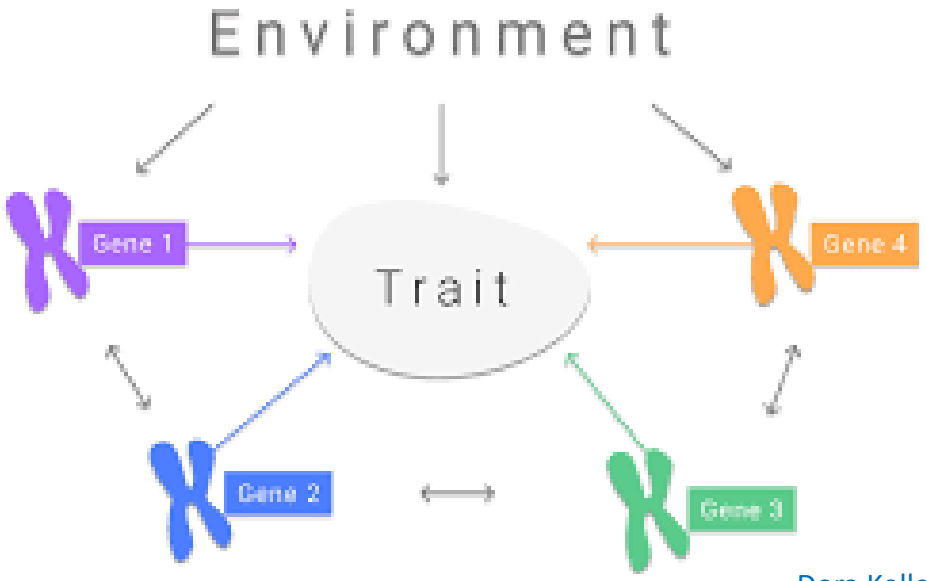
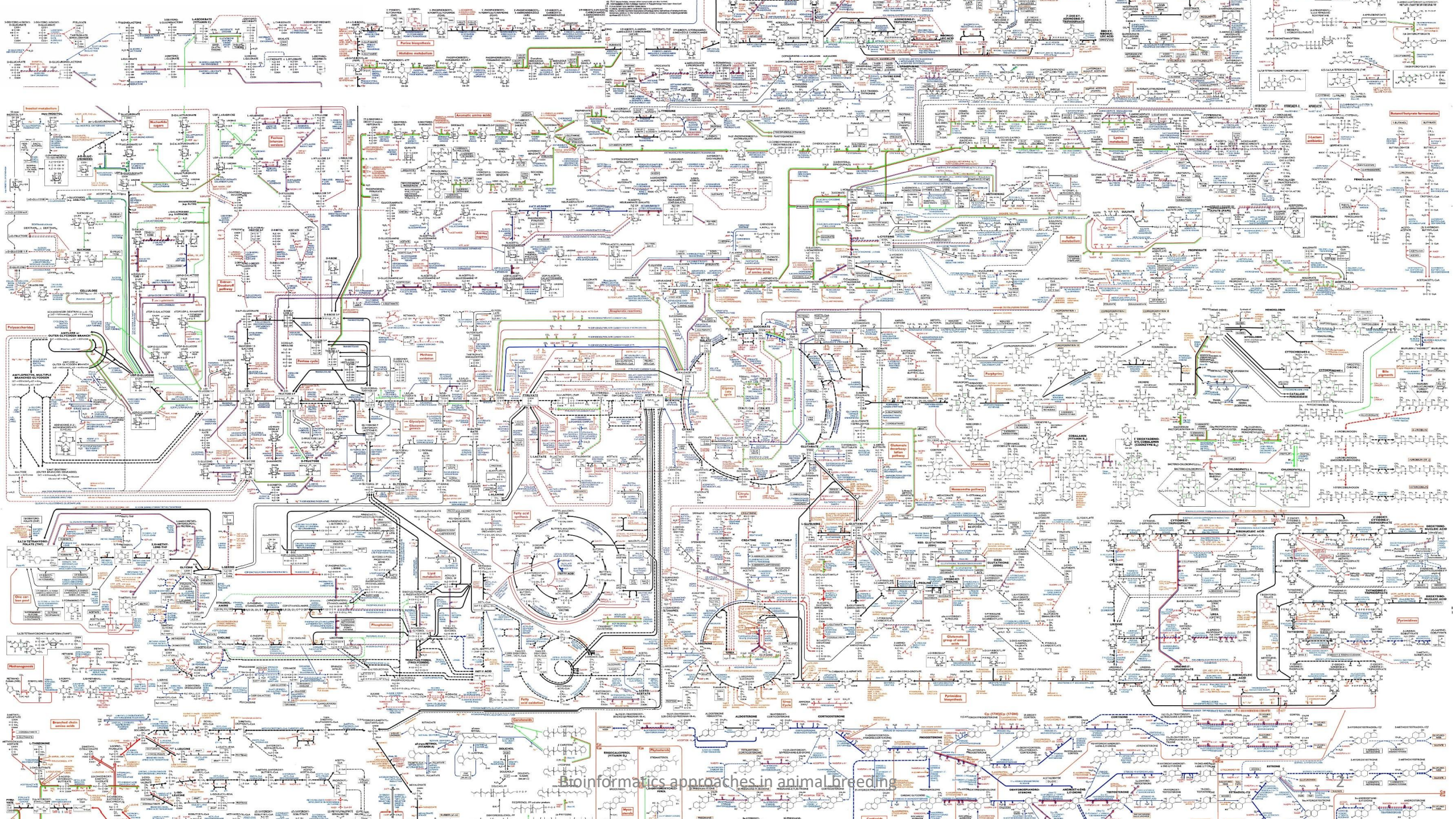
Genetic background of complex traits



