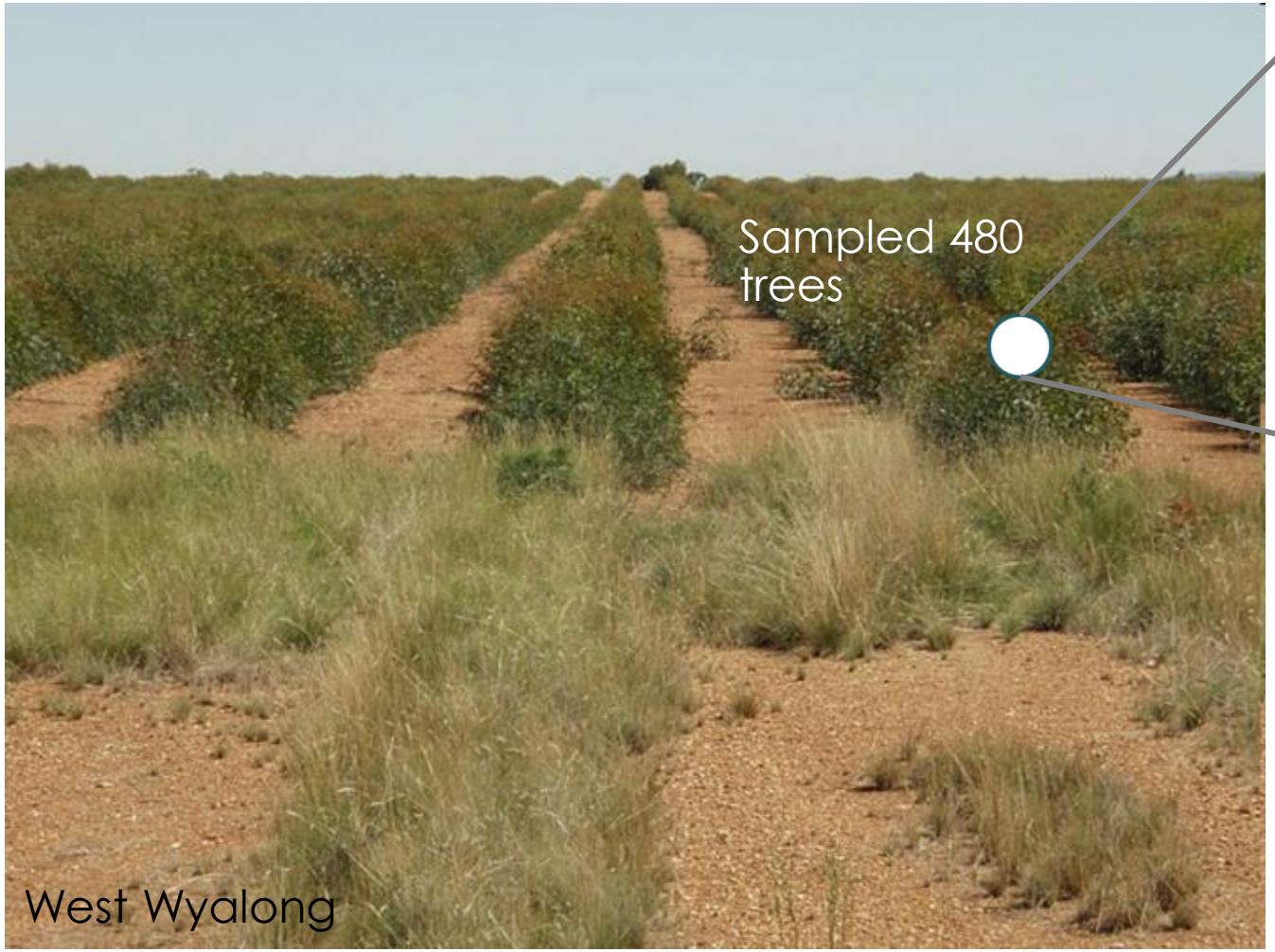
sequencing

Detect variants

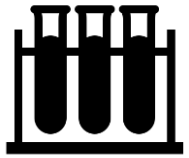
Genotype matrix

	snp1	snp2	snp3	snpP
S1	A/A	C/T	A/T	G/G
S2	A/C	T/T	A/A	G/G
S3	A/A	C/C	A/T	G/G
S480	A/A	C/C	A/T	T/T

