



Centre for
Tropical Livestock
Genetics and Health

Introduction to GWAS

**Ozzie Matika, Chrissy Rochus
& Isidore Houaga**

Roslin Institute, University of Edinburgh

Slides adapted from Ivan Pocrníć





Centre for
Tropical Livestock
Genetics and Health

Basics already covered

- **VERY BASIC Genetics ...!**
- **Genetic models for complex traits**
 - Heritabilities and variance partitioning
 - **SNP associations**
 - **Genomic selection**
- **What geneticists do ...**

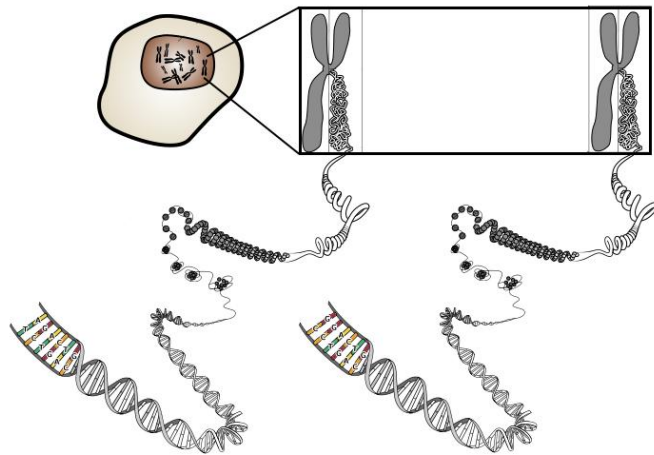


Refresher on genomic data

DNA, $\sim 3 \times 10^9$ base pairs x 2

SNP array genotyping $\sim <100$ to 10^{5-6} variants

Sequencing $\sim 10^{7-8}$ variants



Haplotypes

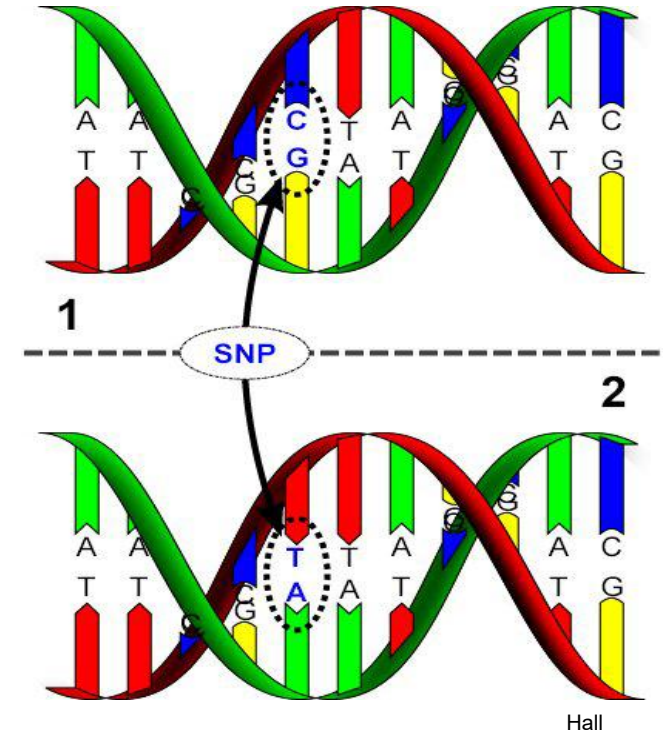
Id1	0	0	1	1	1	1	0	0	...
Id1	1	0	1	1	1	0	0	0	...
Id2	0	0	0	1	1	1	1	1	...
Id2	0	0	0	1	1	0	0	0	...

...

Genotypes

Id1	1	0	2	2	2	1	0	0	...
Id2	0	0	0	2	2	1	1	1	...

...





Refresher on genomic data

The sequences of >150,119 genomes in the UK biobank

<https://doi.org/10.1038/s41586-022-04965-x>

~600M SNPs (representing 7% of all possible human SNPs)

~60M indels

~1M structural variants

~3M microsatellites

Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Genomic Prediction from Whole Genome Sequence in Livestock: the 1000 Bull Genomes Project

B.J. Hayes^{1,2,3}, I.M. MacLeod^{3,4}, H.D. Daetwyler^{1,2,3}, P.J. Bowman^{2,3}, A.J. Chamberlain^{2,3}, C.J. Vander Jagt^{2,3}, A. Capitan^{5,6}, H. Pausch⁶, P. Stothard⁷, X. Liao⁷, C. Schrooten⁸, E. Mullaart⁸, R. Fries⁶, B. Guldbrandtsen⁹, M.S. Lund⁹, D.A. Boichard⁵, R.F. Veerkamp¹⁰, C.P. VanTassell¹¹, B. Gredler¹², T. Druet¹³, A. Bagnato¹⁴, J. Vilkki¹⁵, D.J. deKoning¹⁶, E. Santus¹⁷, and M.E. Goddard^{2,3,4}.



Simple Genetics: Mutations

- **Base change every 500-1000 bases**
 - A A G T A C A T G G C
 - A A **A** T A C A T G G C
 - A A G T A **T** A T G G C
 - A A G T A C A T **C** G C
 - A A G T A C A T G G **A**
- **Mutations within genes:**
 - **No effect**
 - **Altered amino acid (hence protein)**
- **Mutations outside genes**
 - **No effect**
 - **Altered expression of nearby gene**



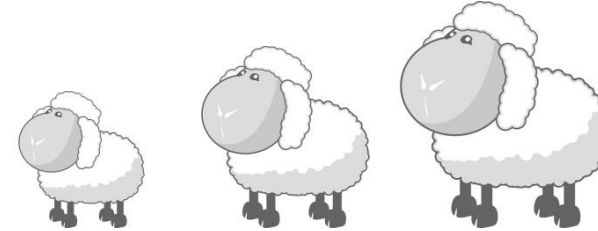
SNPs and SNP Chips

- **SNP = Single Nucleotide Polymorphism**
 - **Specific base where animals differ (G,C,A,T)**
- **Typical SNP chips ~50,000 SNPs**
 - **50k simultaneous genotypes**
- **Typical genome ~ 3,000,000,000 bases**
 - = 1,000,000 pages of info
 - = 2,000 airport blockbusters
 - **Info every 60,000 bases**
- **50k SNP chips captures ~ 1% of all SNPs**



“Mutation-Dependent” Genetics

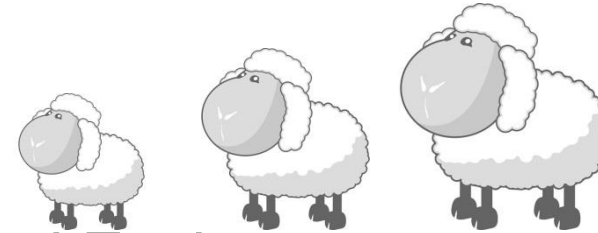
- **Consider a population:**
 - Each individual is different
 - Differences due to **Genetics** and **Environment**
- **Consider genetics:**
 - All individuals have the **same** genes
 - But they have different **variants** of each gene





“Mutation-Dependent” Genetics

- Consider a population:
 - Each individual is different
 - Differences due to Genetics and Environment
- Consider genetics:
 - All individuals have the same genes
 - But they have different variants of each gene
- **Consider different traits:**
 - **For some: between-animal variation due to variants at a few genes**
 - **For others: variants at many genes are important**





Centre for
Tropical Livestock
Genetics and Health

Motivation for genome-wide association studies

Concept is “simple”:

- Find genes related to the traits of interest
- Get the insight into genetic architecture of the trait

Several strategies exists; linkage analysis, association mapping, ...