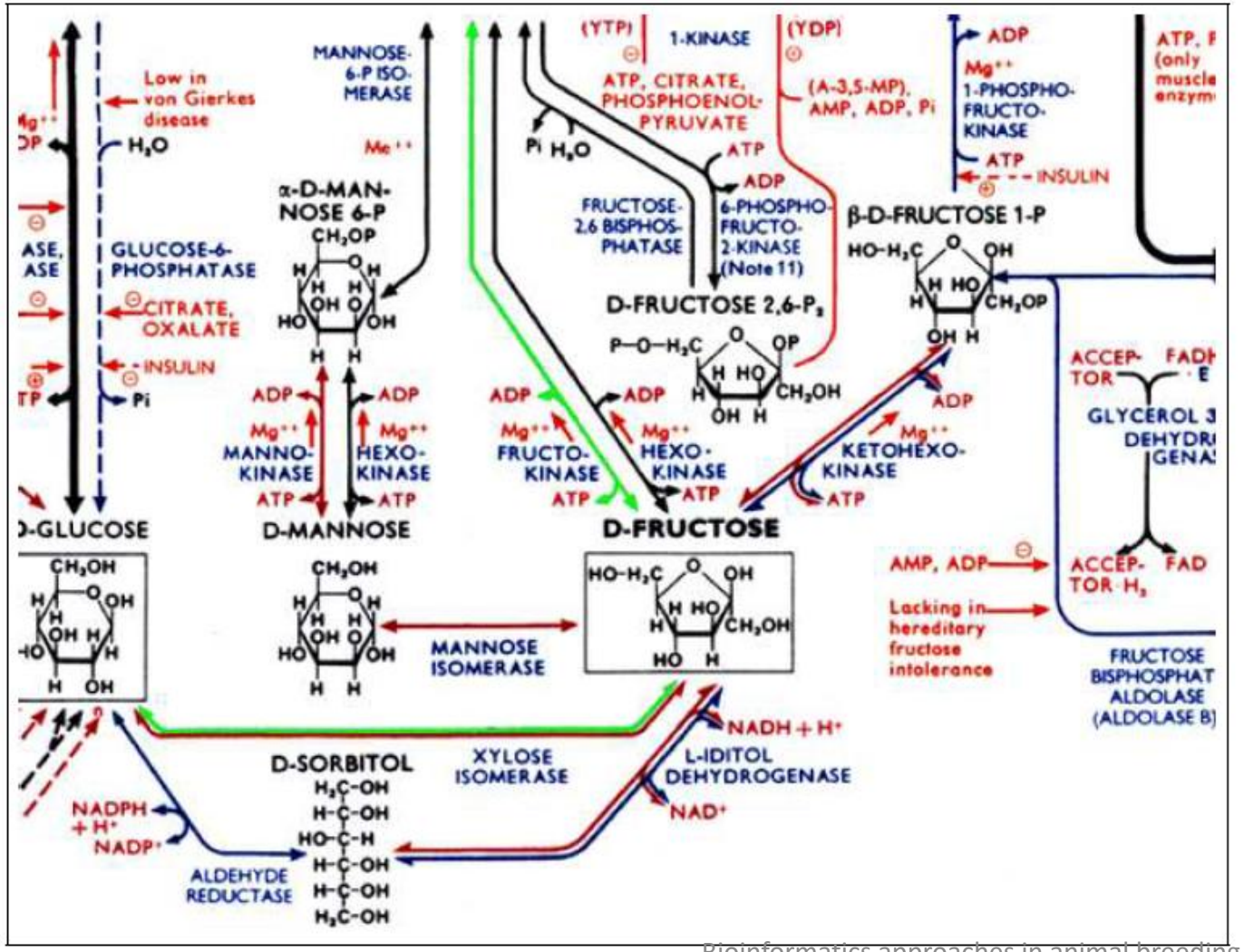
Dora Koller

Peter DOVC

University of Ljubljana, Biotechnical Faculty



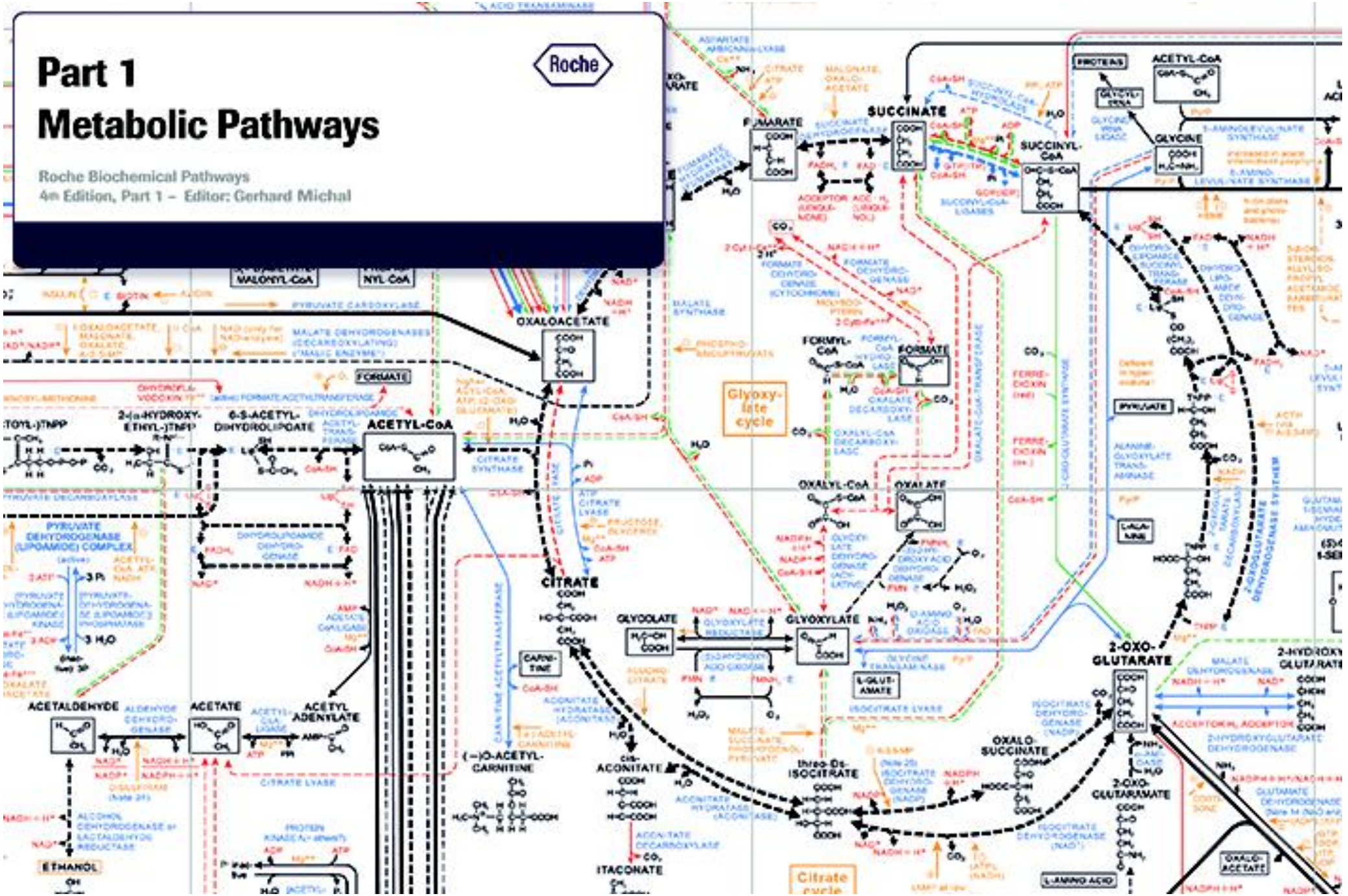
The Biochemical Pathways chart



- The original large poster was drawn up by Gerhard Michal of the Boehringer Mannheim company that gives a cross-section of general metabolism in various species and organs. It's a classic poster that adorns the walls wherever biochemists are found.
- The electronic version of the Biochemical Pathways was produced in the Cllins Laboratory at Cornell University.

Part 1
Metabolic Pathways

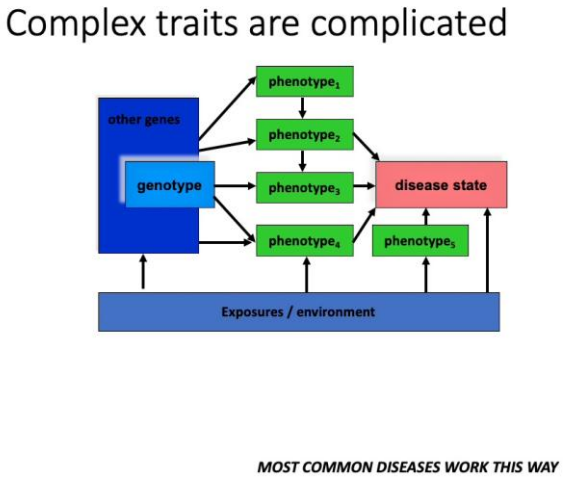
Roche Biochemical Pathways
4th Edition, Part 1 – Editor: Gerhard Michal



Complex or quantitative traits

Examples of complex traits in different fields:

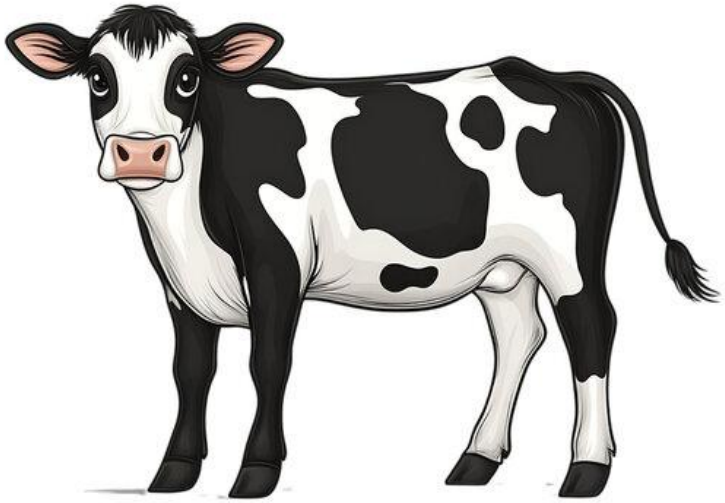
- medicine: e.g. diabetes
- agriculture: e.g. milk yield
- evolution: e.g. body size



What is the base of their complexity?

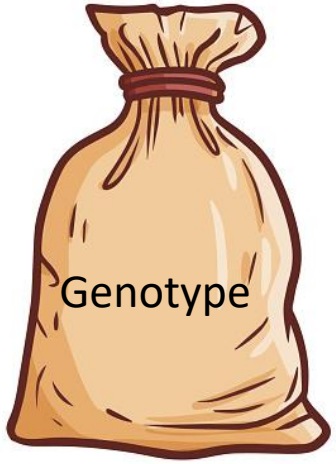
- They are controlled by many genes and by environmental factors
- Although genetics of quantitative traits has been studied for over 100 years, very few of polymorphisms that cause variation in these traits were known until recently.

Black box approach



Phenotype

=



+

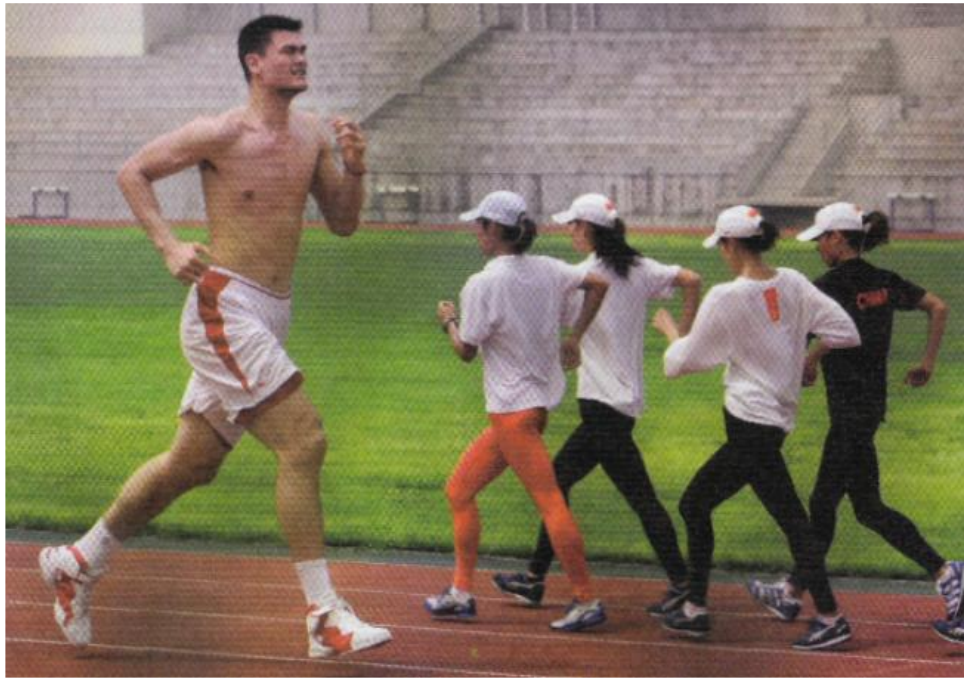
Simple genetic model for study of complex traits

$X = \bar{X} + g + e$

$x = g + e$

Yao Ming: 229cm

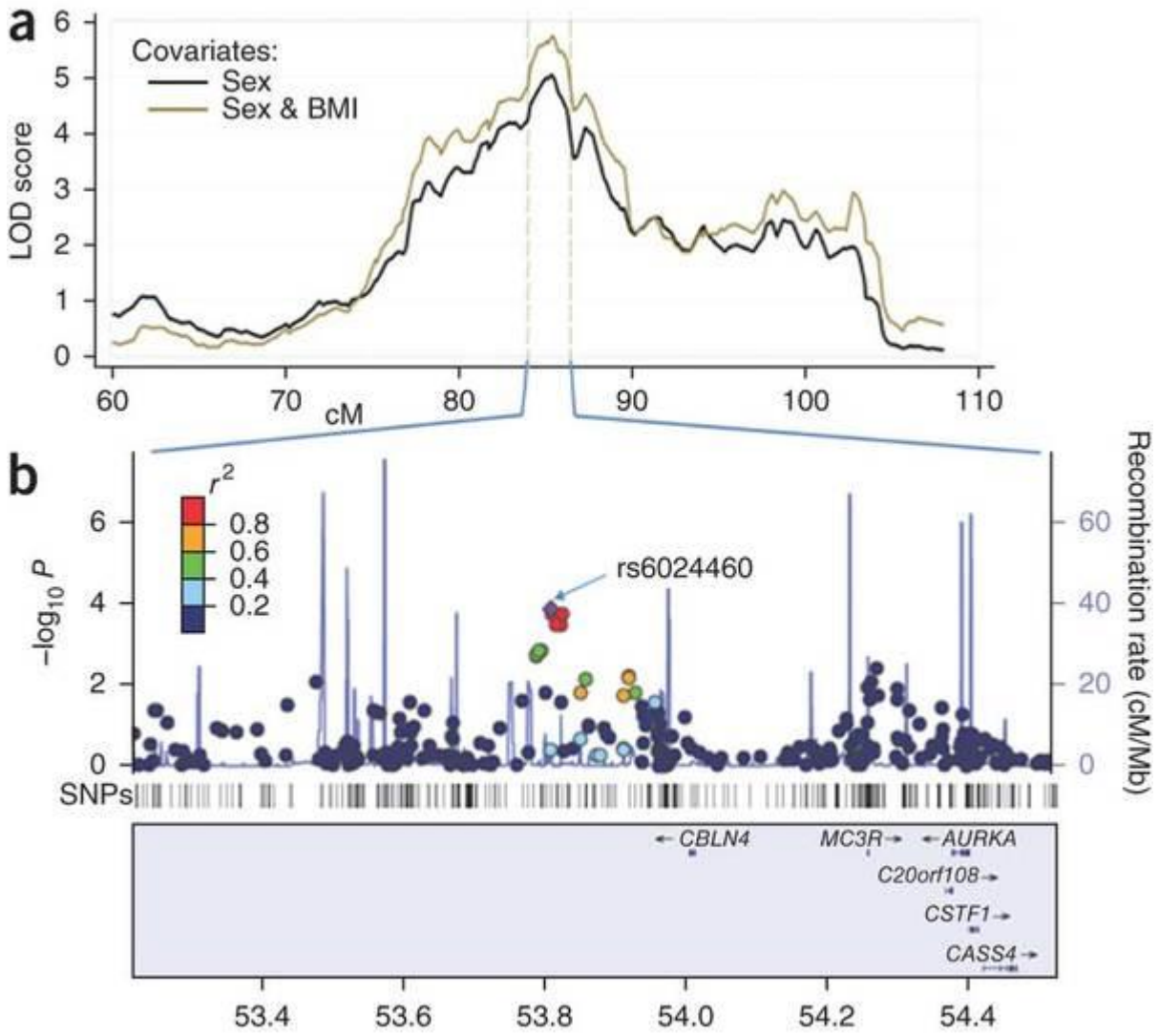
$g + e = 59$cm
Mean of clones: 212cm
$E(e) = 0$
$212 - 170 = 42$cm
$g = 42$cm
$229 = 170 + 42 + 17$