2 Genotyping



Snap freeze with liquid Nitrogen



Extract DNA

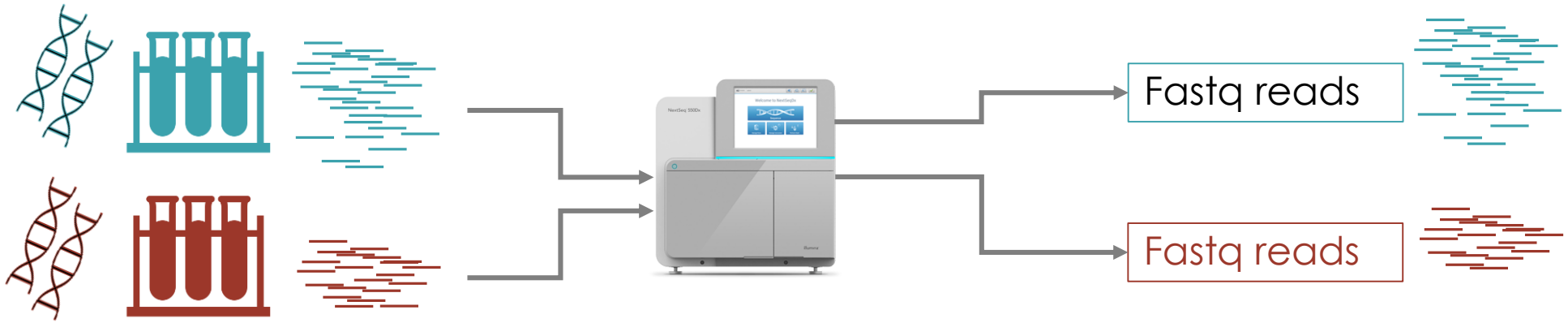
Detect variants

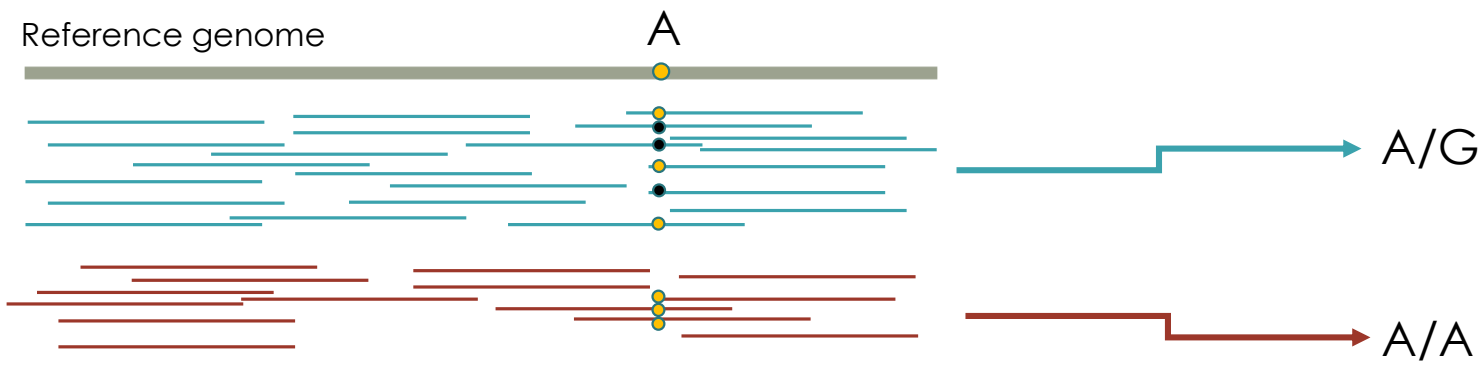
Genotype matrix

	snp1	snp2	snp3	snpP
S1	A/A	C/T	A/T	G/G
S2	A/C	T/T	A/A	G/G
S3	A/A	C/C	A/T	G/G
Sm	A/A	C/C	A/T	T/T

Open slide master to edit

2 Genotyping





OAK RIDGE National Laboratory

Genotyping pitfall: More accessions or more sequencing depth?