With availability of genome-wide dense SNP chips

- Test associations between SNP and phenotype → GWAS
- Where significant associations → SNP in LD with QTL



Genetic Models

- Consider a single locus (or SNP)

Genotypes: tt tT TT

$\text{freq.}(t) = 1-p = q$; $\text{freq.}(T) = p$

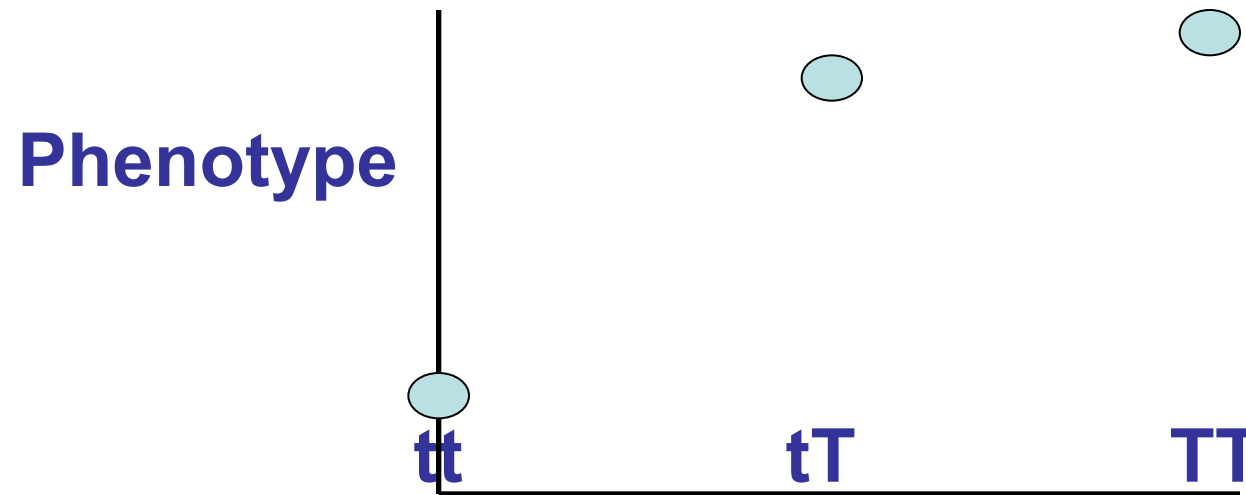


Genetic Models

- Consider a single locus (or SNP)

Genotypes: tt tT TT

freq.(t) = $1-p = q$; freq.(T) = p



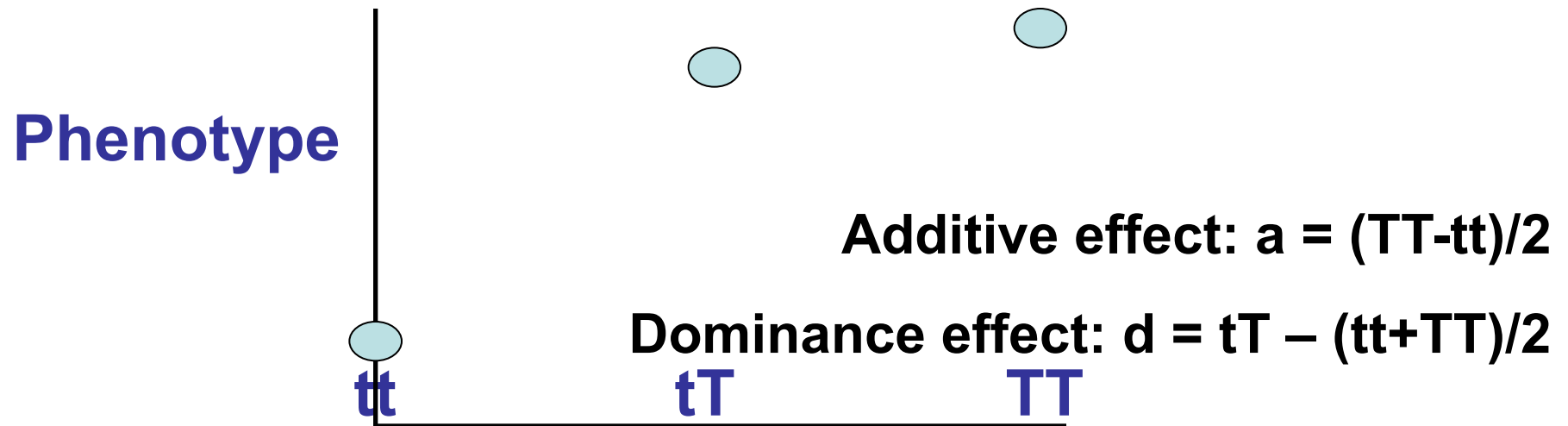


Genetic Models

- Consider a single locus (or SNP)

Genotypes: tt tT TT

freq.(t) = $1-p = q$; freq.(T) = p



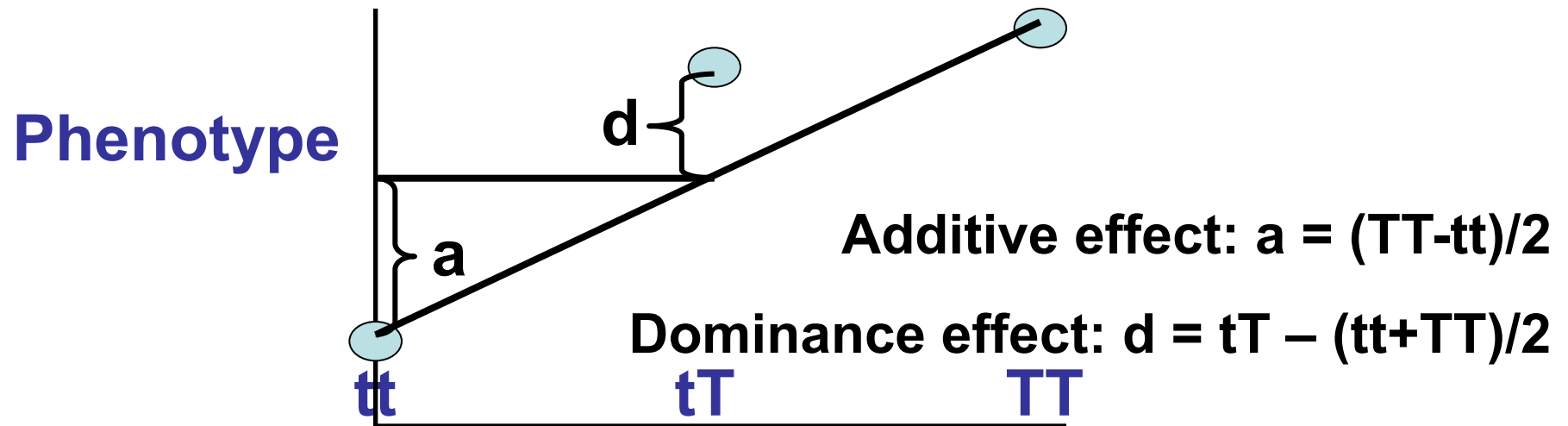


Genetic Models

- Consider a single locus (or SNP)

