

Associative Transcriptomics workshop

Part One

Andrea Harper, Zhesi He

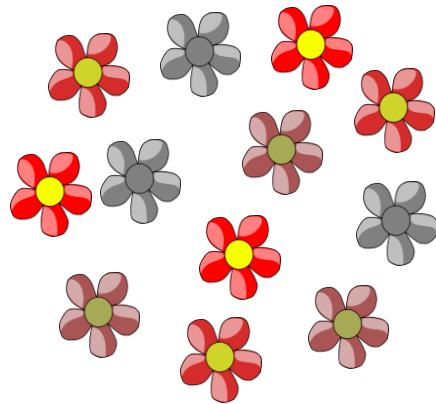
Introduction to AT

What is Associative Transcriptomics?

- Makes use of diversity panels instead of mapping populations to associate markers with traits
- Uses transcriptome data to associate traits with gene expression as well as gene sequence variation
- Theory based on Genome-Wide Association Studies (GWAS)...

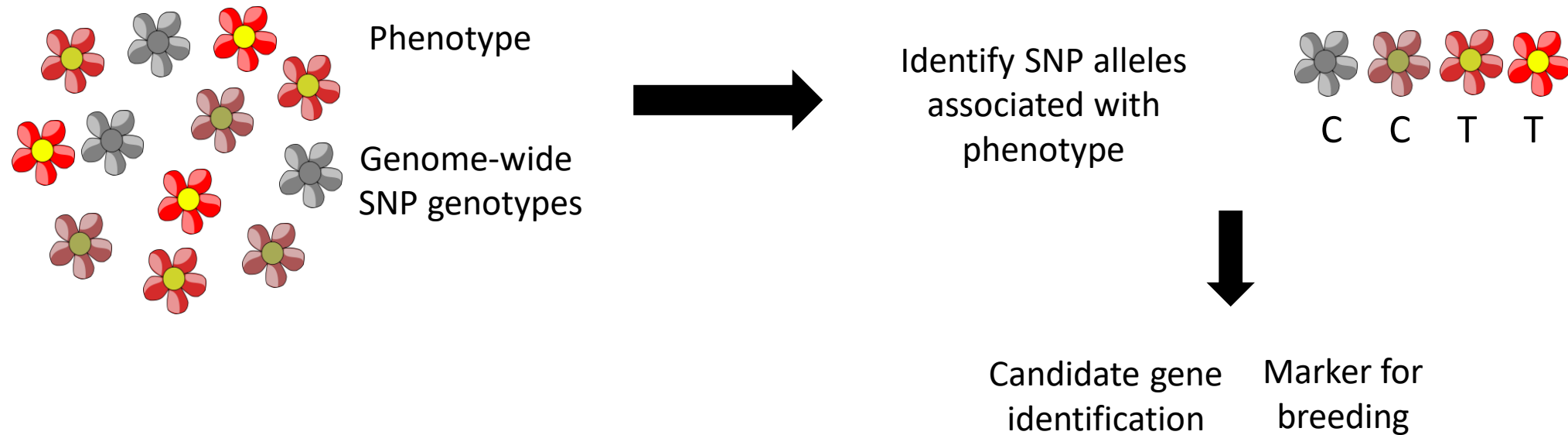
GWAS

- Identifies sequence markers (usually SNPs) that are significantly associated with a trait of interest across a diversity panel of individuals



GWAS

- Identifies sequence markers (usually SNPs) that are significantly associated with a trait of interest across a diversity panel of individuals



AT uses RNA-Seq data

- Associative Transcriptomics (AT) uses RNA-Seq data which enables us to look at both **gene sequence** and **gene expression** variation
- Short read Illumina RNA-Seq data from each sample is first aligned to a reference
- PORI Reference genome based on A and B global pan-genomes
 - 121,188 genes drawn from 12 genomes
- Once aligned, normalised expression is quantified from number of reads aligned to each gene
- SNP variants are also detected by comparing to the reference sequence
- SNP and gene expression matrices are then outputted in the correct format for AT analysis

SNP matrix

[illegible]

GEM matrix


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	V1	a-0001001	a-0001002	a-0001003	a-0001004	a-0001005	a-0001006	a-0001007	a-0001008	a-0001009	a-0001010	a-0001011	a-0001012	a-0001013	a-0001014	a-0001015	a-0001016
2	A01p000010.1_BraZAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	A01p000020.1_BraZAA	0.02864	0	0.143896	0.595278	0.113068	0.15603	0.029304	0.113446	0.085871	0.057336	0.578921	0.107854	0.260575	0	0.056979	0.08
4	A01p000030.1_BraZAA	0	0	0	0	0	0	0	0	0.12022	0	0	0	0	0	0	0
5	A01p000040.1_BnaBNP	0	0	0	0.059036	0	0	0	0.295336	0	0	0.060284	0	0	0.22278	0	0
6	A01p000100.1_BnaNYA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	A01p000140.1_BnaZSA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	A01p000150.1_BraZAA	0	0	0	0.147285	0	0	0	0	0	0	0	0	0	0	0	0
9	A01p000190.1_BraZAA	0.053167	0	0	0	0	0	0	0	0	0.053218	0.053734	0	0	0	0.052887	0
10	A01p000210.1_BraZAA	0	0	0.254468	0	0	0	0	0.752329	0.253094	0.253486	0	0.238414	0	0.472919	0.251908	1.23
11	A01p000260.1_BraZAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	A01p000270.1_BraZAA	0.261921	0.030038	0.058487	0.172823	0.114892	0	0.267993	0.086457	0.058171	0.407826	0.500019	0.493172	0.185345	0.135869	0	0.17
13	A01p000300.1_BraZAA	0.060145	0	0	0	0	0	0	0	0	0	0	0	0	0.224637	0	0
14	A01p000320.1_BraZAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	A01p000340.1_BraZAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	A01p000350.1_BraZAA	0	0	0	0	0	0	0	0	0	0	0.033809	0	0	0	0	0
17	A01p000360.1_BraZAA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	A01p000370.1_BraZAA	0.57739	0	0	0.190489	0	0	0	0	0	0	0	0	0.087553	0	0	0
19	A01p000410.1_BraZAA	0	0	0	0	0	0	0.108599	0	0.424305	0	0	0	0	0	0	0
20	A01p000430.1_BraZAA	0.264858	0.136687	0	0	0	0	0	0	0	0	0	0	0.120486	0	0	0

PORlindia_RPKM

Trait Data – template for PORI submission

York_Accession	PORI_sample_code	PORI_entry_name	Trait_1	Trait_2	etc
a-0001001	NB-2	CHUTKI			
a-0001002	NB-3	Sej-2(1)			
a-0001003	NB-4	BEC-161			
a-0001004	NB-5	D-247			
a-0001005	NB-6	BNF-5			
a-0001006	NB-7	LEH-1			
a-0001007	NB-8	D-205			
a-0001008	NB-9	GM-1			
	NB-10	RLM-619			
a-0001009	NB-11	BJ-1			
a-0001010	NB-12	SEJ-2(2)			
a-0001011	NB-13	RH-30			
a-0001012	NB-14	Pusa Jai Kisan			
a-0001013	NB-15	Laha T-59			
	NB-16	Pusa barani			
a-0001014	NB-17	Pusa bold			
a-0001015	NB-18	Varuna			
	NB-21	RL-1359			
a-0001016	NB-23	B. juncea from Kanpur			
a-0001017	NB-24	Proagro-4			
a-0001018	NB-25	IM-2004			
a-0001019	NB-27	Krishna			
a-0001020	NB-28	RLM-514			
a-0001021	NB-29	Kranti			
a-0001022	NB-30	Rajat			
a-0001023	NB-31	Rohini			
a-0001024	NB-32	RH-819			