N accessions

Or…

2N accessions

- **High** sequencing depth per sample
- Higher genotype accuracy
- Lower genetic diversity
- Lower statistical power

- **low** depth per sample
- Lower genotype accuracy
- Higher genetic diversity
- Higher statistical power

3 modeling

metabolite matrix

	M1	M2	M3	Mn
S1	0	300	12	65
S2	20	500	90	0
S3	2	50	250	0
S480	120	100	15	0

Genotype matrix

	snp1	snp2	snp3	snpP
S1	A/A	C/T	A/T	G/G
S2	A/C	T/T	A/A	G/G
S3	A/A	C/C	A/T	G/G
S480	A/A	C/C	A/T	T/T

	snp1	snp2	snp3	snpP
S1	0	1	1	0
S2	1	2	0	0
S3	0	0	1	0
S480	0	0	1	2

3 modeling

metabolite matrix

	M1	M2	M3	Mn
S1	0	300	12	65
S2	20	500	90	0
S3	2	50	250	0
S480	120	100	15	0

Genotype matrix

	snp1	snp2	snp3	snpP
S1	A/A	C/T	A/T	G/G
S2	A/C	T/T	A/A	G/G
S3	A/A	C/C	A/T	G/G
Sm	A/A	C/C	A/T	T/T

	snp1	snp2	snp3	snpP
S1	0	1	1	0
S2	1	2	0	0
S3	0	0	1	0
Sm	0	0	1	2

Filtering:

- Remove rare SNPs
- Remove SNPs that are likely to be erroneous

3 modeling

metabolite matrix

	M1	M2	M3	Mn
S1	0	300	12	65
S2	20	500	90	0
S3	2	50	250	0
S480	120	100	15	0

Genotype matrix

	snp1	snp2	snp3	snpP
S1	A/A	C/T	A/T	G/G
S2	A/C	T/T	A/A	G/G
S3	A/A	C/C	A/T	G/G
Sm	A/A	C/C	A/T	T/T