Study of complex traits is technology driven

- Development of **SNP chip technology** enabled determination of genotype at thousands of single nucleotide positions.
- SNP markers might be neutral polymorphisms with no effect on the traits studied, but LD between the SNP and the causal polymorphism generates an association between the traits and some SNPs.
- GWAS which use a genome-wide panel of SNPs revealed thousands of associations between SNPs and complex traits.
- GWAS are intended to map the causal polymorphisms to a genomic region but not to identify the causal gene/mutation.

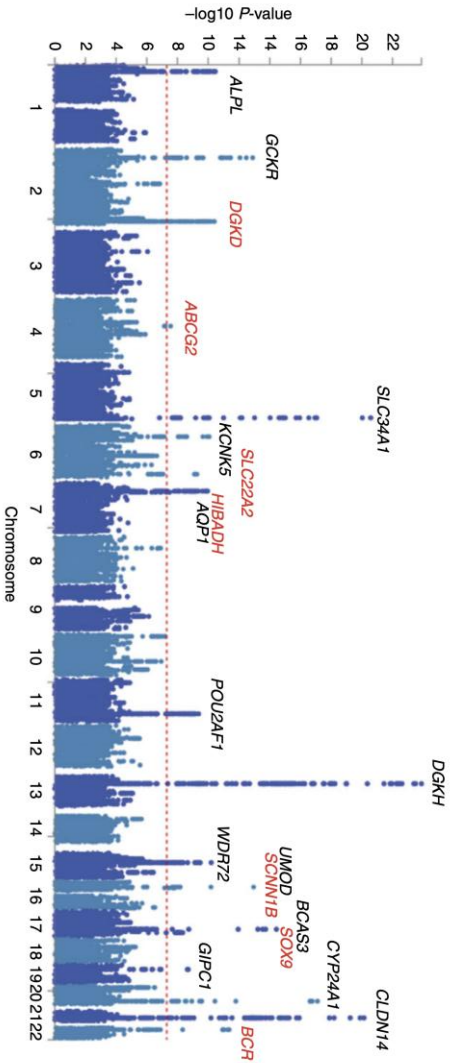
QTL studies

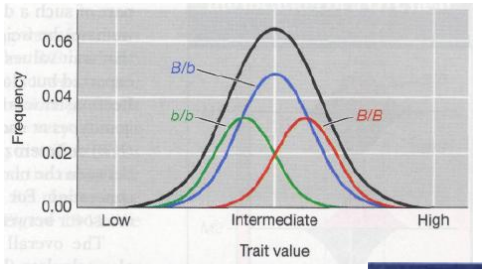


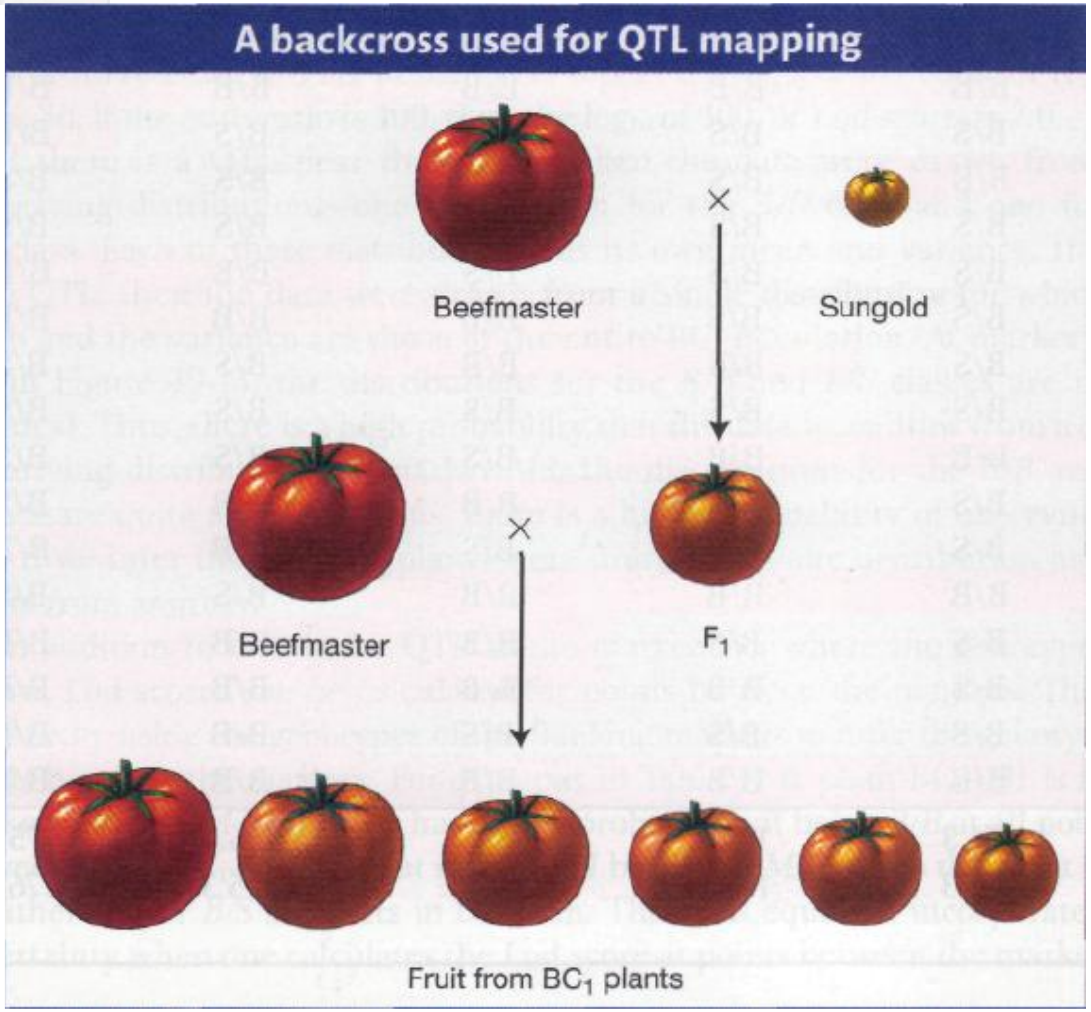
- In the first stage MS and SNPs were used
- QTL intervals are quite big (several cM)
- Discovered QTL are often population specific
- Analysis of QTL is hampered by epistasis, population stratification, non-random mating, environmental effects…

How the data from GWAS can be used?

- They can be used to predict future phenotypes. In agriculture, it is usually the phenotype of the offspring of animals or plants that we wish to predict. Individuals with the best **breeding value** can be selected as parents of the next generation. In human medicine, this might be the probability that a person will develop type 2 diabetes in the future. Although the SNPs used may have no causal relationship with the trait, they may still be useful for prediction due to their LD with causal variants.
- GWAS data are used to map the causal variants to a region of the genome. We expect that this information will help to identify the causal polymorphism. This increases our understanding of the biology of complex traits and may suggest methods of controlling them such as new drug targets.
- GWAS data provide an overview of the genetic architecture of complex traits that is useful in medicine, agriculture and evolution. We would like to know how many polymorphisms control a trait, what are their effects and allele frequencies, the LD between them, and how they evolve.