Genotypes: tt tT TT

freq.(t) = $1-p = q$; freq.(T) = p



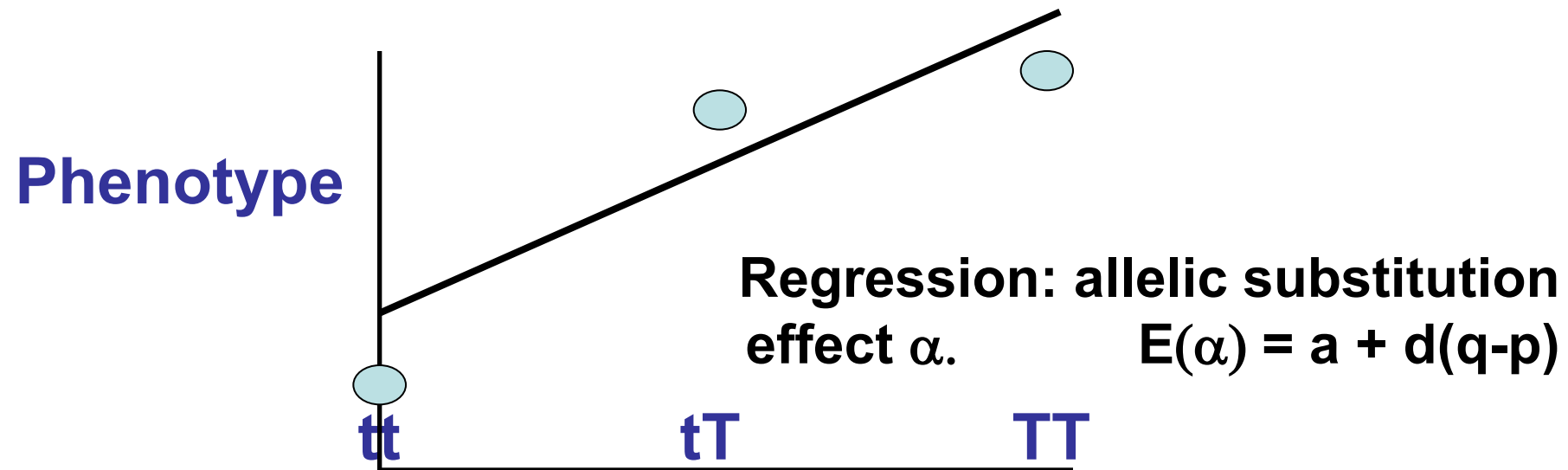


Genetic Models

- Consider a single locus (or SNP)

Genotypes: tt tT TT

freq.(t) = $1-p = q$; freq.(T) = p





Genetic Models

- Consider many loci

$$P = \mu + G + E$$



Genetic Models

- Consider many loci

Phenotype

Environment

$$P = \mu + G + E$$

$$\text{Genotype} = \sum g_j \quad (j=1, m \text{ loci}) \quad (i=1, n \text{ animals})$$



Genetic Models

- Consider many loci

Phenotype

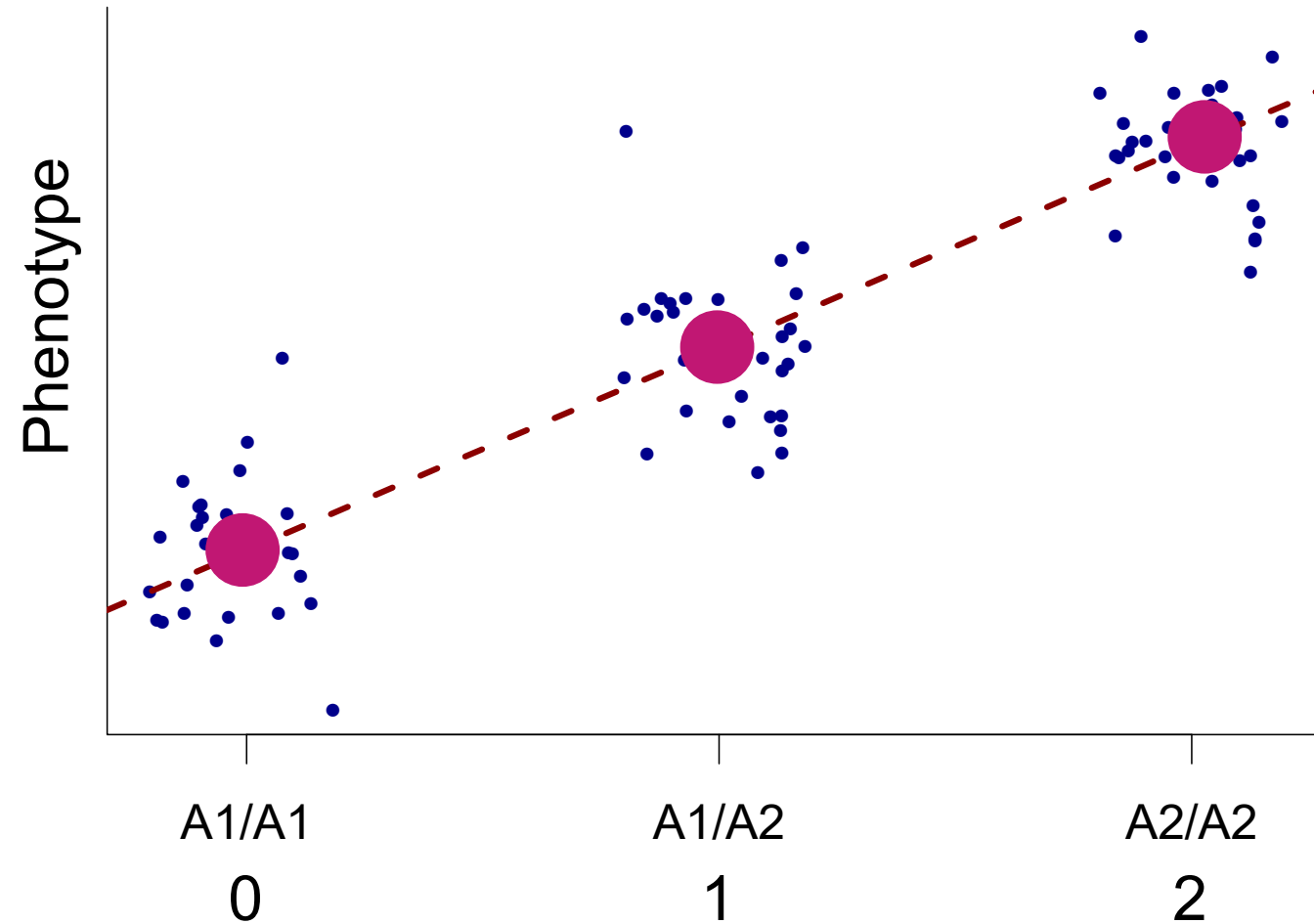
Environment

$$P = \mu + G + \text{SNP}_j + E$$

$$\text{Genotype} = \sum g_j \quad (j=1, m \text{ loci}) \quad (i=1, n \text{ animals})$$