	snp1	snp2	snp3	snpP
S1	0	1	1	0
S2	1	2	0	0
S3	0	0	1	0
Sm	0	0	1	2

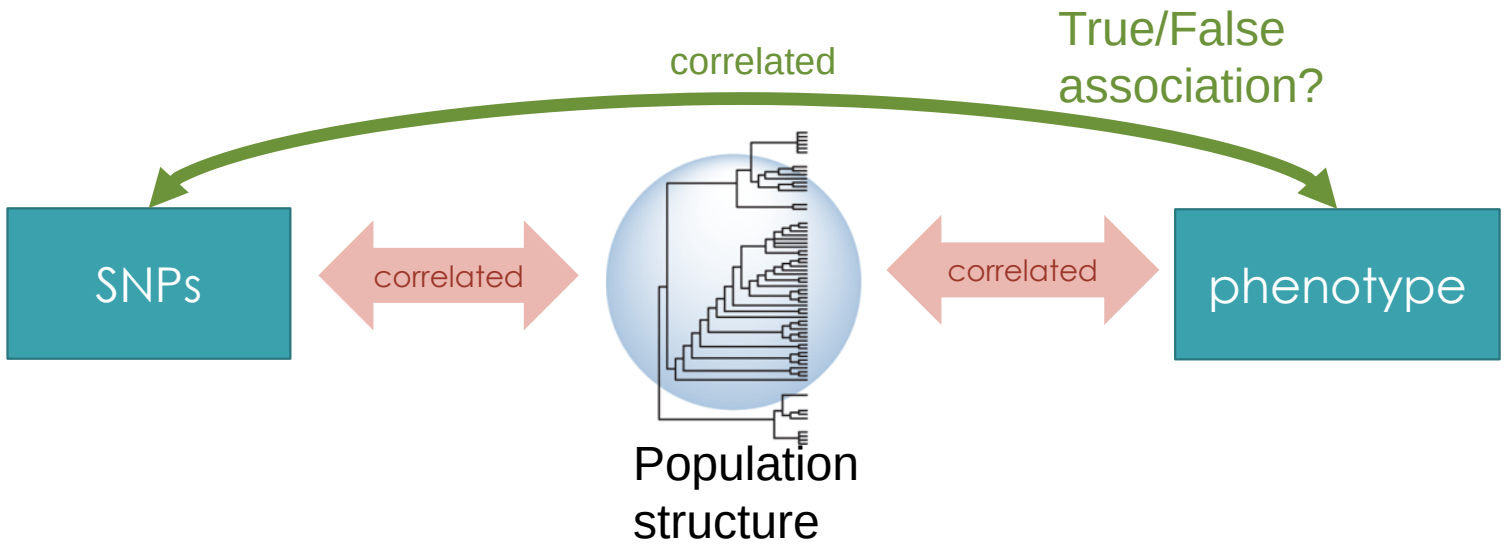
Linear model: $Y = u + b_1 x_1 + e \rightarrow$ P-value

Problem with our model

$$Y = u + b_i x_i + e$$

Population structure

- Samples are often genetically related to each other (even if you think they aren't!), which means they are not independent and can cause false positive SNP associations.