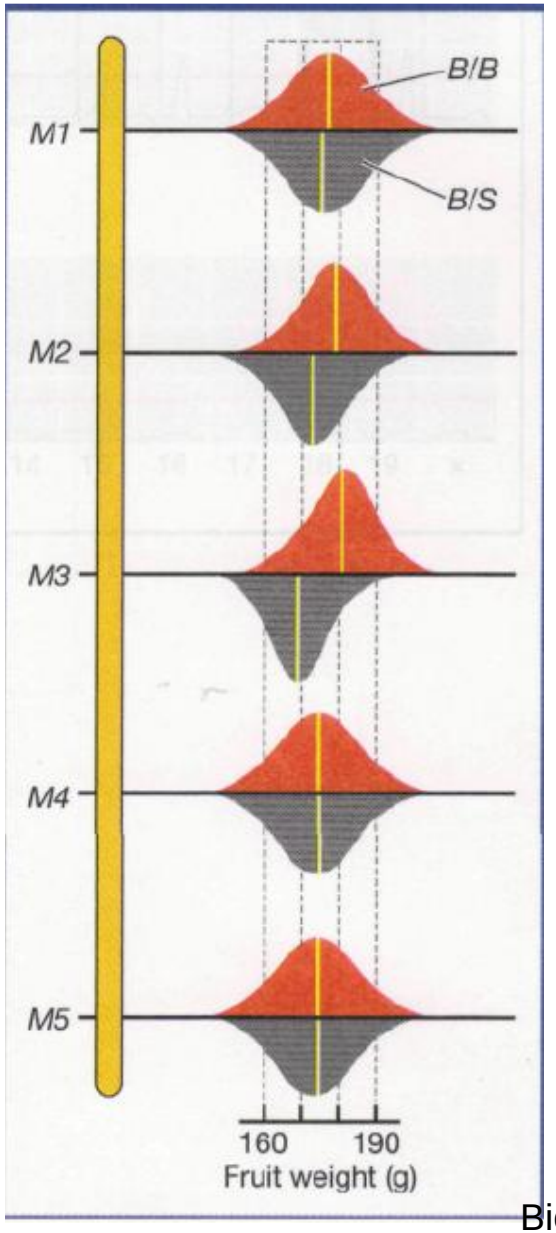

0.06
0.04
0.02
0.00
Frequency
B/b
b/b
B/B
Low
Intermediate
High
Trait value


A backcross used for QTL mapping
Beefmaster
Sungold
Beefmaster
F1
Fruit from BC1 plants

QTL mapping

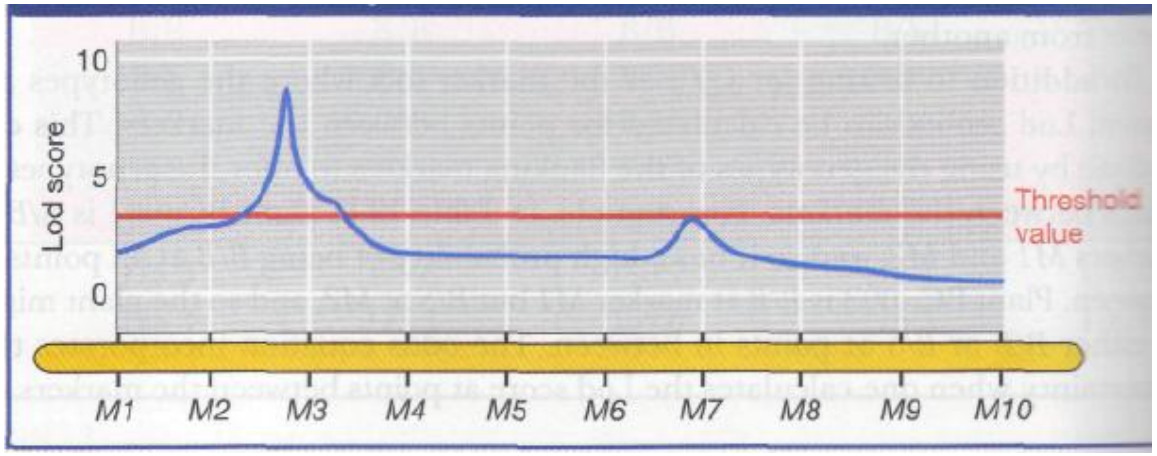
Plant	Fruit wt. (g)	Markers				
		M1	*M2*	*M3*	*M4*	*M5*
Beefmaster	230	B/B	B/B	B/B	B/B	B/B
Sungold	10	S/S	S/S	S/S	S/S	S/S
BC_1-001	183	B/B	B/B	B/B	B/S	B/S
BC_1-002	176	B/S	B/S	B/B	B/B	B/B
BC_1-003	170	B/B	B/S	B/S	B/S	B/S
BC_1-004	185	B/B	B/B	B/B	B/S	B/S
BC_1-005	182	B/B	B/B	B/B	B/B	B/B
BC_1-006	170	B/S	B/S	B/S	B/S	B/B
BC_1-007	170	B/B	B/S	B/S	B/S	B/S
BC_1-008	174	B/S	B/S	B/S	B/S	B/S
BC_1-009	171	B/S	B/S	B/S	B/B	B/B
BC_1-010	180	B/S	B/S	B/B	B/B	B/B
BC_1-011	185	B/S	B/B	B/B	B/S	B/S
BC_1-012	169	B/S	B/S	B/S	B/S	B/S
BC_1-013	165	B/B	B/B	B/S	B/S	B/S
BC_1-014	181	B/S	B/S	B/B	B/B	B/S
BC_1-015	169	B/S	B/S	B/S	B/B	B/B
BC_1-016	182	B/B	B/B	B/B	B/S	B/S
BC_1-017	179	B/S	B/S	B/B	B/B	B/B
BC_1-018	182	B/S	B/B	B/B	B/B	B/B
BC_1-019	168	B/S	B/S	B/S	B/B	B/B
BC_1-020	173	B/B	B/B	B/B	B/B	B/B
Mean of B/B	–	176.3	179.6	180.7	176.1	175.0
Mean of B/S	–	175.3	173.1	169.6	175.3	176.4
Overall mean	175.7					

LOD score

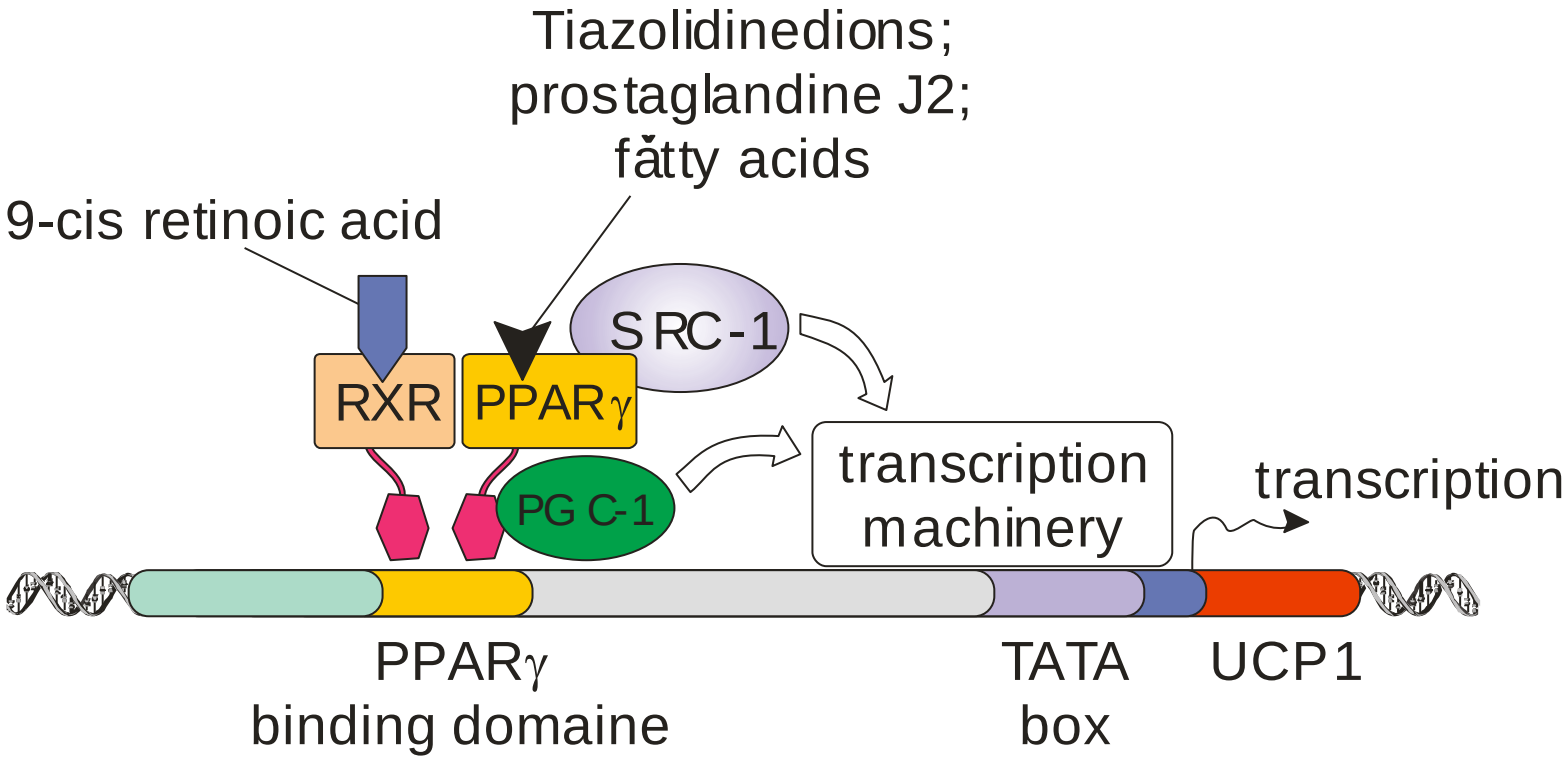


$$\text{odds} = \frac{\text{Prob (data|QTL)}}{\text{Prob (data|no QTL)}}$$

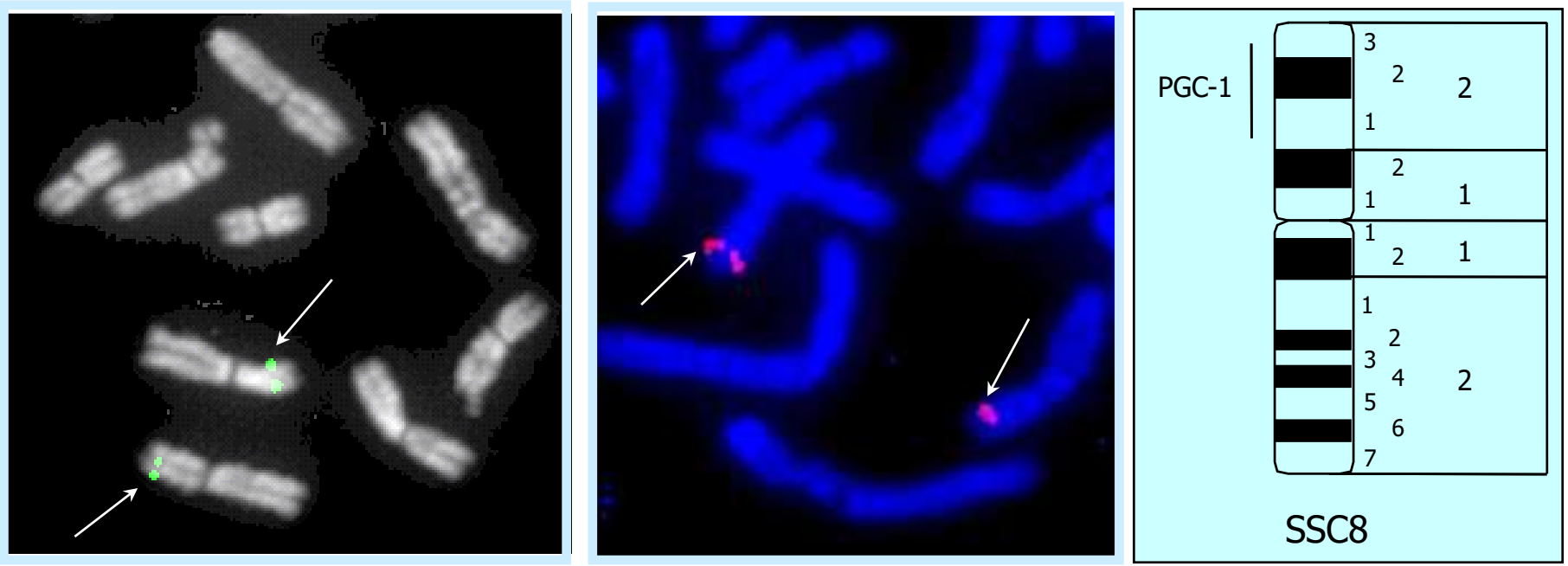
	Fruit weights			Effects	
	B/B	*B/S*	*S/S*	A	D
QTL 1	180	170	160	10	0
QTL 2	200	185	110	45	30