Basic Principal





Linear Mixed Models

- $Y = Xb + Zu + Wc + \dots$
- b = vector of fixed effects, X = design matrix
 - INCLUDES SNP
- u = vector of random genetic effects, Z = design ...
- c = vector of another random effect
 - E.g. litter effect, maternal, QTL, ...
- Estimate b , u , c , etc and variance components
 - Estimated u = Estimated Breeding Values (EBV)



Linear Mixed Models

- $Y = Xb + Zu + Wc + \dots$
- b = vector of fixed effects, X = design matrix
 - INCLUDES SNP
- u = vector of random genetic effects, Z = design ...
- c = vector of another random effect
 - E.g. litter effect, maternal, QTL, ...
- Estimate b , u , c , etc and variance components
 - Estimated u = Estimated Breeding Values (EBV)
- 50,000 SNPs => specialised software needed



SNP Association Issues

- **Linkage Disequilibrium (LD)**
- **Population Structure**
- **Multiple Testing**



Definitions of LD

- **LD is required for both linkage and linkage disequilibrium mapping**
- **Difference:**
 - linkage analysis mapping considers the within family LD
 - extends for 10s of cM and broken down after only a few generations
 - LD mapping requires a marker allele to have within population LD with a QTL allele
 - association must have persisted across multiple generations as a property of the population
 - Which means marker and QTL must be very close



Definitions of LD

		Marker A		
		A1	A2	Freq
Marker B	B1	0.4	0.1	0.5
	B2	0.1	0.4	0.5
	Freq	0.5	0.5	

$$\begin{aligned} D &= \text{freq}(A1_B1) * \text{freq}(A2_B2) - \text{freq}(A1_B2) * \text{freq}(A2_B1) \\ &= 0.4 * 0.4 - 0.1 * 0.1 \\ &= 0.15 \end{aligned}$$



Definitions of LD

- Measuring LD (determines how dense markers need to be for LD mapping)

$$D = \text{freq}(A1_B1) * \text{freq}(A2_B2) - \text{freq}(A1_B2) * \text{freq}(A2_B1)$$

– highly dependent on allele frequencies

- not suitable for comparing LD at different sites

$$r^2 = D^2 / [\text{freq}(A1) * \text{freq}(A2) * \text{freq}(B1) * \text{freq}(B2)]$$



Definitions of LD

		Marker A		
		A1	A2	Freq
Marker B	B1	0.4	0.1	0.5
	B2	0.1	0.4	0.5
	Freq	0.5	0.5	

$$D = 0.15$$

$$r^2 = D^2 / [\text{freq}(A1) * \text{freq}(A2) * \text{freq}(B1) * \text{freq}(B2)]$$

$$r^2 = 0.15^2 / [0.5 * 0.5 * 0.5 * 0.5]$$
$$= 0.36$$



Definitions of LD

- If we assume two loci: one is marker and the other QTL
- r^2 between the two can give the proportion of QTL variance which can be observed at the marker
 - e.g. if variance due to a QTL is 300kg^2 , and r^2 between marker and QTL is 0.2, variation observed at the marker is 60kg^2 .
- The power of LD mapping to detect QTL
 - sample size needs be increased by $1/r^2$ to have the same power as an experiment observing directly the QTL



Definitions of LD

- Multi-locus measures of LD
 - r^2 is easy to calculate, useful and widely used
 - Takes two loci at a time
 - Does not tell us of the causes of LD

Causes of LD



Centre for
Tropical Livestock
Genetics and Health

- **Migration**

- LD artificially created in crosses

- large when crossing inbred lines
 - small in cross breeding of breeds that are similar in gene frequencies
 - Transitional- disappears after a number of generations

- **Selection**

- Selective sweeps

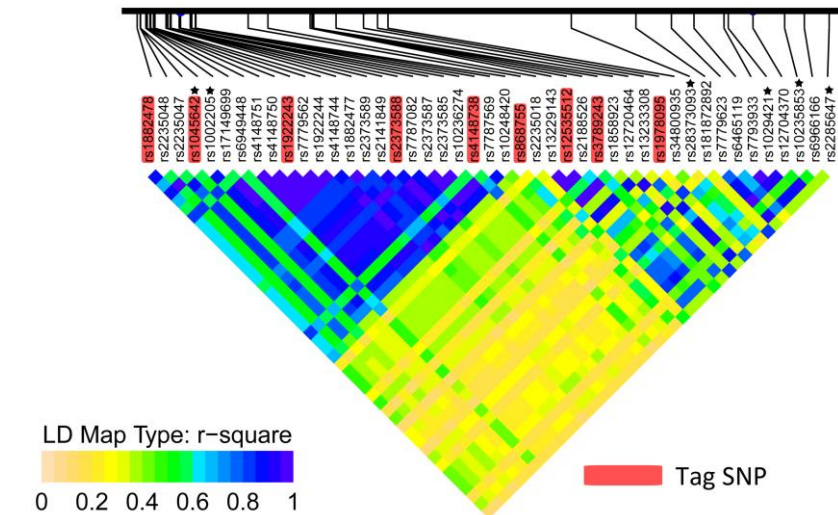
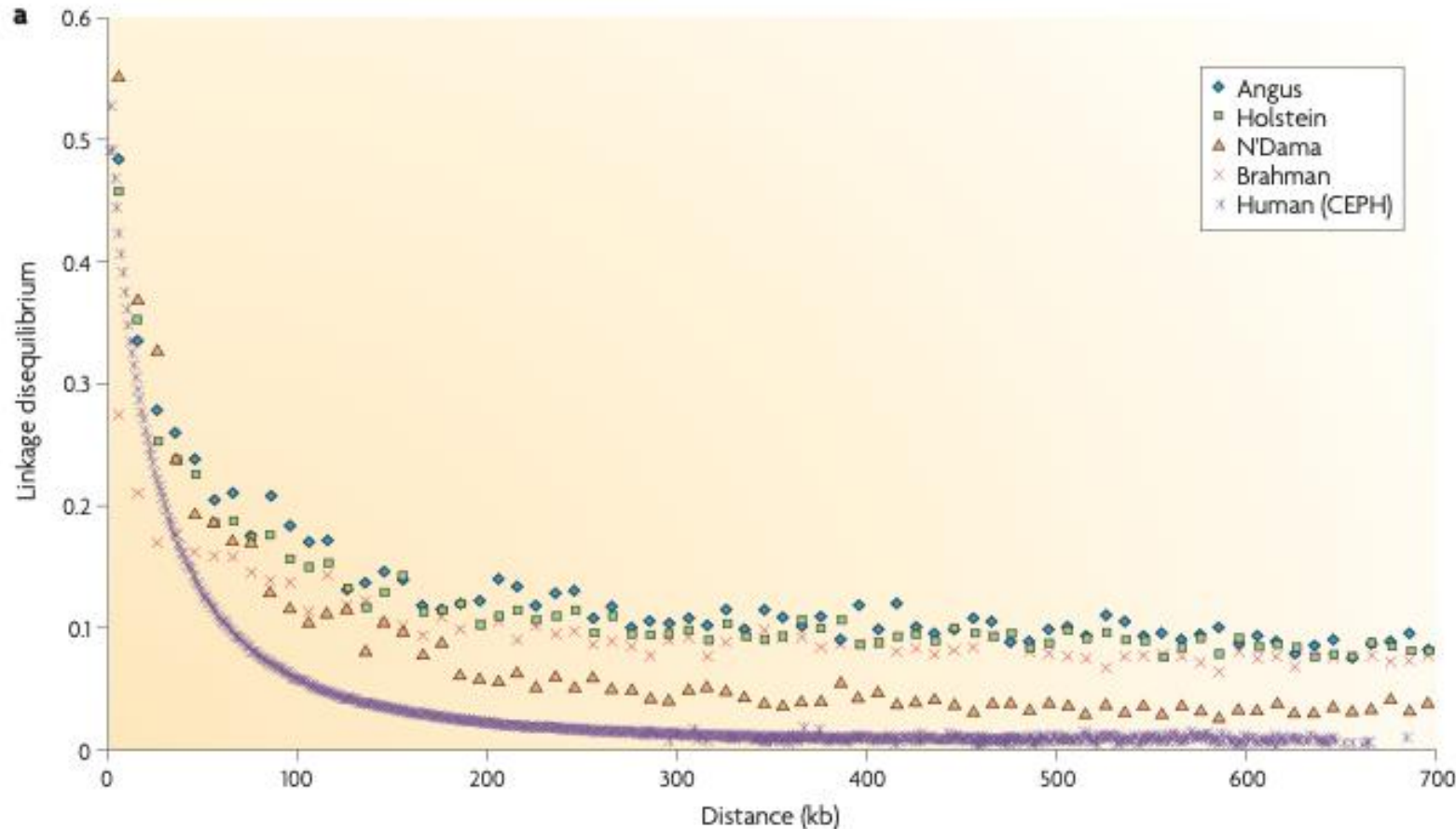
- **Population size**