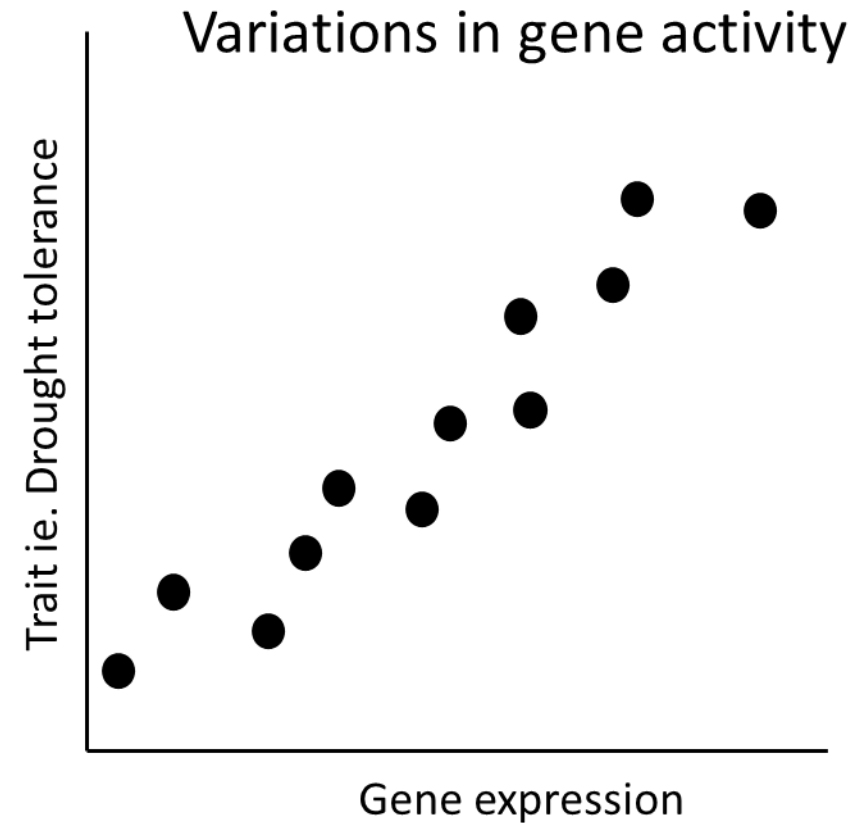
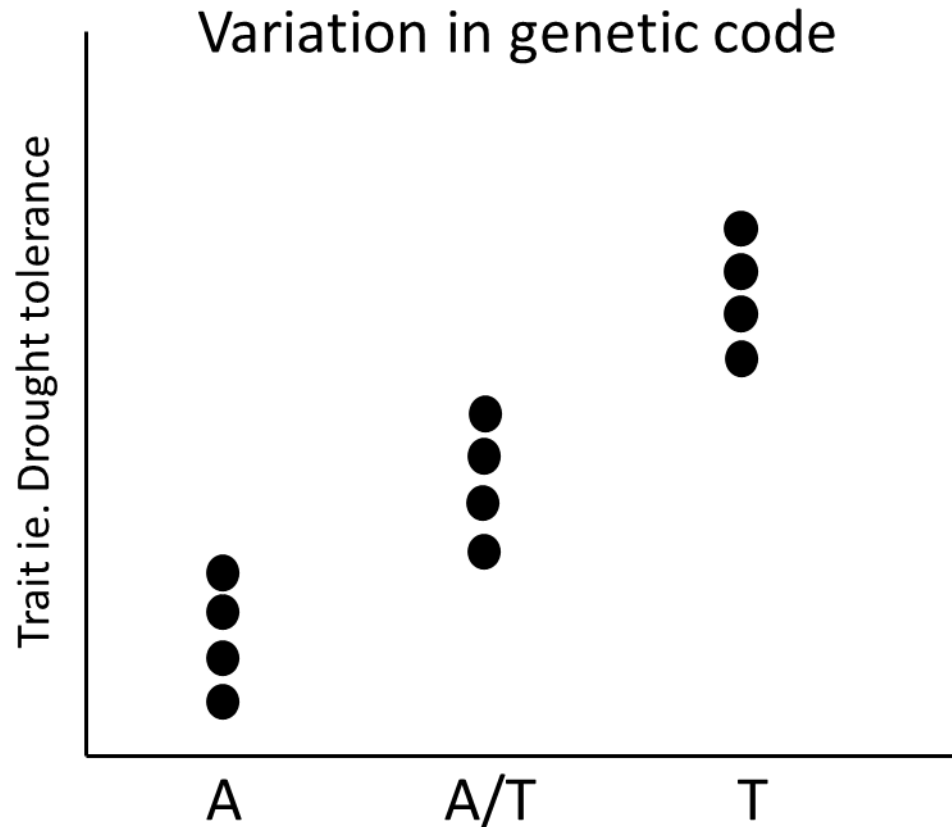
Trait format for input to AT

 *Untitled - Notepad

File Edit Format View Help

Taxa	TSW	Pod_Length	Seeds_Pod	Pct_Oil
a-0001001	4.94	4.7	15	41
a-0001002	3.14	5	10	38
a-0001003	2.62	3.6	16	43
a-0001004	3.66	4	14	40
a-0001005	4.31	4.1	14	39
a-0001006	3.28	3.4	16	40
a-0001007	2.7	4.5	15	39
a-0001008	5.79	4.1	12	39
a-0001010	3.79	4.4	13	39
a-0001011	4.99	4.5	11	40
a-0001012	7.09	7.2	19	38
a-0001013	5.69	3.3	12	38
a-0001014	5.42	5.8	12	40
a-0001015	5.08	6	13	40
a-0001016	5.05	5.3	13	38
a-0001017	5.83	5	14	37.43
a-0001018	5.78	5	14	39
a-0001019	4.5	3.6	12	37
a-0001020	3.75	4.6	12	40
a-0001021	3.9	5.7	13	41
a-0001022	4.66	3.5	10	39
a-0001023	4.91	4.3	14	39
a-0001024	4.98	4.8	14	39
a-0001027	3.93	4.6	15	40
a-0001030	5.56	4	11	38
a-0001031	5.02	5.7	13	41
a-0001032	6.89	3.67	13	38
a-0001033	6.42	3.9	12	38
a-0001034	6.58	4	12	39

AT aims to identify genetic markers associated with the trait of interest (TOI)