PGC-1 is co-activator of PPARγ and regulates UCP1 expression



Chromosomal assignment of the porcine *PGC-1* to SSC8, p2.1-2.3



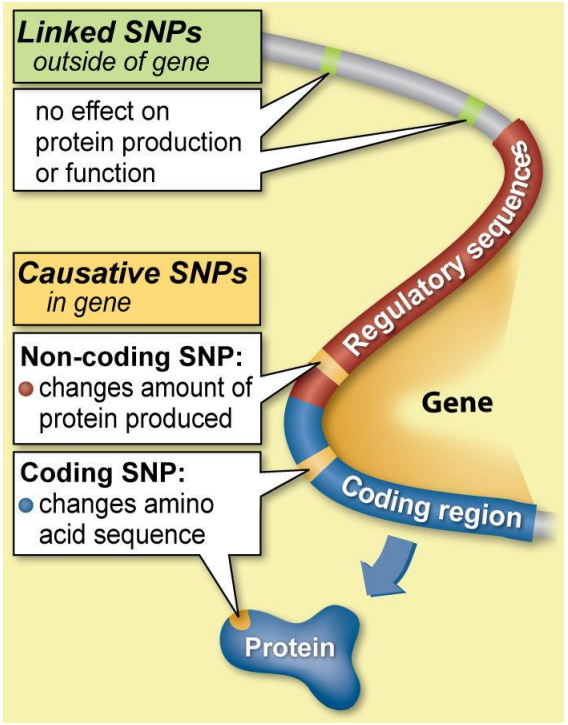
Fluorescent *in situ* hybridisation (FISH)

Transversion A - T at position 1290 in the exon 8 causes amino acid substitution Ser – Cys and is breed specific

BAC230p21	GACCAGAGCTACC
Swedish landrace	GACCAGAGCTACC
German landrace	GACCAGAGCTACC
Pietrain	GACCAGAGCTACC
Duroc	GACCAGAGCTACC
Yorkshire	GACCAGAGCTACC
BAC234005	GACCAGTGCTACC
Goettinger mini pig	GACCAGTGCTACC
Mangalica	GACCAGTGCTACC

Different design of the GWAS for different purposes

- Mapping causal polymorphisms is usually done by fitting one SNP at a time in a regression model.
- Predicting genetic value is most often done by assuming all SNPs have an effect drawn from the same normal distribution.
- SNPs can be assumed to have an effect drawn from a mixture of normal distributions with increasing variances. They can be used for genomic prediction, mapping of causal variants and inference on the genetic architecture of complex traits.



Prediction of genetic value

- In many datasets, the number of SNPs (p) is greater than the number of individuals with records (n). Consequently, if the effects of the SNPs on the trait are treated as **fixed effects** in a multiple regression analysis, there is **no unique solution**.
- The total variance explained by all SNPs (a result of their effect sizes and allele frequencies) must be **less** than the total genetic variance, this represents a restriction of effect sizes.
- More accurate predictions can be obtained by treating the effect sizes **as random variables** drawn from a distribution which is consistent with the total genetic variance. The best prediction of a random variable (g) from a set of predictors (x) is the expected or average value of g conditional on the values observed for x.
- As there are typically thousands of SNPs, the distribution of SNP effects on a trait is simply the distribution of their thousands of effects: many may have no effect at all and some may have a large effect.

Prediction of genetic value

- p(y/b, x) is the likelihood of the phenotypes (y) given the genotypes of the individual (x) and the effects of the SNPs (b). This method was invented by Meuwissen et al. and is called genomic selection or **genomic prediction**.
- The statistical analysis resulting from applying this prediction rule depends on the prior distribution chosen for SNP effects (b).
  - b is assumed to be normally and independently distributed with a mean of 0 and a variance that is same for all SNPs: this method is an example of best linear unbiased prediction (BLUP).
  - Distribution of b can be a mixture of zero and a t distribution. The 'Bayes R' method is a further development, which uses a prior distribution for b which is a mixture of four normal distributions each with zero mean but with variances of 0, 0:0001 σ^2 g, 0:001 σ^2 g and 0:01 σ^2 g.
- **While BLUP is a linear method in that b is estimated by a linear combination of the phenotypic data y, the Bayes R method is non-linear in y**.

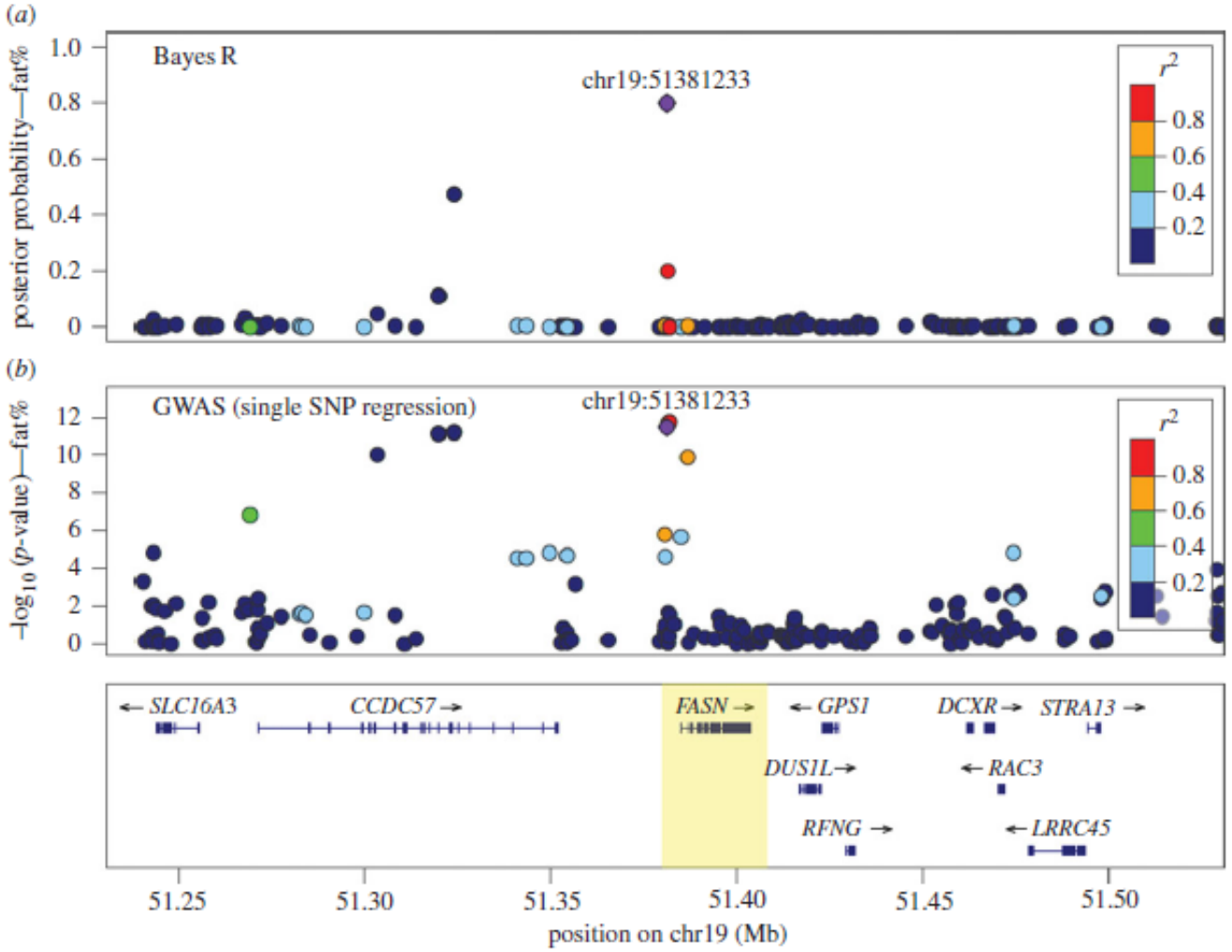
Distribution of SNP effects

- The BLUP prior corresponds to a ‘pseudo-infinitesimal’ model in which all polymorphic sites in the genome have an effect on every trait, and all effects are of similar magnitude and very small. For instance, if there are 1 million SNPs, each one is assumed to explain approximately 10^{-6} of the genetic variance σ^2 g. As a consequence, **all estimated SNP effects shrink towards 0 when BLUP is used.**

- Bayes R model allows the distribution of SNP effects to depart from this pseudo-infinitesimal distribution, with **some SNPs having zero effect and some SNPs having a large effect on the trait.**

Genome-wide analysis of bovine milk fat percentage

(results for a region around the FASN gene)



(a) Posterior probability that an SNP has a nonzero effect from Bayes R where all SNPs are fitted in the model simultaneously.

(b) (b) -$\log_{10}$ p-value from GWAS single SNP regressions.