- Livestock have small effective pop. Size >> High LD
 - Humans have large effective pop. Size >> Low LD



Foundation is linkage disequilibrium (LD)

Non-random association of alleles between loci

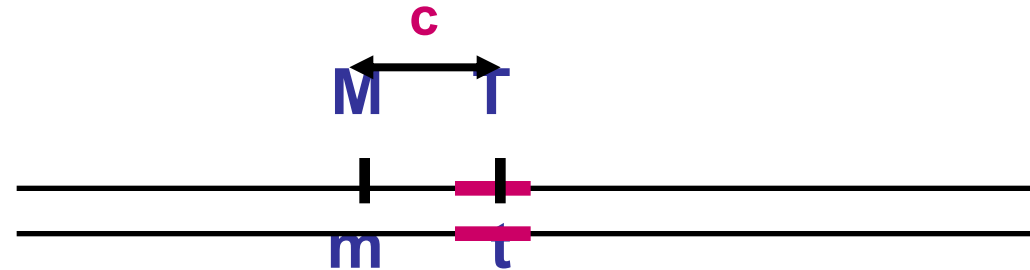


Wikipedia



SNP Association Issues

- LD



LD = correlation (r) M/m with T/t

$r \downarrow$ as $c \uparrow$

$E(r \mid c)$ ~ age of mutations
~ effective population size
~ population history, including selection

At population level: SNP & Causative mutation must be close



SNP Association Issues

- **Population Structure (stratification)**
 - Cause of MANY false positive associations



SNP Association Issues

- **causes of Population Structure:**
 - **Anything!!**
 - Breed
 - Families
 - Selection
 - Biased sampling
 - **Often hard to spot**



SNP Association Issues

- **Population Structure solution**
 - Use known pedigree or SNP Genotypes to detect & remove structure
- **$Y = Xb + Zu + \beta.SNP$**
- **β = regression (as before)**
- **u = vector of genetic effects**
 - **$\text{Var}(u) \sim A\sigma_a^2$ or $K\sigma_a^2$**
 - **Where K is matrix of covariances of genotypes**
 - **Where A is matrix of covariances by pedigree**



SNP Association Issues

- **Multiple Testing**
 - E.g. 50k or 750k statistical tests
 - Many “significant” results by chance alone
 - The most significant results have upwards bias
 - **Validate results in independent sample to get unbiased**



SNP Association Issues

- **Multiple Testing**
 - E.g. 50k or 750k statistical tests
 - Many “significant” results by chance alone
- **Solution: stringent thresholds**
 - Bonferroni: e.g. $0.05/n$ → 1×10^{-6} for 50k SNPs
 - Empirical thresholds from permutation

Need large studies to have power to declare associations significant (unless SNP effect large)



SNP Association Examples

- Same as for heritabilities
 - FIT SNP as fixed effect as well
- Additive variance due to SNP = $2pq[a + d(q-p)]^2$
- SNP heritability = $2pq[a + d(q-p)]^2/V_p$



- **Association analysis**

- Simple Mixed Model (animal model) – as in h^2 analyses:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

- Add one or all SNPs:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\mathbf{b} + \beta_i m_i + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\mathbf{b} + \sum \beta_i m_i + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

GWAS

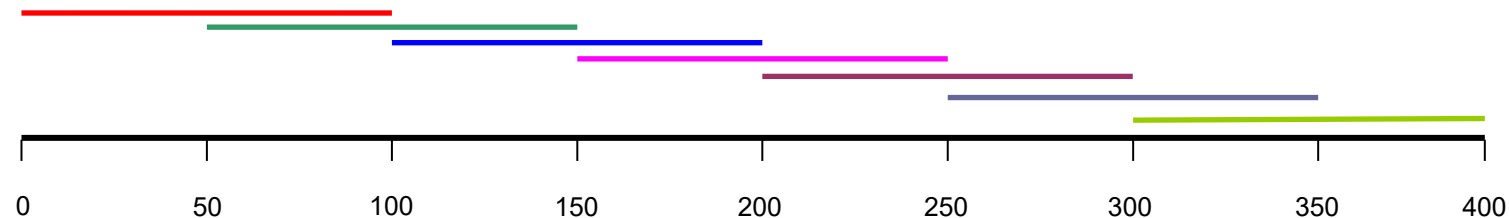
Relationship
matrix **G**

Genomic selection



Regional Heritability Mapping (RHM)

- Variance component approach
- Fit joint effects of all loci within a genomic region
- Each chromosome is divided into windows of a pre-defined number of SNPs:
 - In our case, 100 SNPs and window shifted every 50 SNPs





Regional Heritability Mapping (RHM)

$$y = Xb + Za + Zw + e$$

overall genetic effect

regional combined genetic effect

$$h^2 = \frac{(\sigma_a^2 + \sigma_w^2)}{(\sigma_a^2 + \sigma_w^2 + \sigma_e^2)}$$

Total h^2

$$h_w^2 = \frac{\sigma_w^2}{(\sigma_a^2 + \sigma_w^2 + \sigma_e^2)}$$

Regional h^2



GWAS methods

Single-marker regression

Multiple-marker model

- Multiple regressions
- Fit as a random effects via Bayesian regression methods
- Borrowed from GS
- Note equivalence; GWAS by GBLUP



Multiple testing correction

Bonferroni correction

- Significance level / Number of tests
- Conservative
- For 50 K SNP: $0.05/50000 = 10^{-6}$
- For 1 M SNP: $0.05/10^6 = 5 \times 10^{-8}$

Other options

- Limiting the false discovery rate (FDR)
- Permutation testing
- Effective number of independent tests
- Bayesian approach
- ...