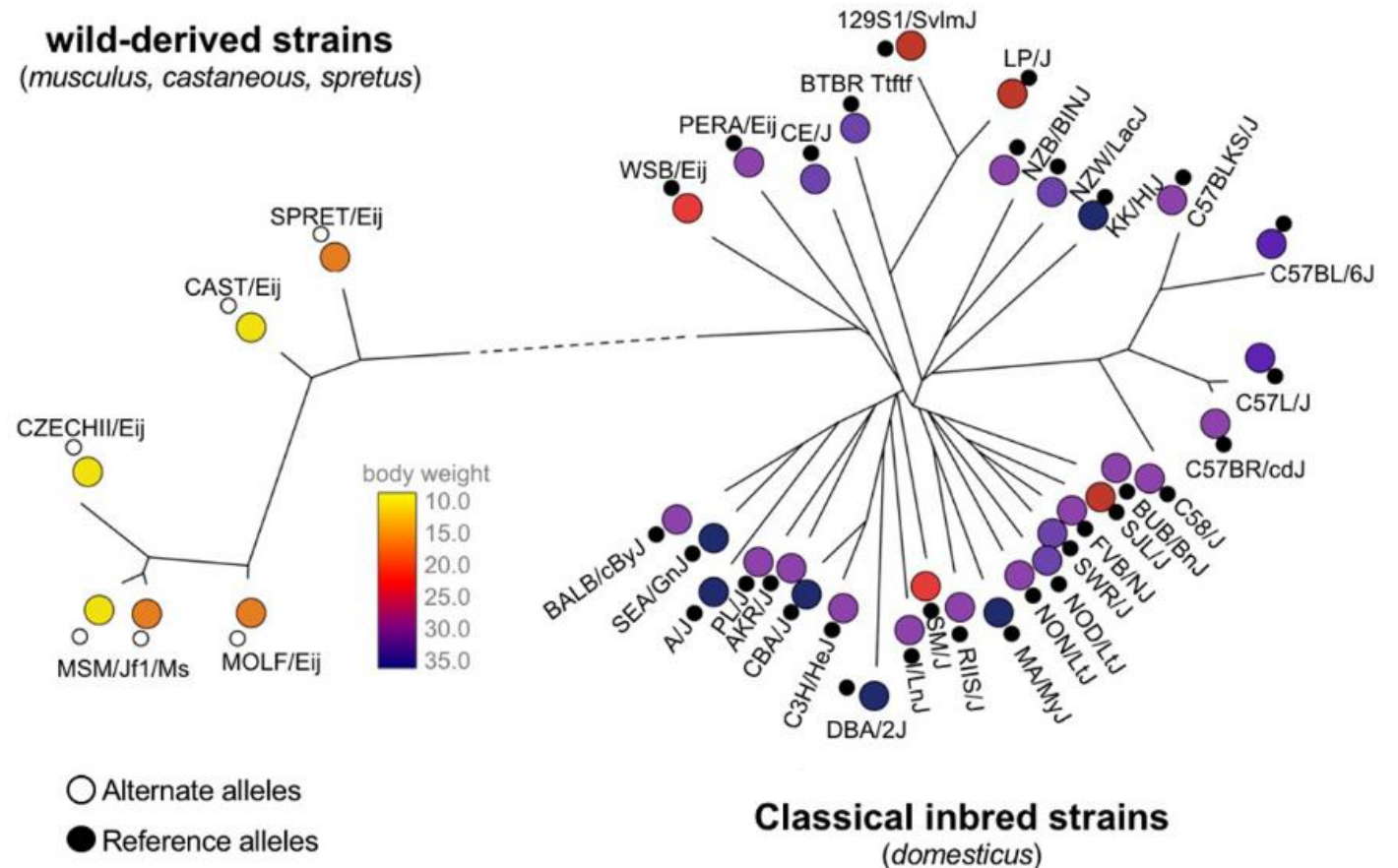


Population Structure

Population structure can lead to false positives

Population structure introduces uncertainty about associations

Is the marker really associated with the trait, or are some alleles just more common in one population than the other?



AT linear models account for population structure

- As simple linear models (such as simple linear regressions between trait and genetic variant) are prone to false positives, we use more complex types of model
- Both SNP and GEM models incorporate population structure, and SNP model also incorporates pairwise relatedness in a mixed linear model

There are various ways to calculate population structure

- Populations with different ancestry should have different allele frequencies
- Two general ways of identifying structure
 - Model-based clustering (ie. STRUCTURE)
 - Principal components analysis (PCA)
- Model-based clustering is slower, but can be easier to interpret
- PCA is much faster, but does not consider prior information about your data

Population Structure

STRUCTURE

- Model-based Bayesian clustering algorithm
- Assumes a model with K populations, and attempts to assign individuals to a population
- Run different numbers of population clusters, for several iterations to calculate the most likely number
- The resulting Q matrix assigns ancestry of every individual to these population clusters (ie. 40% of Individual 1's background is from cluster 1, 60% cluster 2 etc.)
- This is a good method for population structure analysis, and the results are easy to interpret, but it is very slow!
- Also, other, simpler methods perform equally well (maybe better!) at controlling for population structure in association analyses...

Population Structure

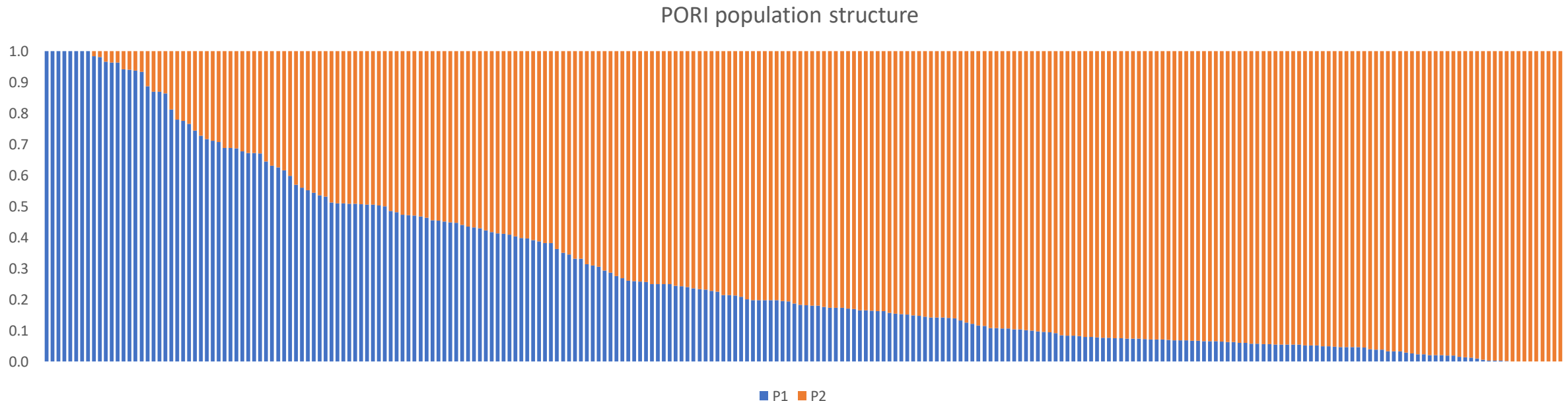
PCA

- Method of reducing data complexity when many variables are measured
- Creates new variables (principal components; PCs) that are linear combinations of the old ones, and describe the variation in the data
- When PCA is used for population structure analysis, it often separates individuals by geography or crop type
- In association analysis, we can use the top few PCs as a covariate instead of a Q matrix
- However, choosing how many PCs to use is difficult
- Also, PCs are not as easy to interpret as a Q matrix