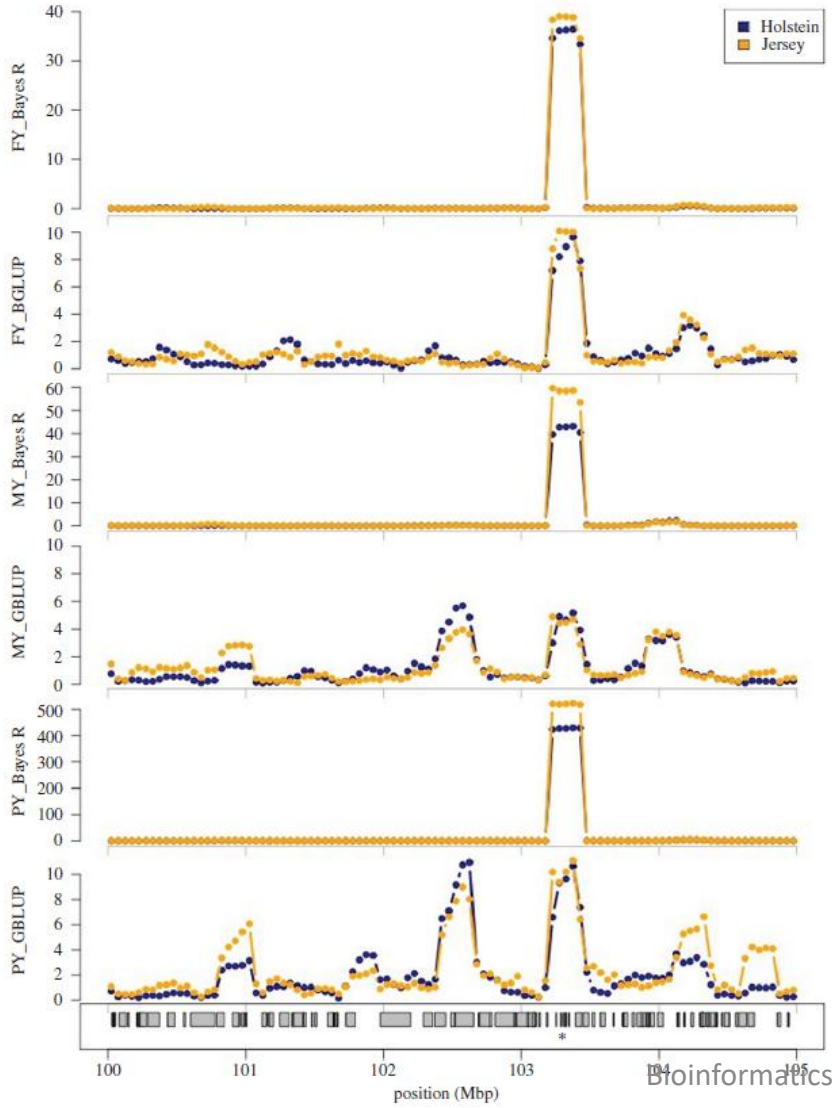
The top Bayes R variant is annotated (with base pair position) and shown as a purple diamond, and the strength of LD (r^2) between this and all other variants is colour coded.

In the Bayes R model, a single SNP just upstream of the FASN gene has the highest posterior probability, while in the GWAS model there are several SNP extending across to the CCDC57 (coiled-coil domain containing 57) gene region with almost equally significant effects.

FASN is a key enzyme in *de novo* fatty acid biosynthesis and highly expressed in lactating bovine mammary tissue.

Goddard et al, 2016

Local GEBV variance near the PAEP gene for FY, MY and PY using Bayes R and BLUP



Goddard et al, 2016

GEBV variance in overlapping 250 kb windows for Holstein and Jersey reference animals from SNP effects estimated from the multi-breed reference population.

- Traits: FY, fat yield; MY, milk yield; PY, protein yield.
- The position of PAEP on BTA11 is marked (*).

History of complex traits in medicine

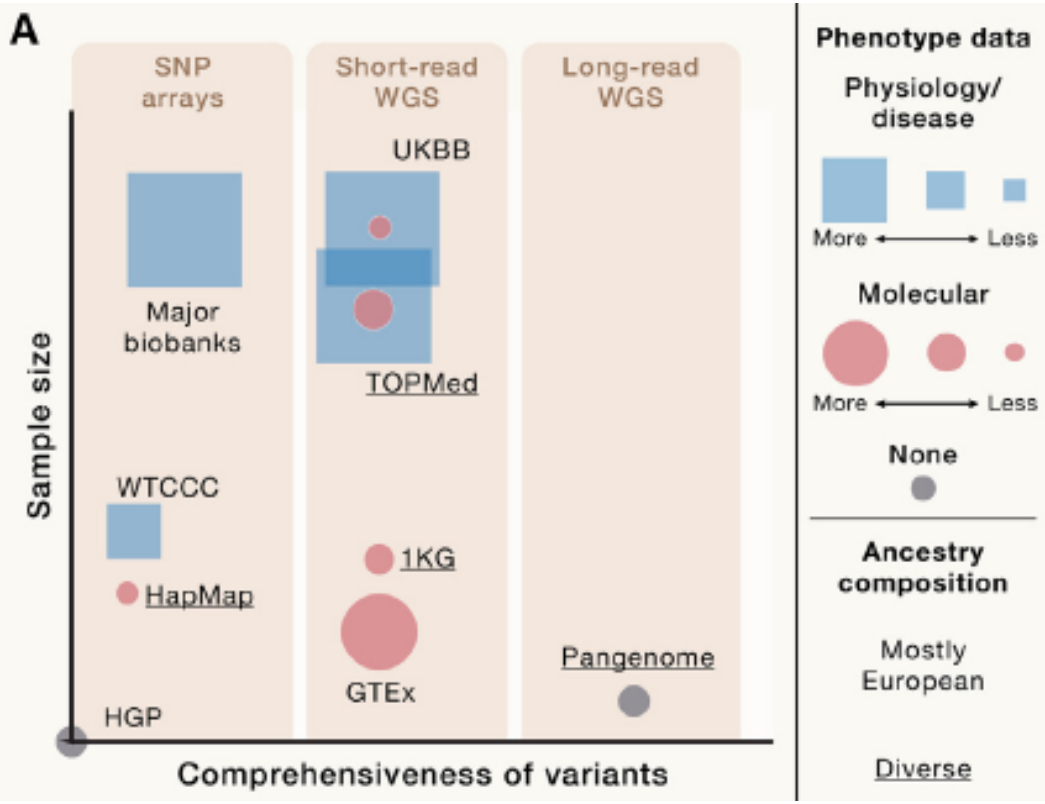
The progress during the 20th century in understanding the structure and content of genetic informatin can be devided in four phases:

- discovery of chromosomes
- defining the molecular structure of DNA
- discovery of the molecular machinery of gene function
- determining the sequence of entire genes, scaffolds and genomes

Major achievements in the 21st century:

- characterization of genetic variation in human populations and discovery of their contribution to phenotypic variation

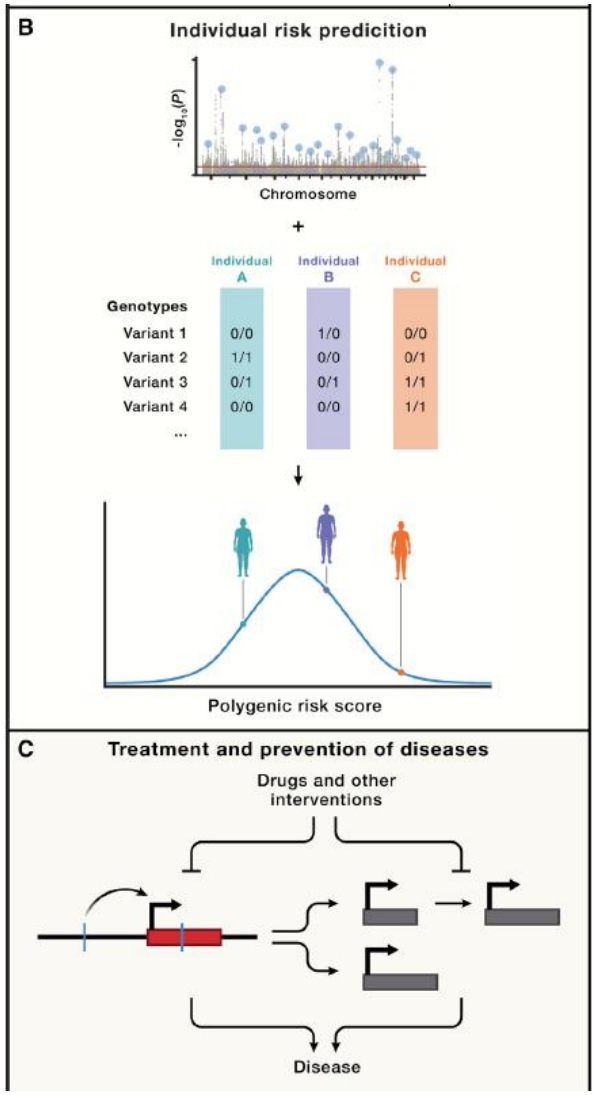
Datasets for human genetics research



Lappalainen et al, 2024

- Growth in human genetics dataset. Relationship between comprehensiveness of the genome analysis (x axis) with the technologies indicated on the top, and the number of donors (y axis). Underlined project names include a relatively balanced representation of individuals from diverse ancestries.
- The projects: Human Genome Project (HGP), HapMap, Wellcome Trust Case Control Consortium (WTCCC), 1000 Genomes (1KG), UK Biobank (UKBB), Pangenome project, Genotype Tissue Expression (GTEx), and Trans-Omics for Precision Medicine (TOPMed), WGS, whole-genome sequencing.

Motivation for human genetics research

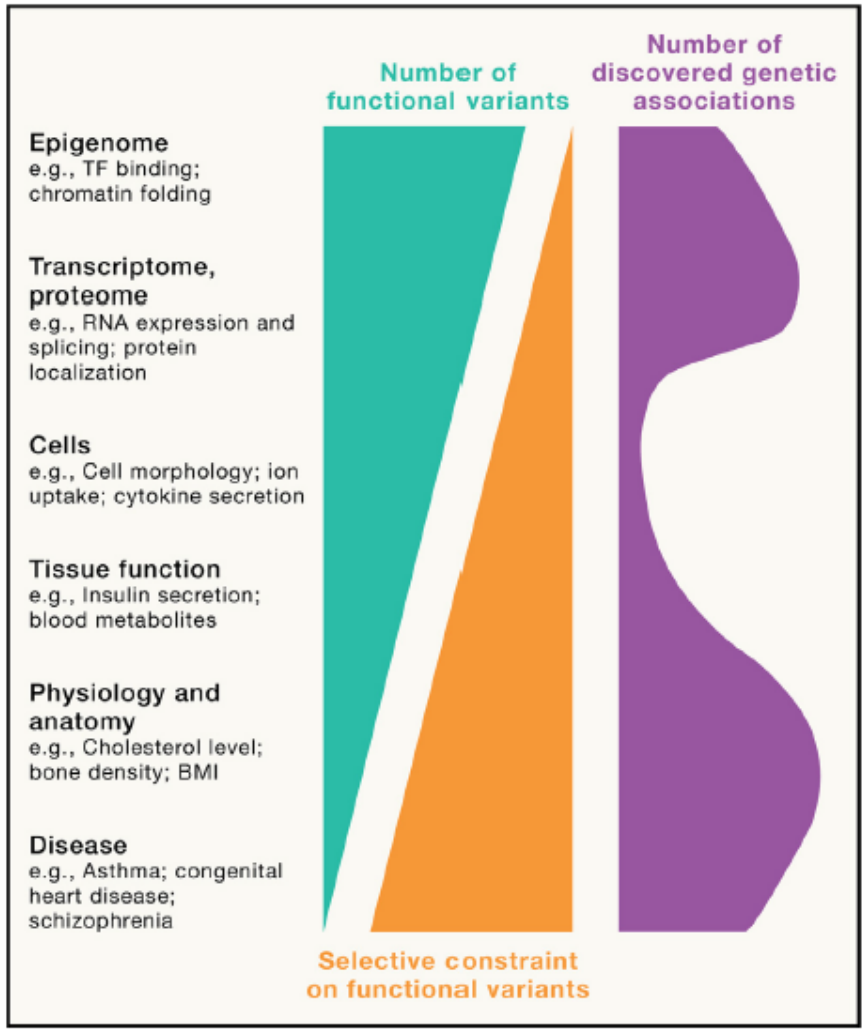


Lappalainen et al, 2024

Illustration of the two complementary approaches how human genetics contributes to human health.