Centre for
Tropical Livestock
Genetics and Health

Importance of quality control

Genotype call rate (SNP, individual)

Minor allele frequency

Departure from Hardy–Weinberg equilibrium

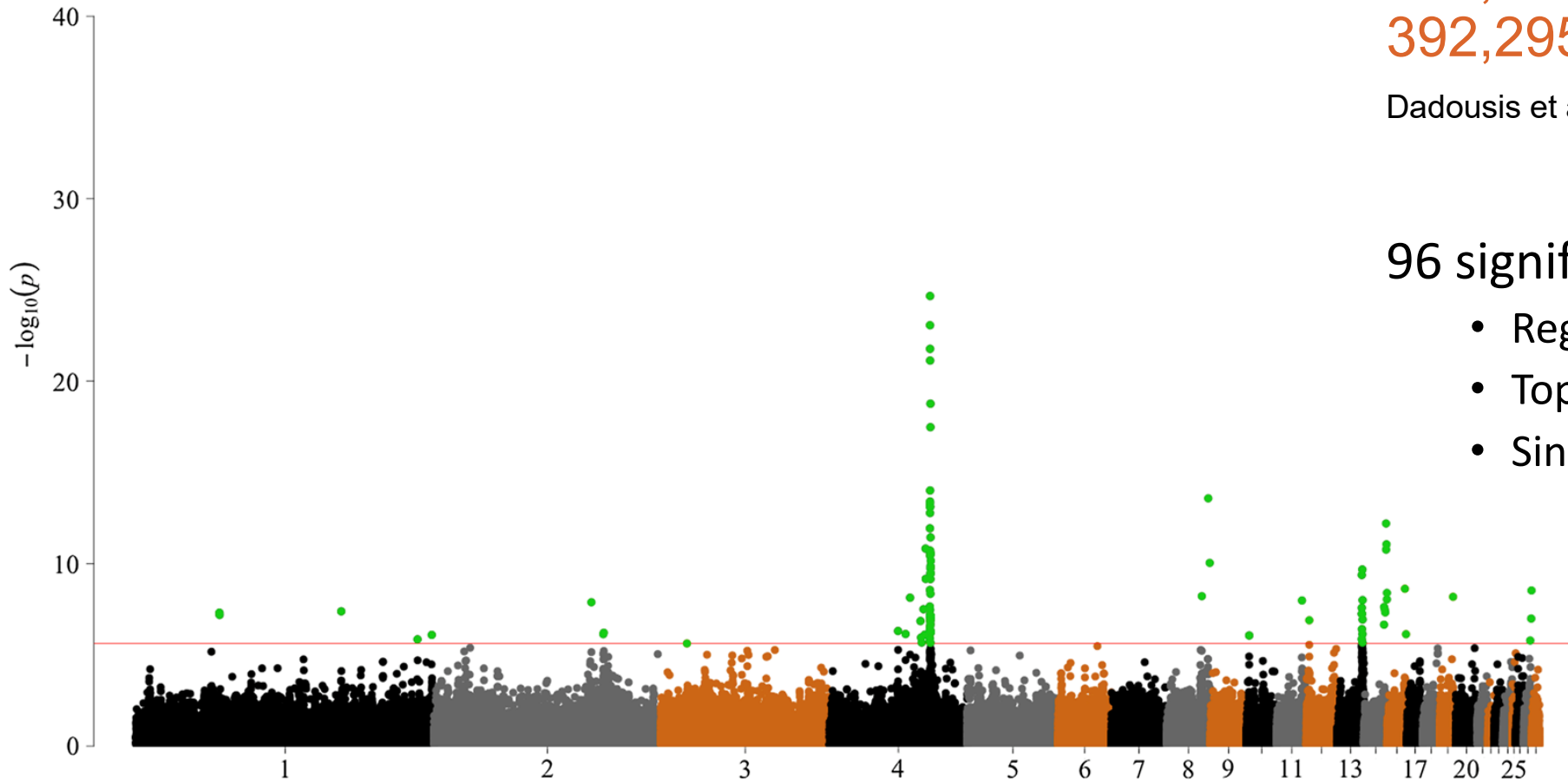
Don't forget QC of phenotypes!