PORI Population Structure

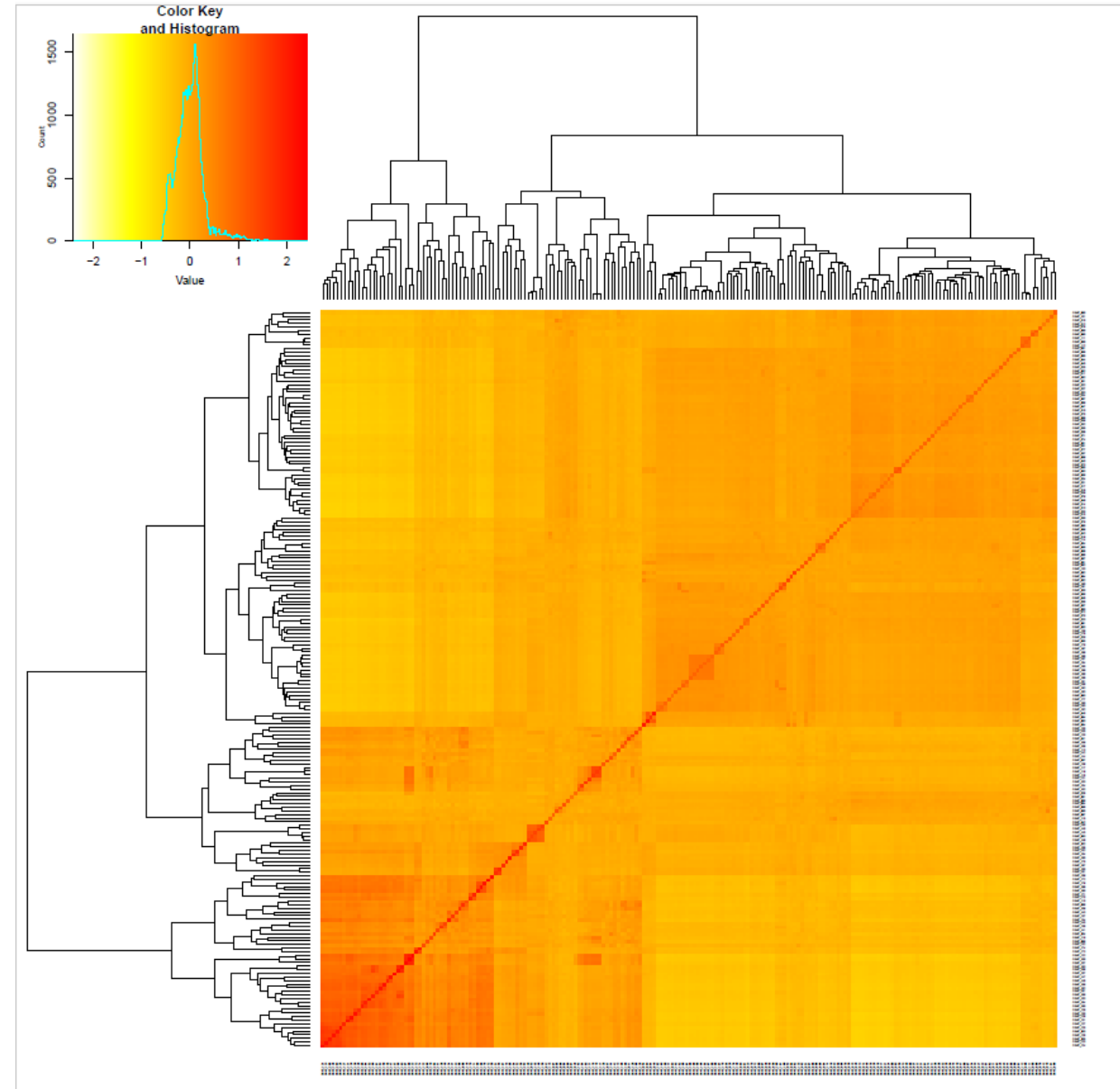
PSIKO

- For AT, we use a PCA method which also estimates the number of population clusters for us, and creates a type of Q matrix
- It is fast *and* easy to interpret!
- PSIKO – DOI:10.1534/genetics.114.171314



Relatedness

- Population structure analysis measures major clustering of individuals
- Relatedness between pairs of individuals also adds noise to the data
- We can simply calculate a measure of similarity between each pair of individuals based on all of their genotypes
- This “kinship” matrix can also be used to minimise false positives in the analysis



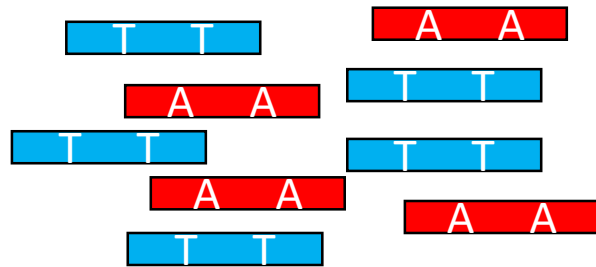
Linkage Disequilibrium

Association approaches use Linkage Disequilibrium (LD) to identify QTL regions

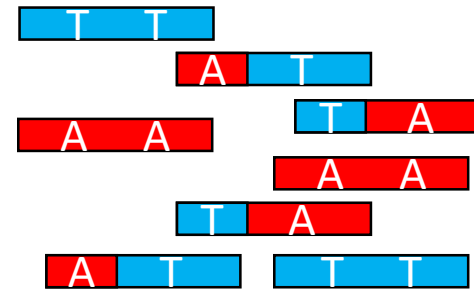
- LD Determines the resolution of association mapping in a population:
 - Long distance LD - Mapping at the centimorgan (cM) distances
 - Short distance LD - Mapping at the base pair (gene) distance
- LD measures the degree of “Non-random association of alleles at two or more loci”

Linkage disequilibrium (LD)

- “Non-random association of alleles at two or more loci”