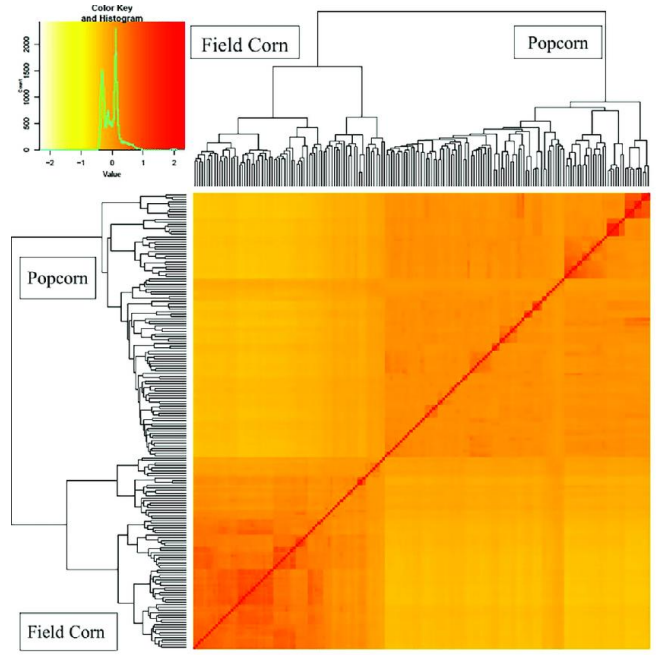


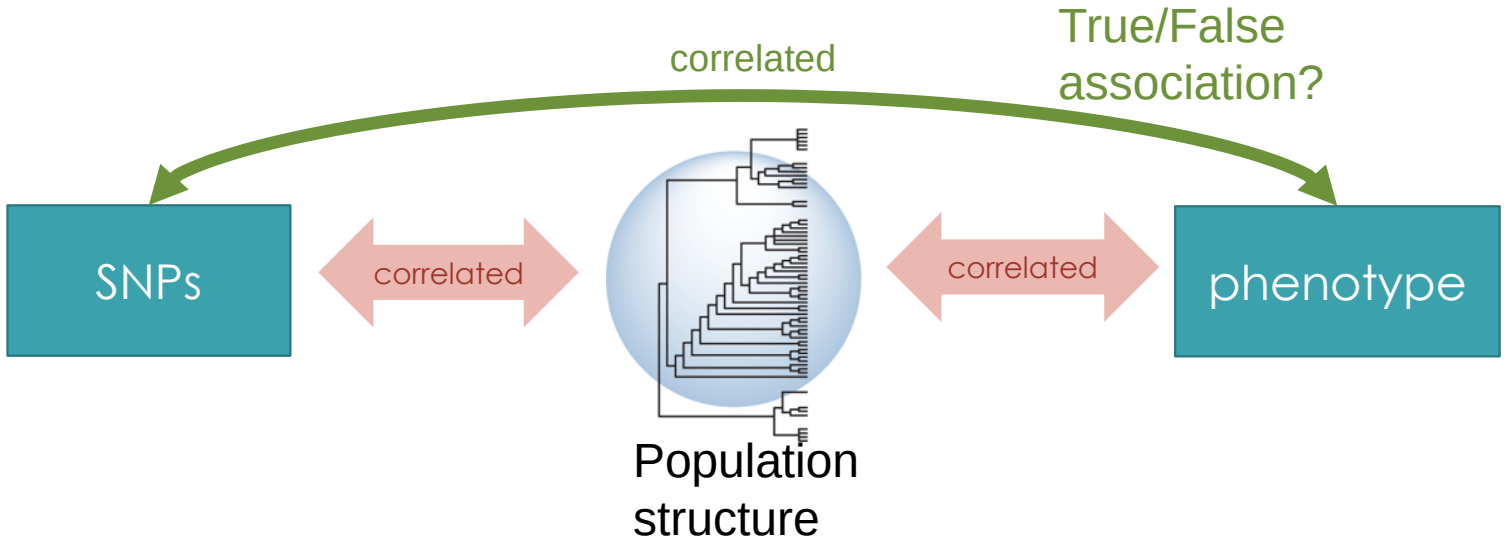
Problem with our model

$$Y = u + b_i x_i + k + e$$

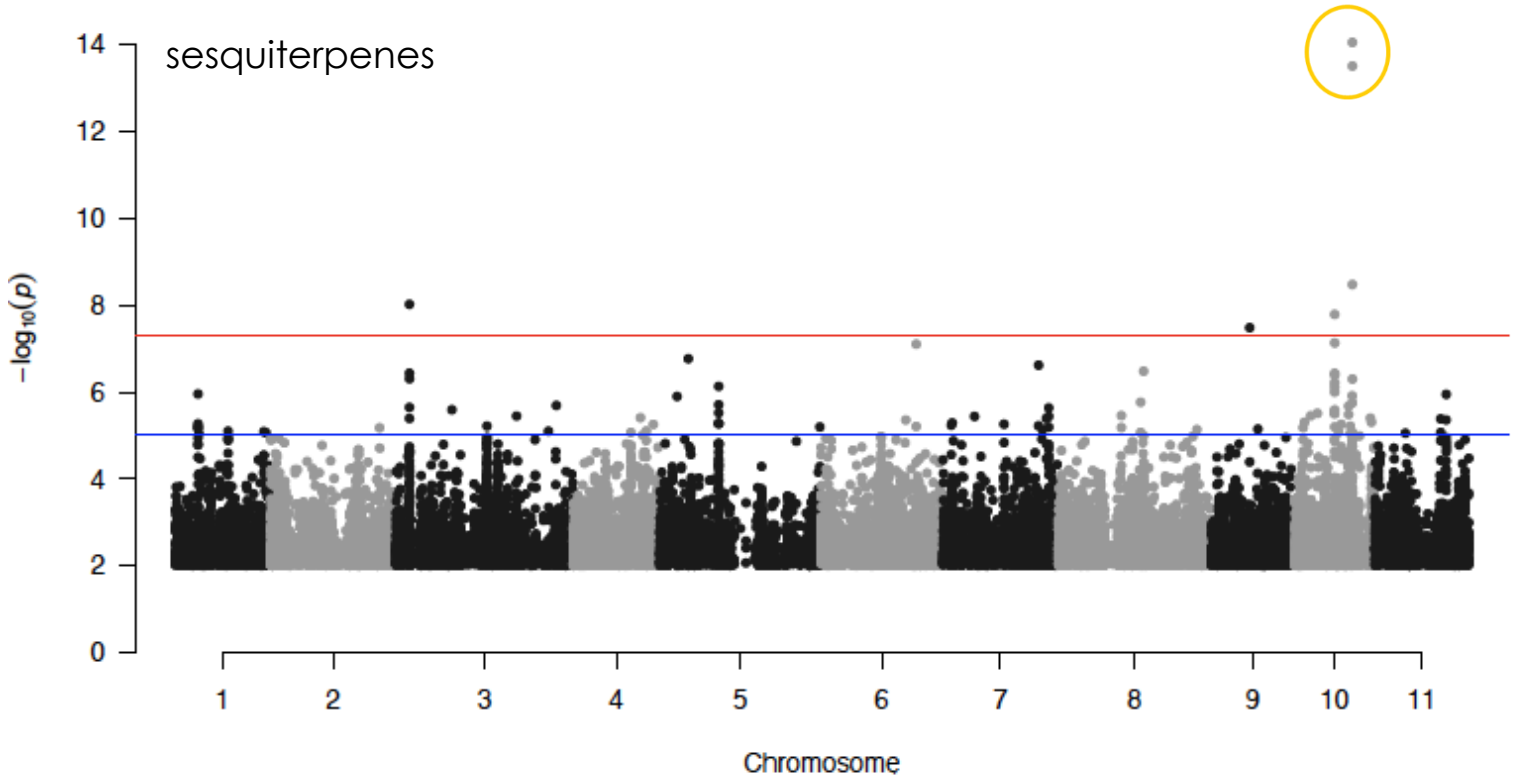
Population structure

- Samples are often genetically related to each other (even if you think they aren't!), which means they are not independent and can cause false positive SNP associations.