Results: Manhattan plot

137,343 broiler chickens
392,295 imputed SNP

Dadousis et al., 2021

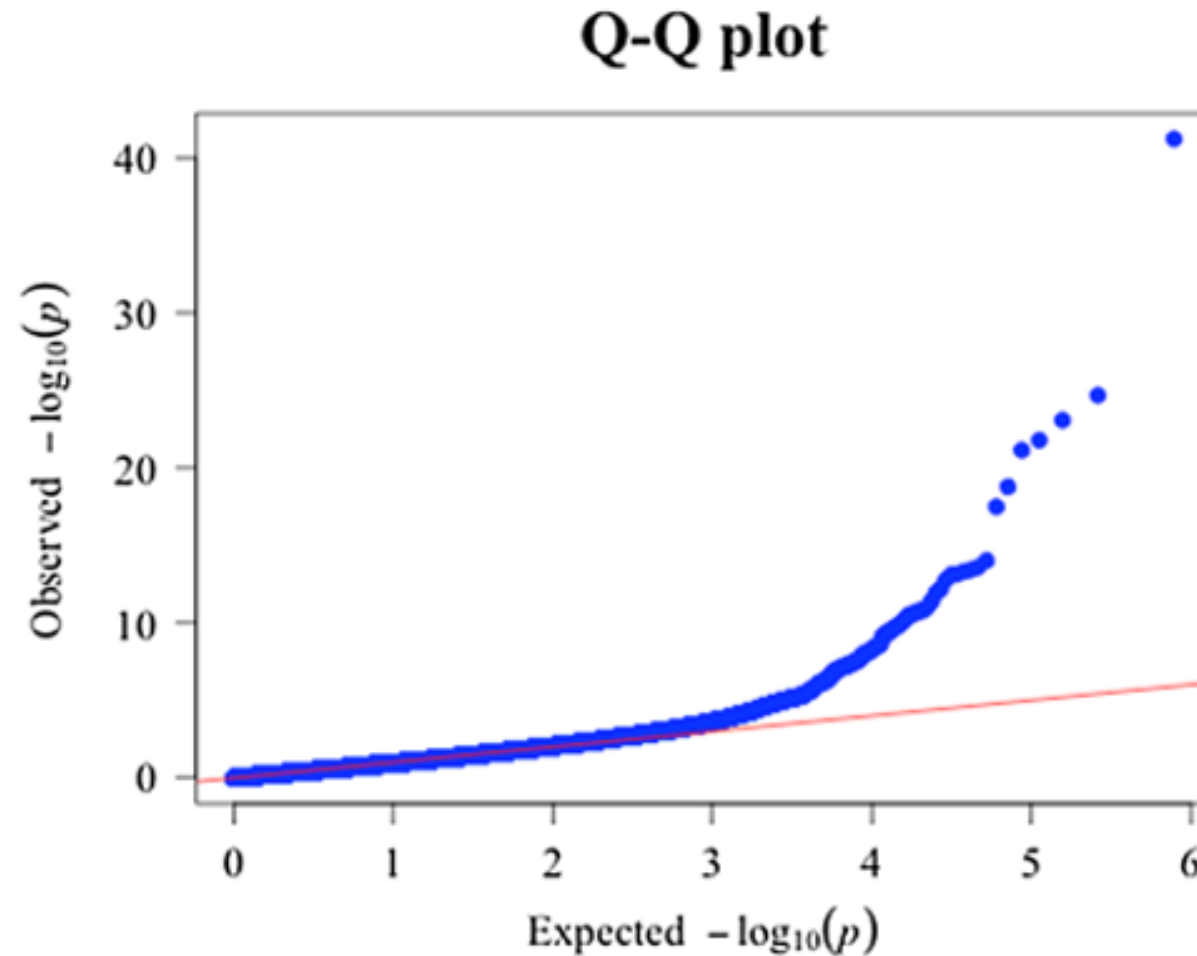


96 significant SNP in 25 regions

- Regions explained 30% V_A
- Top region 4.37% V_A
- Single top SNP 1.9% V_A



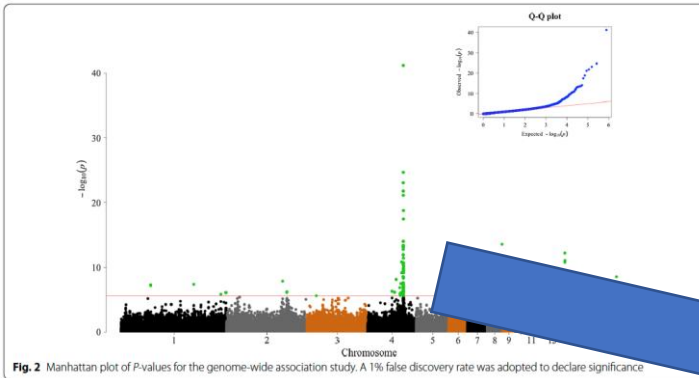
Results: Q-Q (quantile-quantile) plots





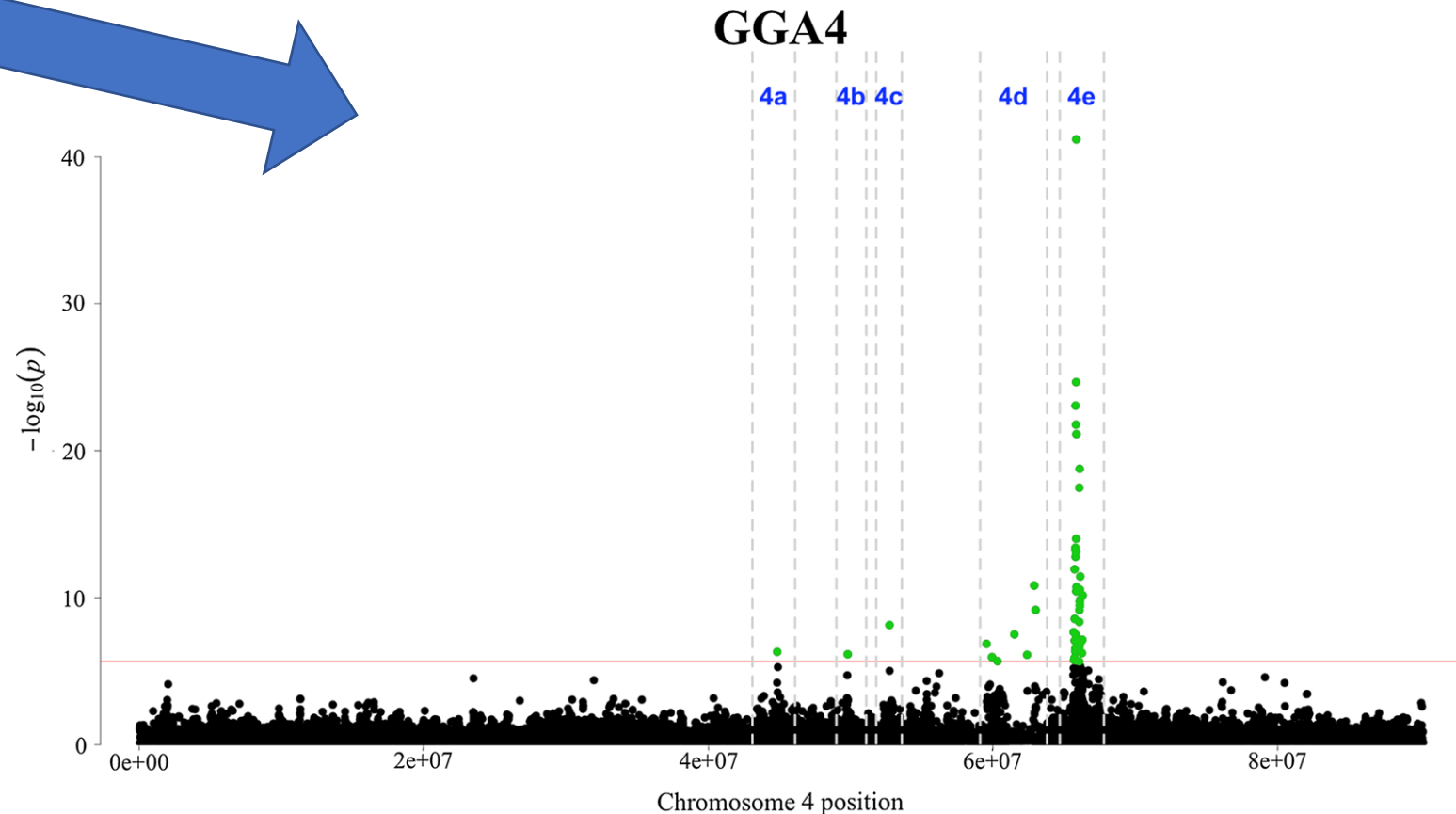
Centre for
Tropical Livestock
Genetics and Health