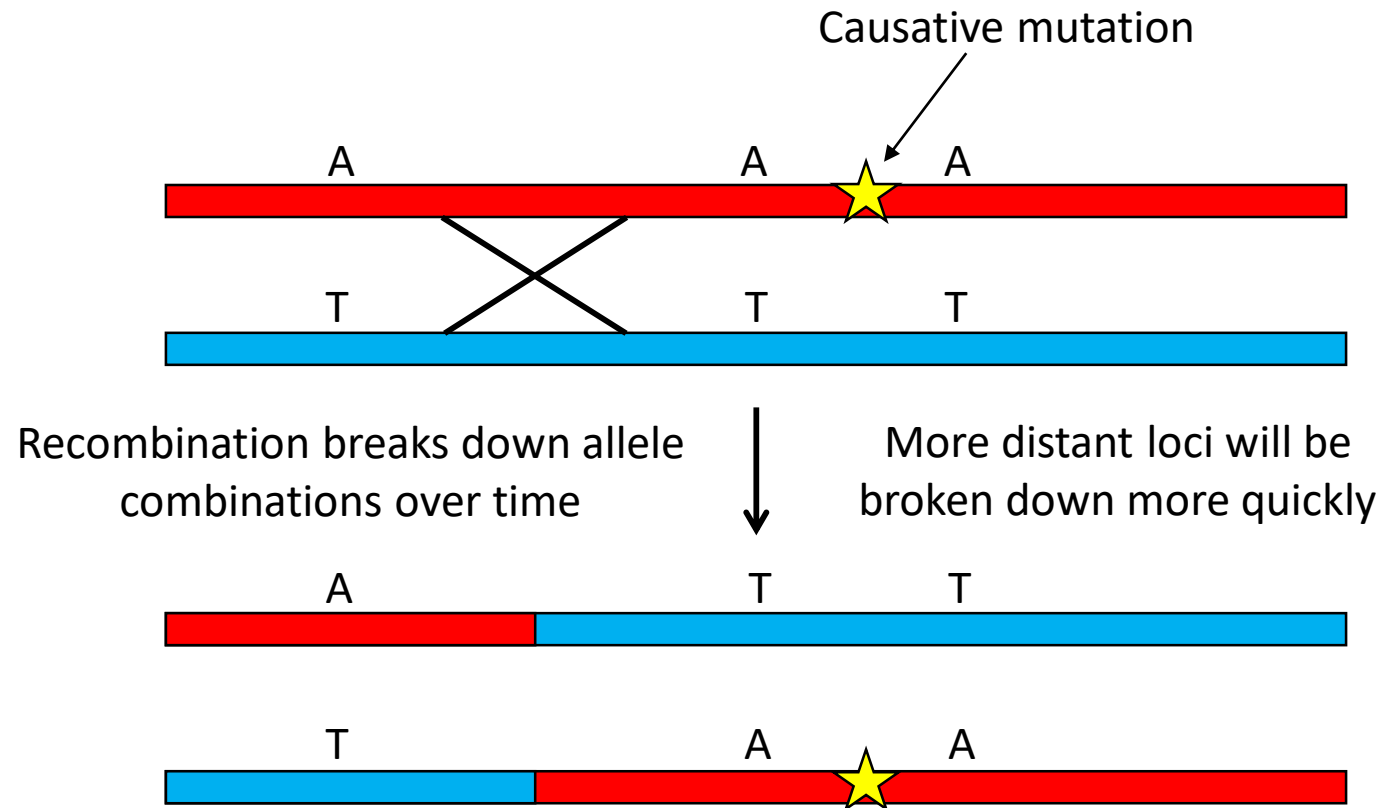
If alleles are ALWAYS seen together in a population then they are in perfect LINKAGE DISEQUILIBRIUM



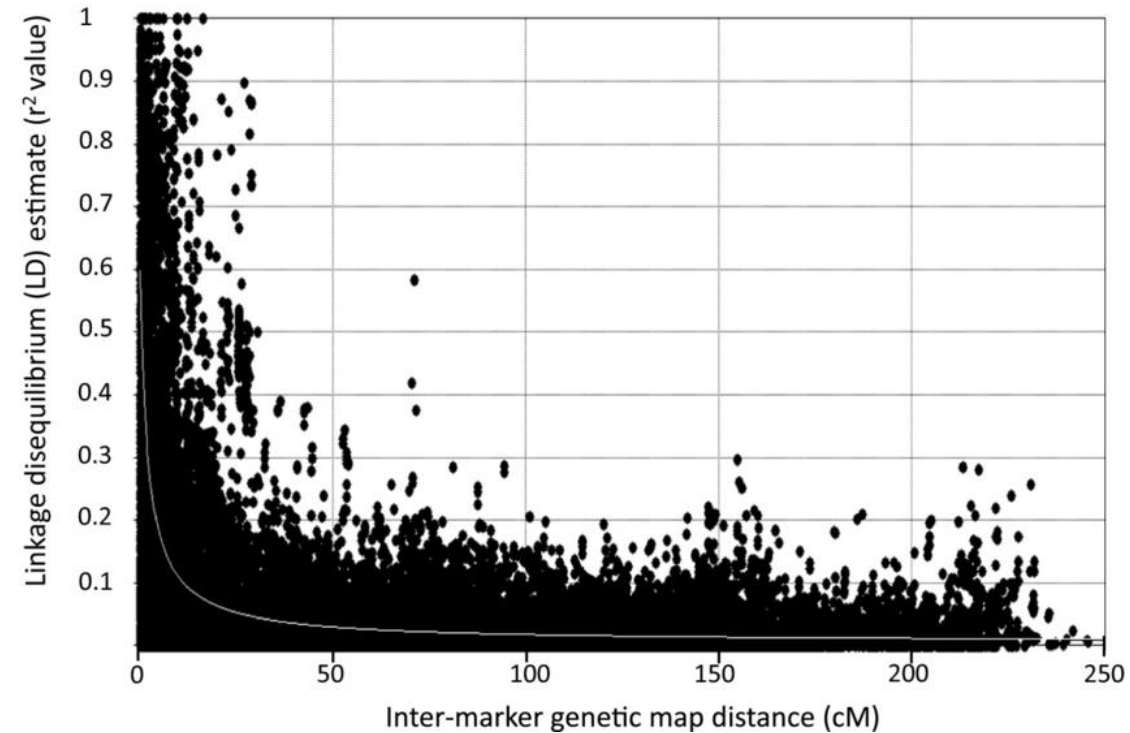
If all combinations of alleles are seen at random in a population then they are in LINKAGE EQUILIBRIUM

Linkage disequilibrium (LD) is broken down by recombination

- “Non-random association of alleles at two or more loci”