- Well-powered GWAS can allow building polygenic risk scores that can be used for personalized disease risk prediction.
- Understanding the functional mechanisms of GWAS loci can allow targeting these mechanisms with drugs and other interventions to prevent or treat disease.

Genetic effects on functions at different levels



- There are large numbers of variants affecting molecular functions of the genome and the cell, many of which have **no or smaller effects** downstream.
- Variants affecting physiological, anatomical, and disease traits can be under direct **natural selection.** The purple graph indicates the success in discovery of genetic associations for molecular traits (captured by molQTL mapping) and for physiological and disease traits (captured by classical GWAS), with a gap in our knowledge of genetic associations for **cellular** and **tissue-level** traits.

Lappalainen et al, 2024

Extreme polygenicity of common traits

A key parameter driving genetic discoveries is the trait polygenicity:

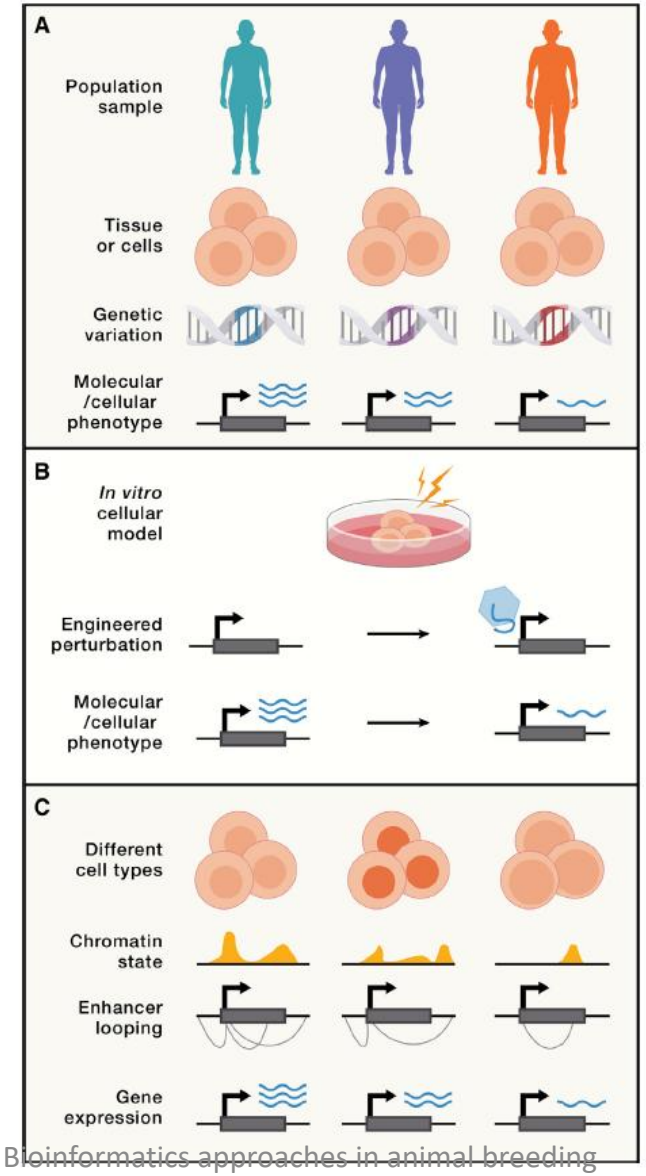
- the total number of causal variants influencing the trait and
- the distribution of their effect sizes.

❖ Highly polygenic traits involve many weak causal variants and require large sample sizes. Polygenicity has been consistently estimated to be very high, ranging from thousands of causal variants to millions. This imply that, for some traits, many causal variants are acting through nearly every gene in the genome on average and implicate more than half of all common polymorphisms.

❖ In general, cellular and pigmentation traits exhibit the lowest polygenicity (hundreds of causal variants), whereas anthropometric and cognitive/behavioral traits exhibit some of the highest estimates (>10,000 effective variants). Although traits with similar heritabilities often exhibited different levels of polygenicity, the variance in polygenicity across traits was generally lower than expected, suggesting that selection, may be acting pleiotropically across traits rather than on any one measured phenotype.

❖ Recently, a GWAS of height in 5.4 million participants demonstrated that 12,111 jointly significant variants explained 40% of the phenotypic variation, lending the first direct evidence for high trait polygenicity. The evolutionary causes of high polygenicity continue to be actively investigated, but the implications are clear: understanding human traits will require distilling the function of tens of thousands of variants.

Approaches for understanding molecular effects of genetic variants



Lappalainen et al, 2024

- (A) molQTL mapping.
- (B) Engineered perturbations of the genome.
- (C) Inference from multi-layered functional omics data.

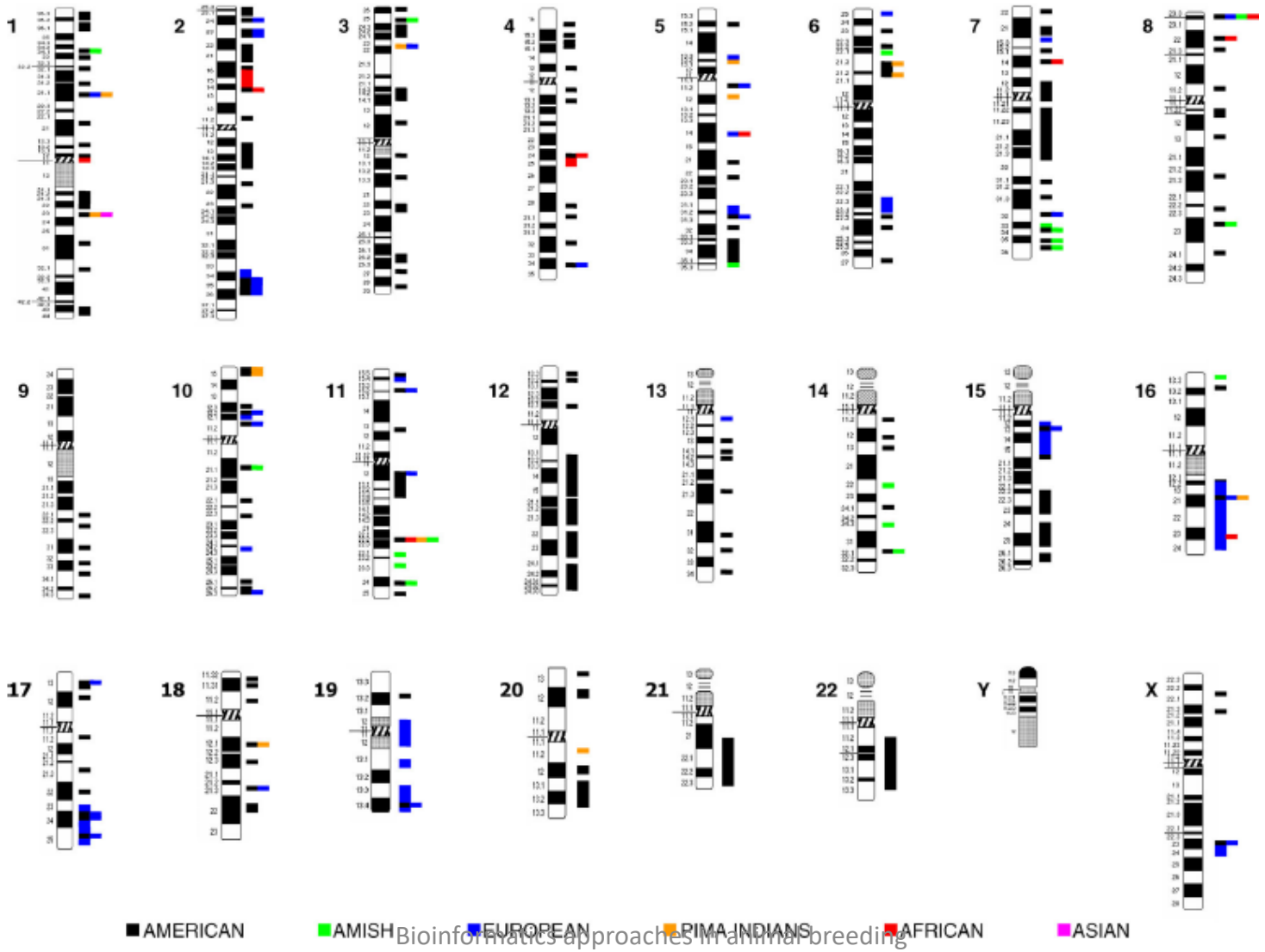
The Human Obesity Gene Map

Evidence from the rodent and human obesity cases caused by single-gene mutations, Mendelian disorders exhibiting obesity as a clinical feature, quantitative trait loci uncovered in human genome-wide scans and in cross-breeding experiments in various animal models, and association and linkage studies with candidate genes and other markers are reviewed.

Forty-seven human cases of obesity caused by single-gene mutations in six different genes have been reported in the literature. Twenty-four Mendelian disorders exhibiting obesity as one of their clinical manifestations have now been mapped. The number of different quantitative trait loci reported from animal models currently reaches 115. Attempts to relate DNA sequence variation in specific genes to obesity phenotypes continue to grow, with 130 studies reporting positive associations with 48 candidate genes.