Results: Manhattan plot



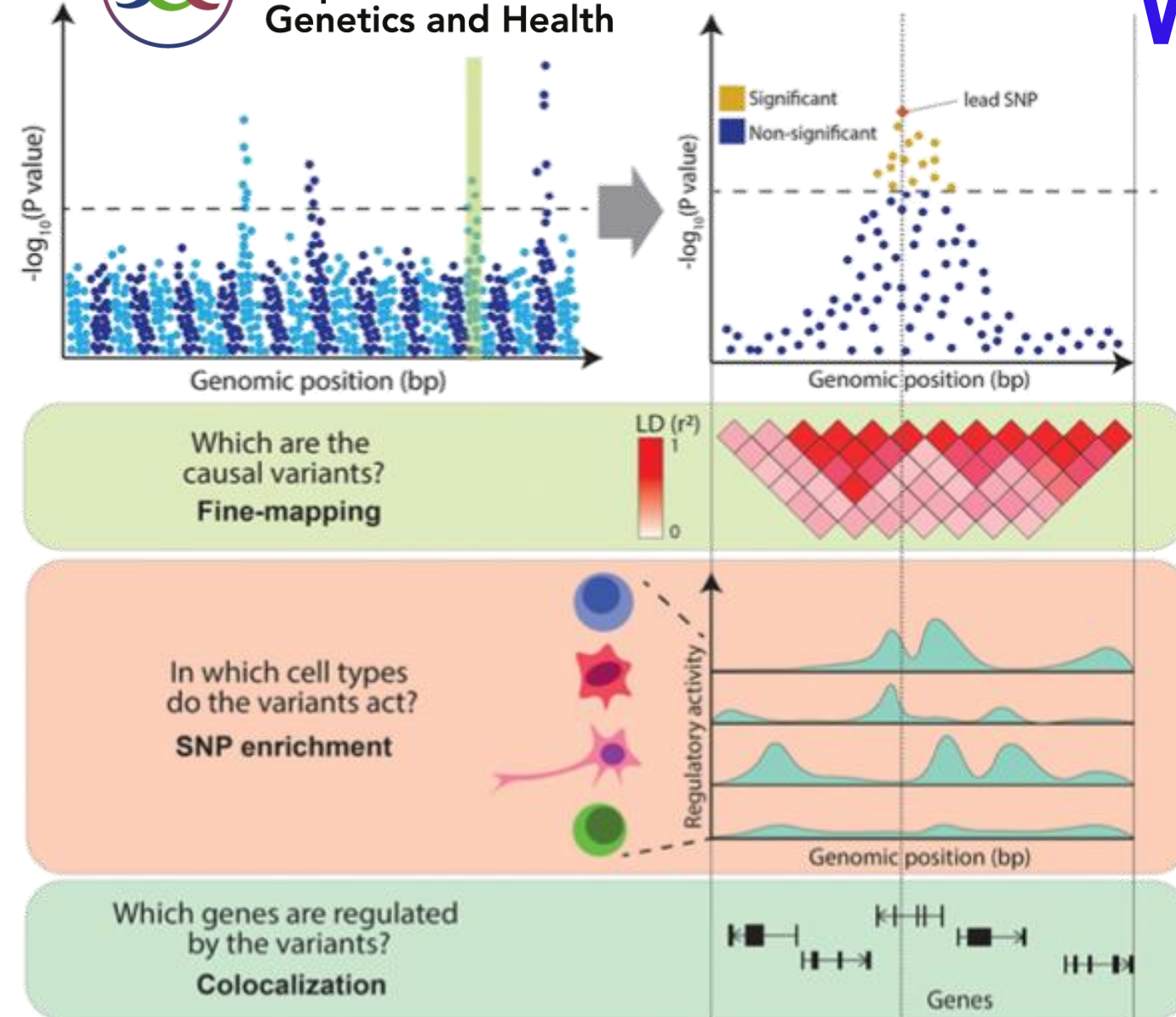
137,343 broiler chickens
392,295 imputed SNP

Dadousis et al., 2021





What's after GWAS?



GWAS association \neq causal variant
Biological validation and causal variants?

Replications? Fine mapping?

Functional analysis? eQTL? Genome editing?

Is GWAS able to enrich the genomic evaluations?



Centre for
Tropical Livestock
Genetics and Health

What is the future of GWAS?

I will leave it up to you


Five Years of GWAS Discovery

Peter M. Visscher,^{1,2,*} Matthew A. Brown,¹ Mark I. McCarthy,^{3,4} and Jian Yang⁵

10 Years of GWAS Discovery: Biology, Function, and Translation

Peter M. Visscher,^{1,2,*} Naomi R. Wray,^{1,2} Qian Zhang,¹ Pamela Sklar,³ Mark I. McCarthy,^{4,5,6} Matthew A. Brown,⁷ and Jian Yang^{1,2}

Evidence for and localization of proposed causative variants in cattle and pig genomes

Martin Johnsson^{1*}  and Melissa K. Jungnickel²

Status and prospects of genome-wide association studies in plants

Laura Tibbs Cortes¹  | Zhiwu Zhang²  | Jianming Yu¹ 



Animal QTL data base



Cattle QTL

There are **163,725** QTL from **1,069** publications curated into the database. Those QTL represent **685** different traits (see **data summary** for details).



Catfish QTL

There are **0** QTL from **0** publications curated into the database. Those QTL represent **0** different traits (see **data summary** for details).



Chicken QTL

There are **15,475** QTL from **362** publications curated into the database. Those QTL represent **442** different traits (see **data summary** for details).



Goat QTL NEW

There are **64** QTL from **3** publications curated into the database. Those QTL represent **21** different traits (see **data summary** for details).



Horse QTL

There are **2,473** QTL from **99** publications curated into the database. Those QTL represent **62** different traits (see **data summary** for details).



Pig QTL

There are **33,540** QTL from **745** publications curated into the database. Those QTL represent **704** different traits (see **data summary** for details).



Rainbow Trout QTL

There are **1,372** QTL from **17** publications curated into the database. Those QTL represent **22** different traits (see **data summary** for most recent updates).



Sheep QTL

There are **3,752** QTL from **201** publications curated into the database. Those QTL represent **274** different traits (see **data summary** for most recent updates).



Top 15 QTL/associations

Traits	Number of QTL
Age at puberty	10,623
Scrotal circumference	10,457
Milk fat percentage	8,117
Milk fat yield	6,957
Milk protein percentage	4,999
Milk C14 index	4,847
Milk kappa-casein percentage	4,836
Metabolic body weight	4,275
Milk yield	3,835
Percentage normal sperm	3,596
Calving ease	3,540
Average daily gain	3,504
Milk myristoleic acid content	3,313
Milk protein yield	3,095
Milk glycosylated kappa-casein percentage	2,753

<https://www.animalgenome.org/cgi-bin/QTLdb/index>



Centre for
Tropical Livestock
Genetics and Health

Examples



Change of perspective

Infinitesimal model:

- Infinite number of additive loci; each with a small effect

Finite loci model:

- Finite amount of DNA and genes
- How many?
- Small/big effect?



Basic statistical model

Phenotype = f(genotype, e)

$$y_i = \mu + bx_i + e_i$$

Simple but many analysis

- 50 K SNP we run 50 K regressions, 100 K, 1M, ...

Put in matrix notation

$$y = 1\mu + xb + e$$
$$y = Xb + e$$



Relatedness and population structure

Cause of spurious associations

Account for it in a model:

- As a random effect;
 - Relationship matrix (**A**, **G**, ?)
- As a covariates;
 - Subpopulations (structured association; STRUCTURE)
 - PCA analysis (top PCs)

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{e} \\ \text{Var}(\mathbf{a}) &= \mathbf{G}\sigma_a^2 \\ \text{Var}(\mathbf{e}) &= \mathbf{I}\sigma_e^2 \end{aligned}$$