LD decays with distance between markers



Major factors affecting LD

- Recombination rate

- Population size

- Inbreeding

- Mutation rate



Mostly out of your control

- Population structure

- Selection



We should be correcting for this

Major factors affecting LD

- Recombination rate
- Population size
- Inbreeding
- Mutation rate

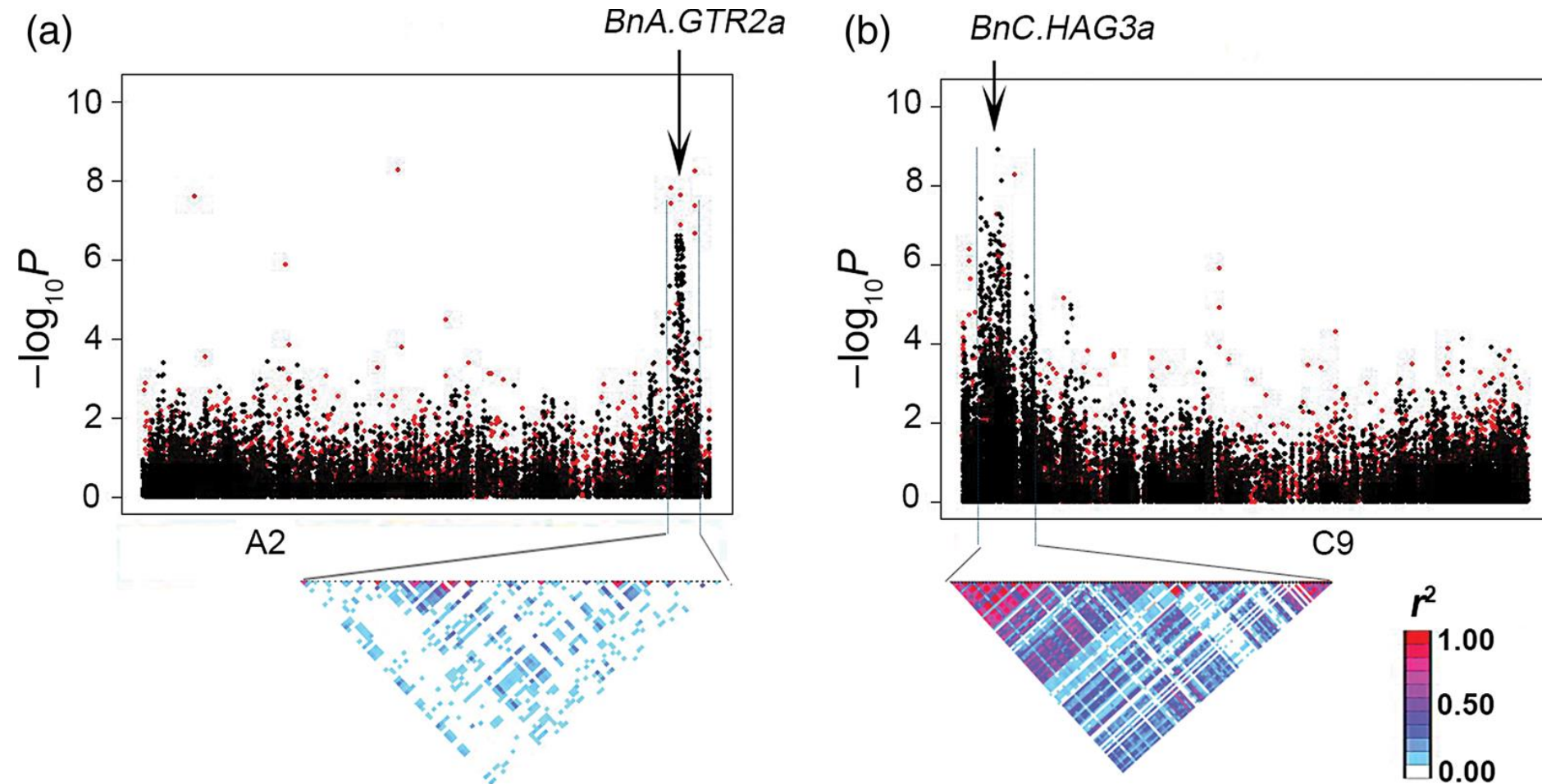
Mostly out of your control

- Population structure
- **Selection**

We should be correcting for this

Could be quite common in crops where desirable traits have been selected

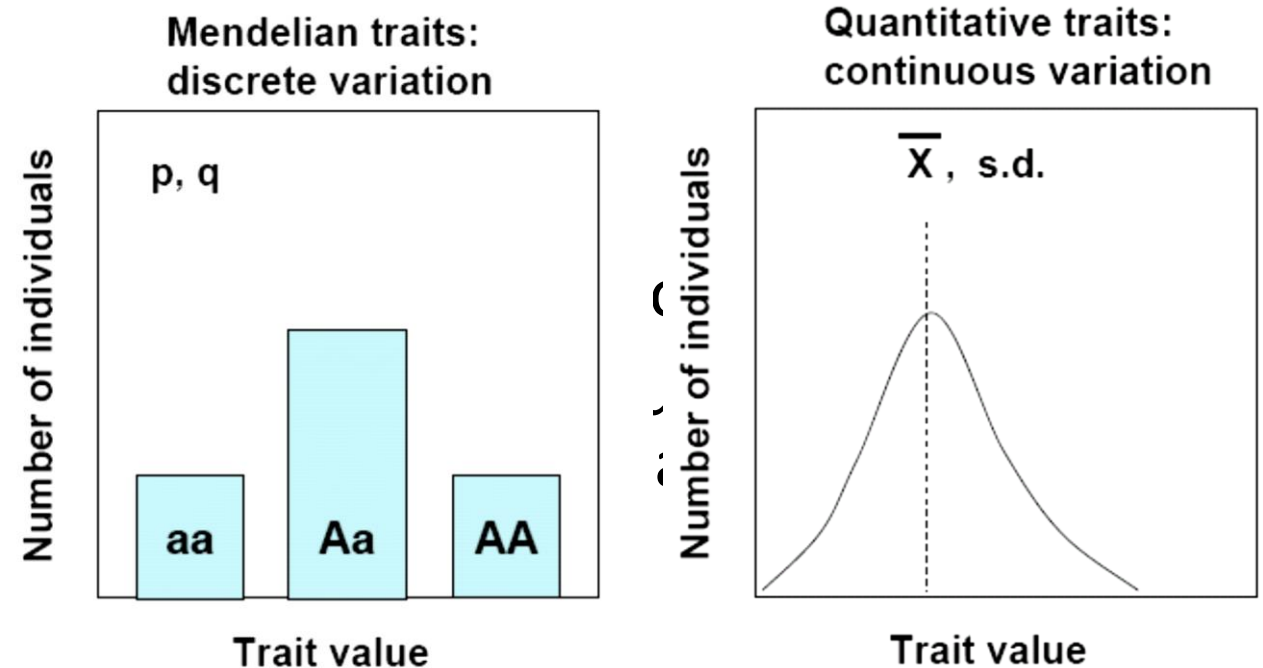
LD affects resolution of association peaks



Optimising Trait Data

Important considerations for your trait data

- Association approaches are ideally suited to complex (polygenic) traits
- However, environment may also contribute to complex phenotypes
- High quality phenotype data (precise, controlled, replicated measurements etc.) will improve analysis