Results: Manhattan plot



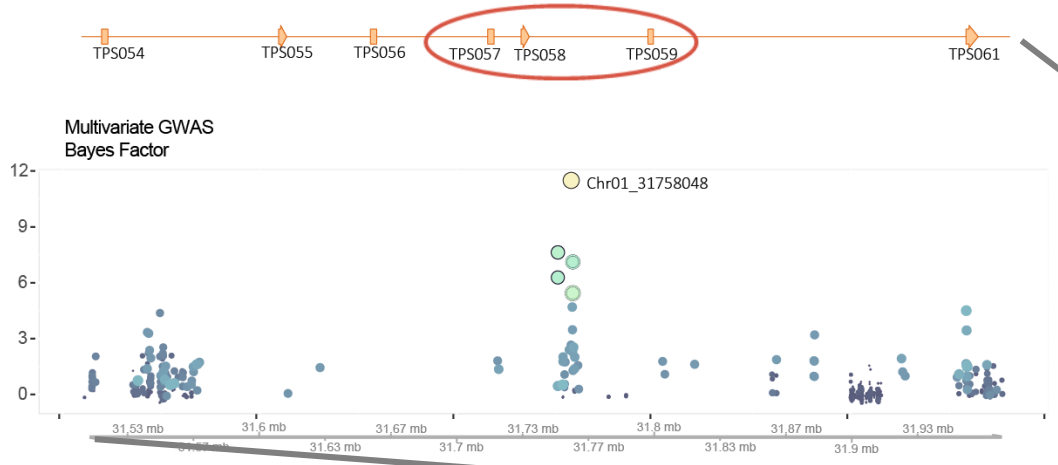
Multiple testing correction threshold

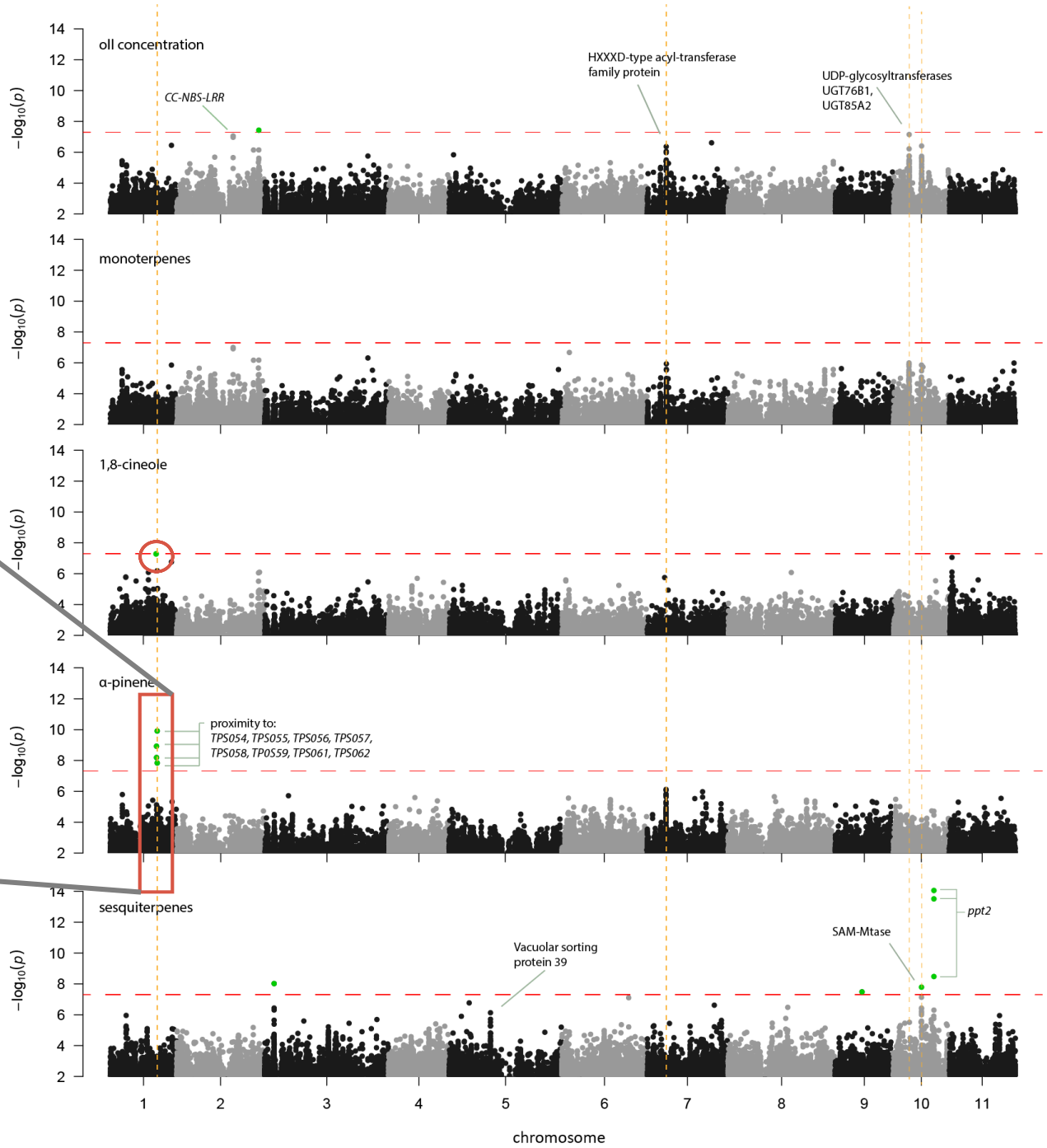
Results

OAK RIDGE National Laboratory

Results: interpretation

Candidate genes!