Finally, 59 loci have been linked to obesity indicators in genomic scans and other linkage study designs. The obesity gene map reveals that putative loci affecting obesity-related phenotypes can be found on all chromosomes except chromosome Y. The number of loci in human genome associated with human obesity phenotypes is now above 250.

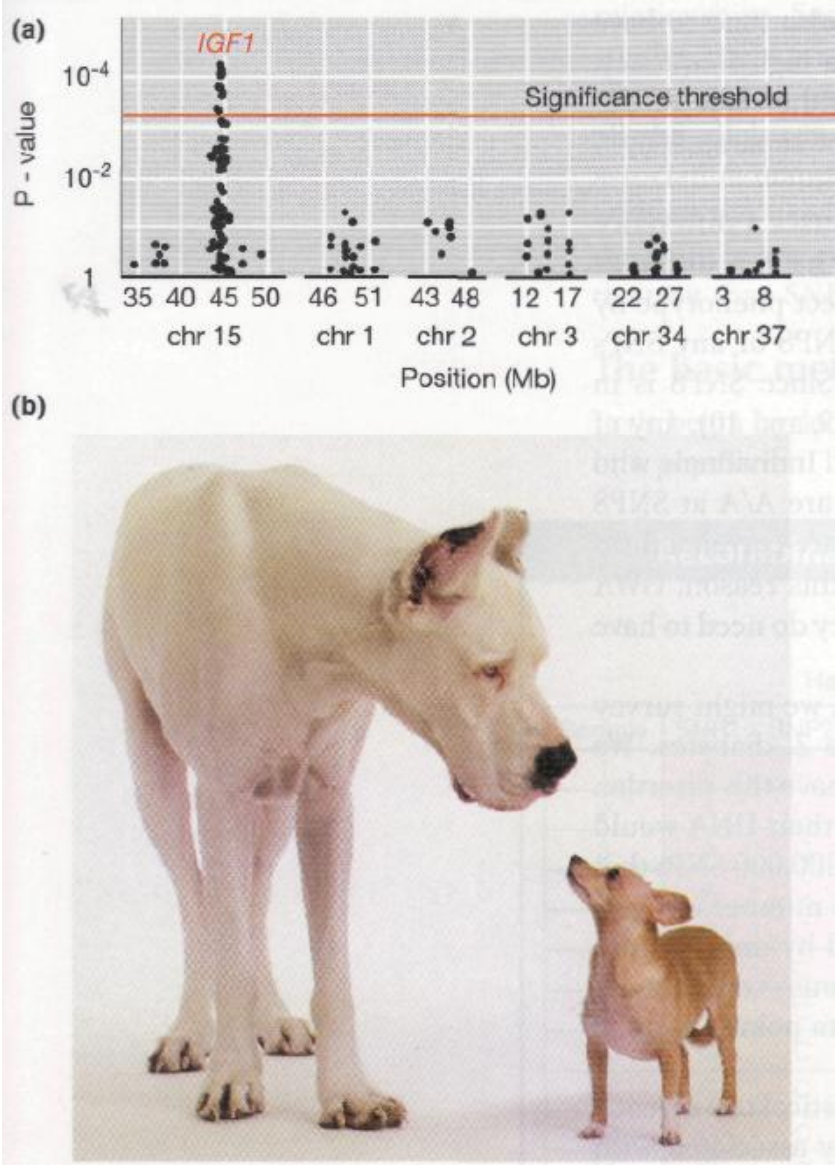
Mapping of obesity QTL in human populations

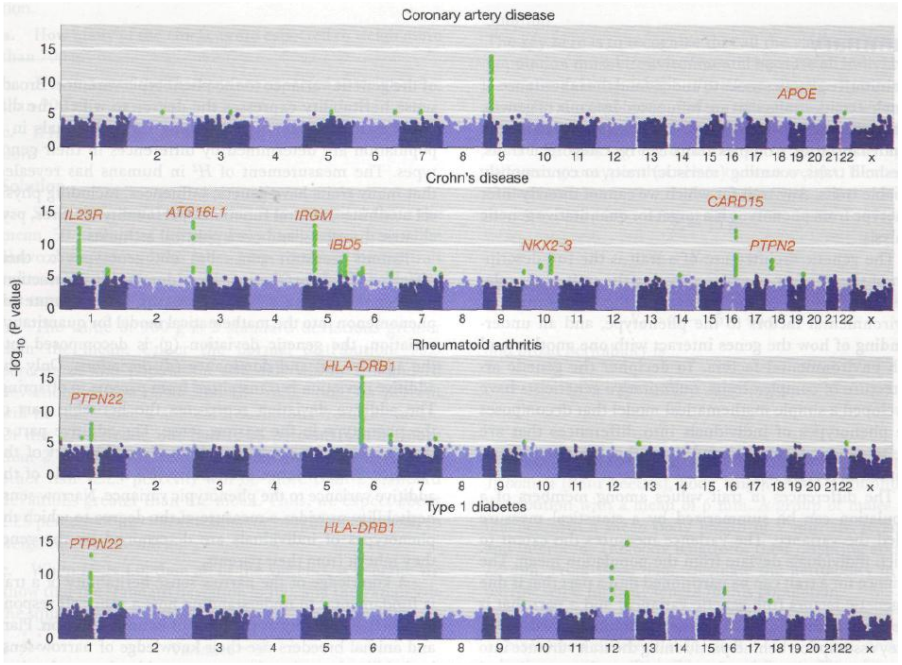


Genetic architecture of complex traits

- Traditional SNP regression analysis of GWAS reports the estimated effect of SNPs that are declared **significant**. However, this information does not give a good description of the genetic architecture of the trait.
- In SNP regression analysis, **very stringent p-values** ($p < 5 \times 10^{-8}$) are used to protect against testing as many as 1 000 000 SNPs for an effect. This has several consequences. The estimated effect of SNPs declared significant is grossly **overestimated**.
- A better estimate of the effect size can be obtained by estimating the effect of significant SNPs in an independent dataset. In this case, the **significant SNPs explain a small proportion of the genetic variance estimated from pedigree or family relationships**. This was called the 'missing heritability' paradox.
- For instance, the genetic variance for height in humans explained by significant SNPs was 5% of phenotypic variance, whereas all the SNPs together explained 45% of the phenotypic variance. The explanation for this difference is simply that **most SNP effects on height are too small** to be significant given the stringent p-value used.

GWAS, geni, bolezni, heritabiliteta

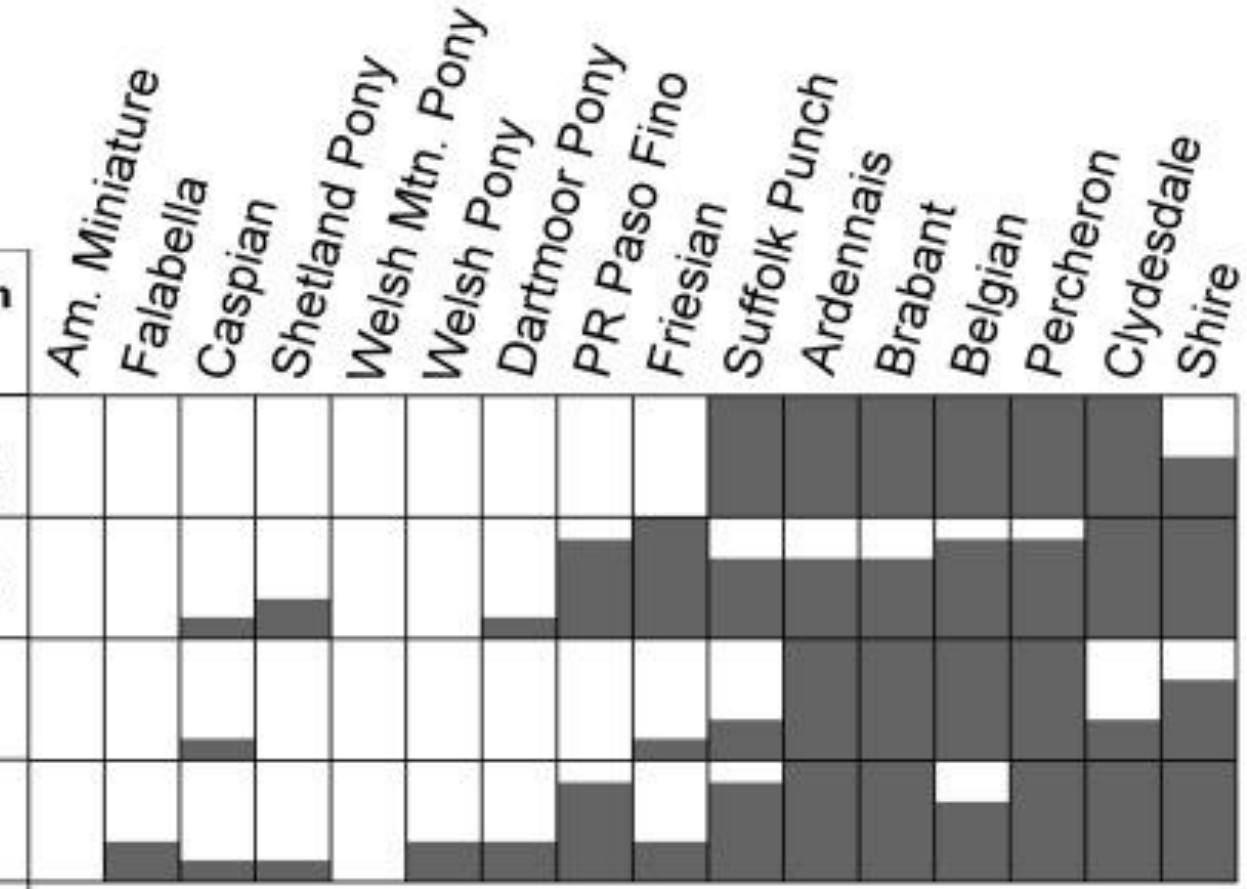




GWAS nam omogočajo mapiranje regij, kjer se nahajajo vzročni geni za pomembne bolezni.

Genetics of body size in horse

Chr.	Position	Alleles		P-value	Fraction of Variance Explained	Genes in Locus
		Little	Big			
3	105,547,002	T	C	1.05×10^{-9}	0.685	*LCORL, NCAPG*
6	81,481,064	C	T	5.21×10^{-7}	0.556	*HMGA2*
9	75,550,059	C	T	3.48×10^{-7}	0.567	*ZFAT*
11	23,259,732	G	A	6.19×10^{-7}	0.589	*LASP1*
4 Loci Together					0.835	



Thanks for your attention!