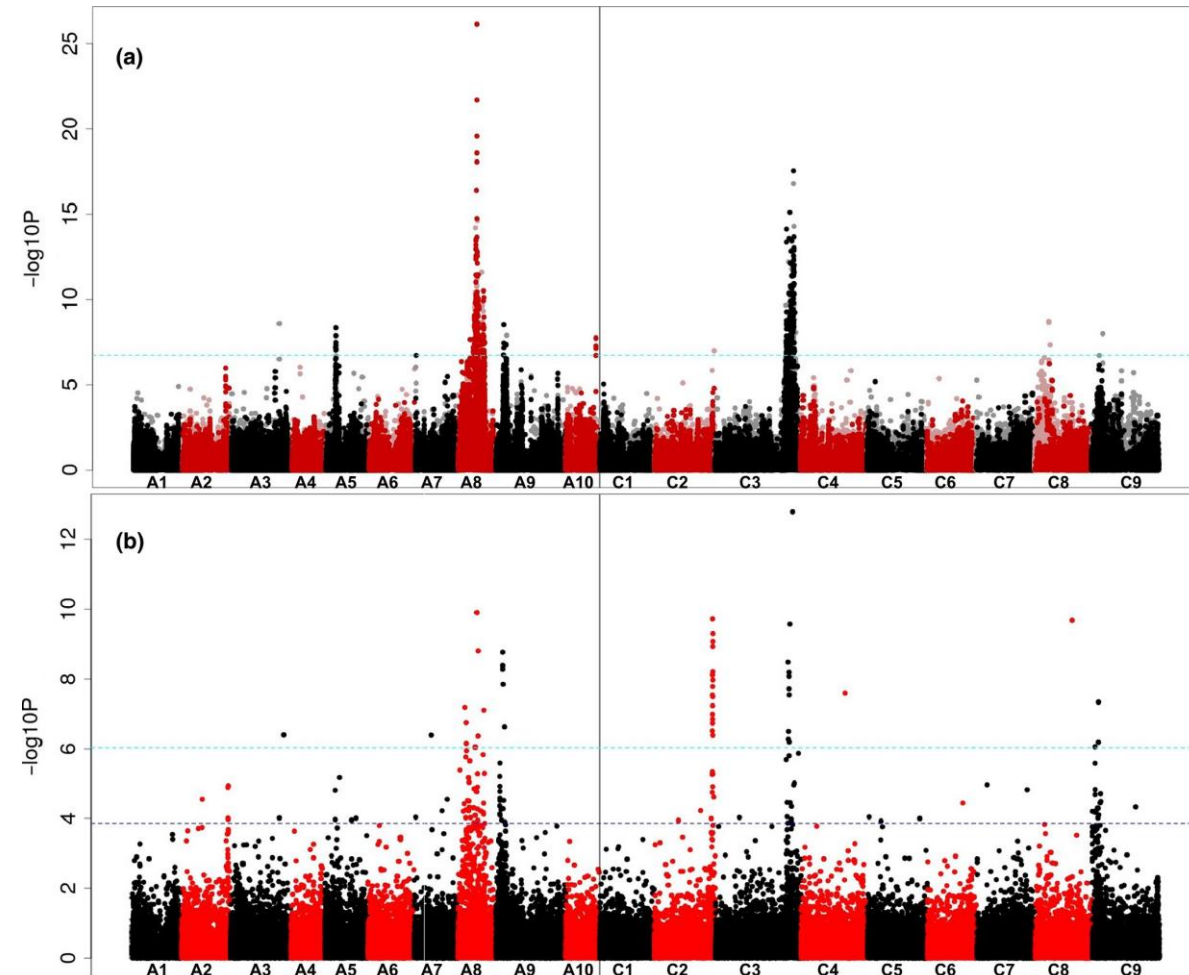


Variation follows a normal distribution if:

- Multiple loci are involved (quantitative)
- Each locus has about equal effect (additive)
- Each locus acts independently (interaction is minimal)

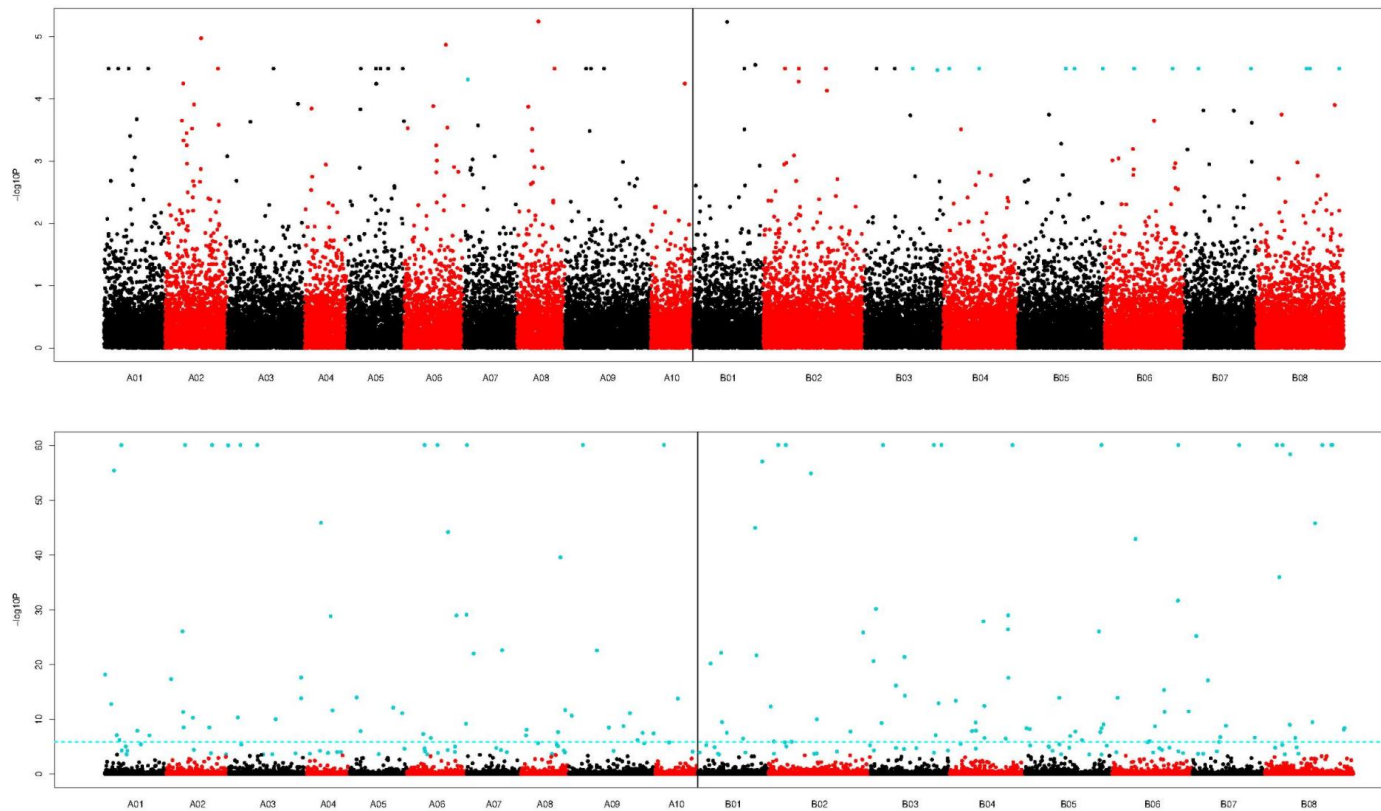
Important considerations for your trait data

- Association approaches may still work if your data is not ideal
 - Controlled by just a few major genes
 - Categorical data
 - Non-normally distributed
- However, results from these data should be interpreted with care (more on that tomorrow!)