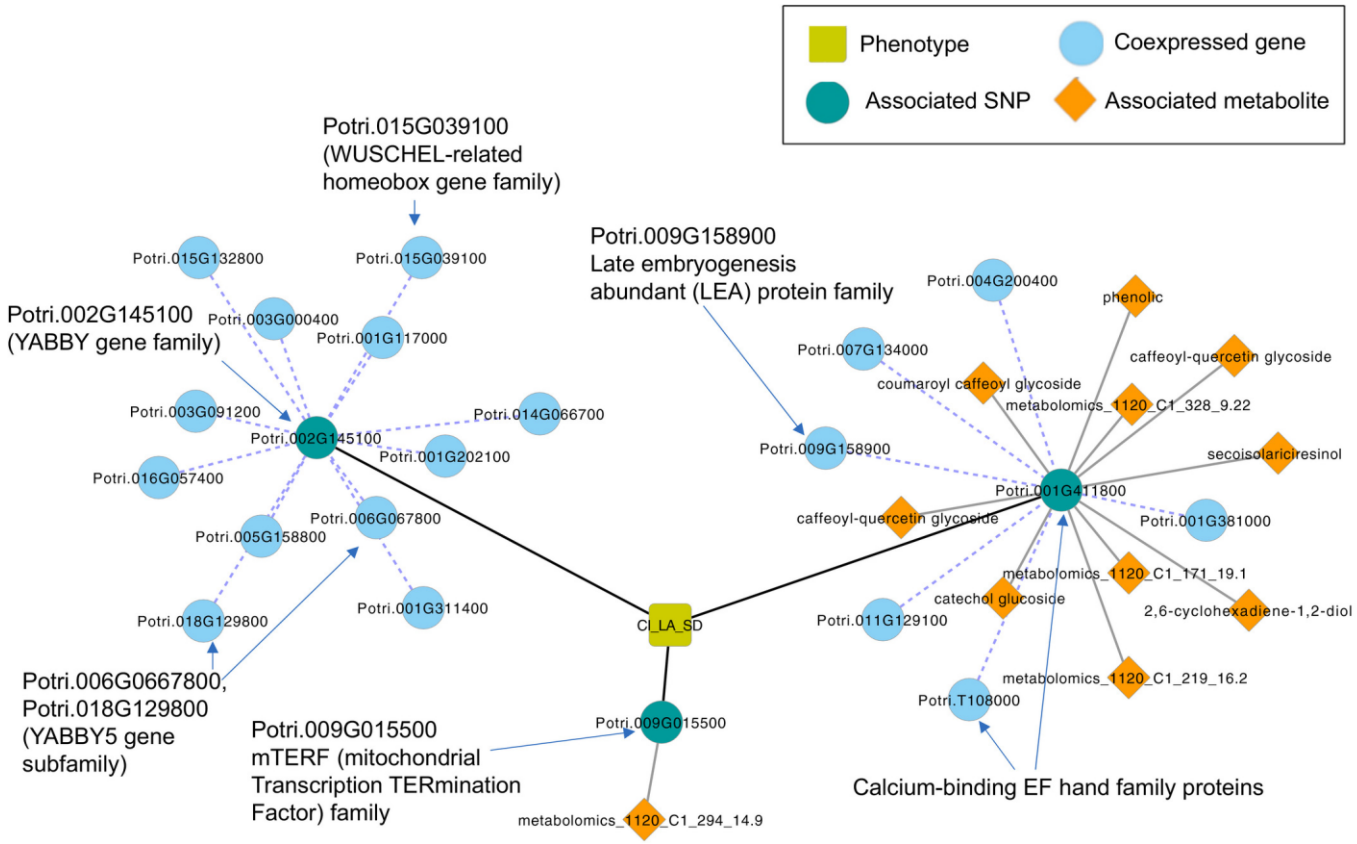




OAK RIDGE National Laboratory

Next Steps

- Validation !!
- Multi-omics integration
- Try other models
- Build the story



OAK RIDGE National Laboratory

Thank You

- Carsten Kulheim
- William Foley
- Carlos Bustos
- Amanda Padovan
- Daniel Jacobson
- Jerry Tuskan

Australian
National
University