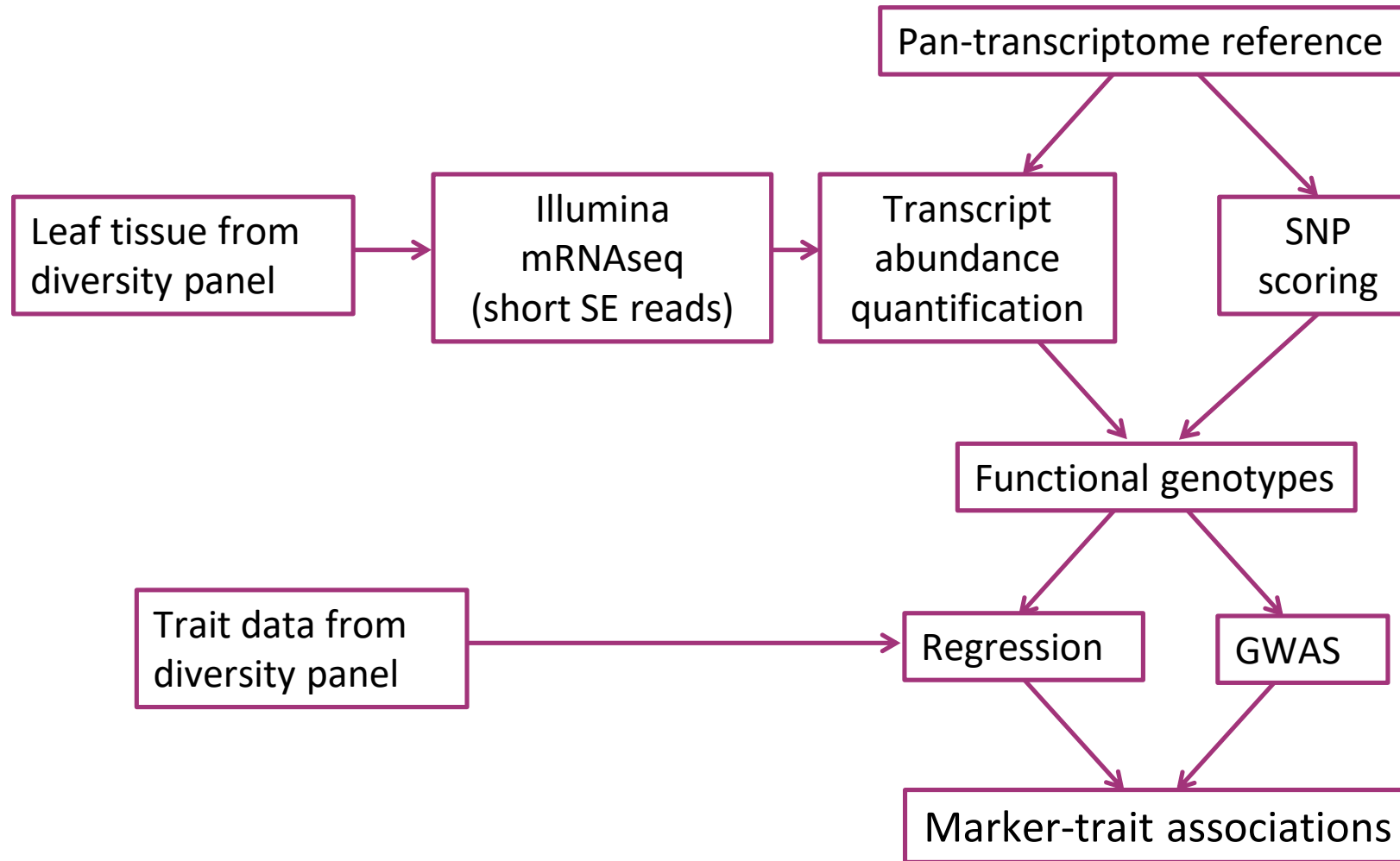


Experimental design considerations

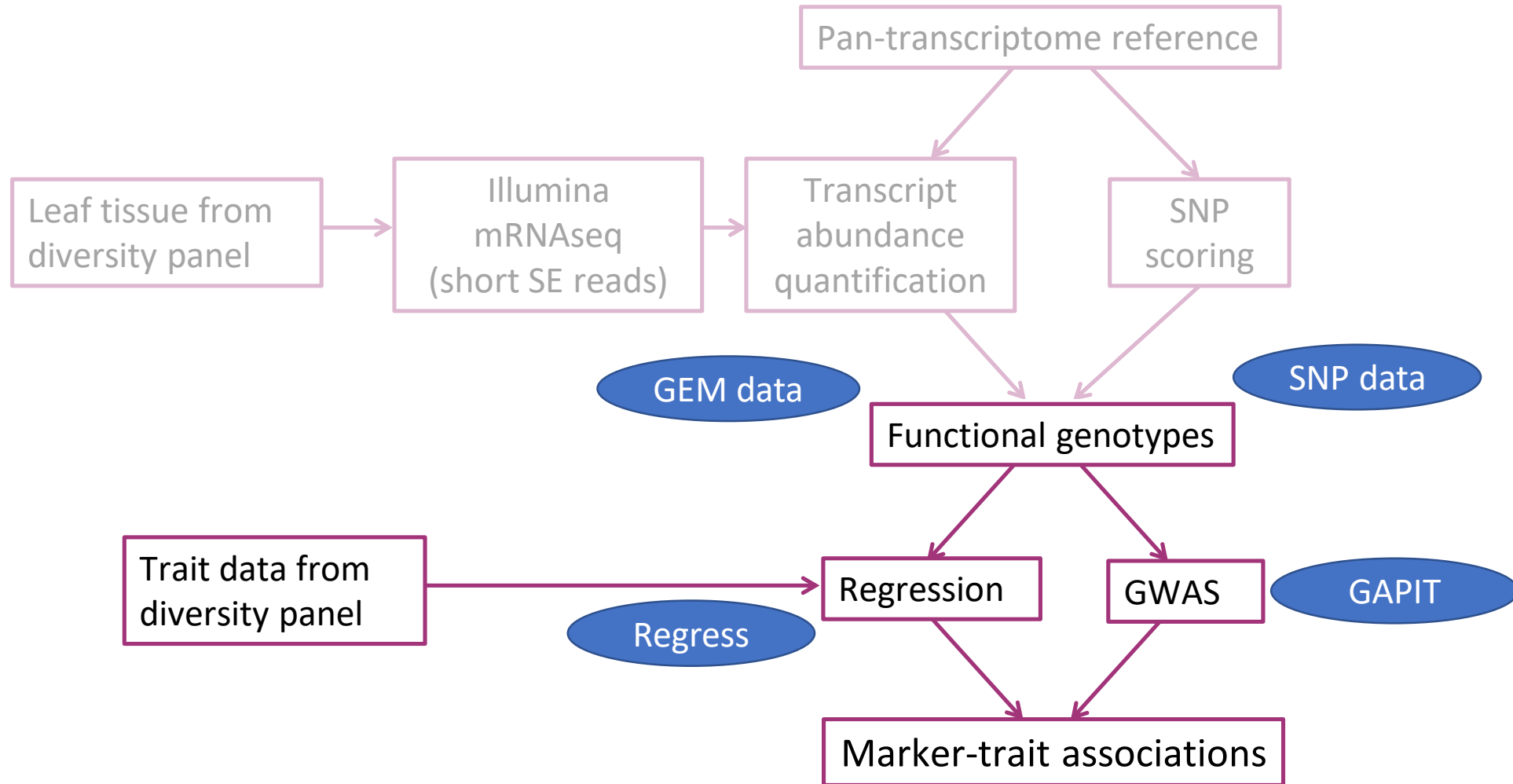
- Outliers in the data can cause big problems (false positives)



AT pipeline



AT pipeline



Inside the black box

- SNP analysis – GAPIT
<https://zzlab.net/GAPIT/>
- GEM analysis – Regress
Bancroft/Harper Lab scripts

GAPIT

- GAPIT is a freely available R package, designed for GWAS analysis
- It does the following:
 - Inputs SNP data table and converts it to a numerical SNP matrix
 - Calculates a kinship (relatedness) matrix
 - Inputs trait and population structure data tables
 - Runs mixed linear model for every SNP, with genotype and population structure as fixed effects, and relatedness as a random effect
 - Generates results files and plots*

*we replot all Manhattan plots so that they can be directly compared with the Regress results

Regress

- Custom script written by Bancroft/Harper Labs
- It does the following:
 - Inputs expression (RPKM) data table
 - Filters out genes with negligible expression levels across the panel
 - Inputs trait and population data
 - Runs linear regression model for every GEM, with genotype and population structure as fixed effects
 - Generates results files and plots

Running AT on the server