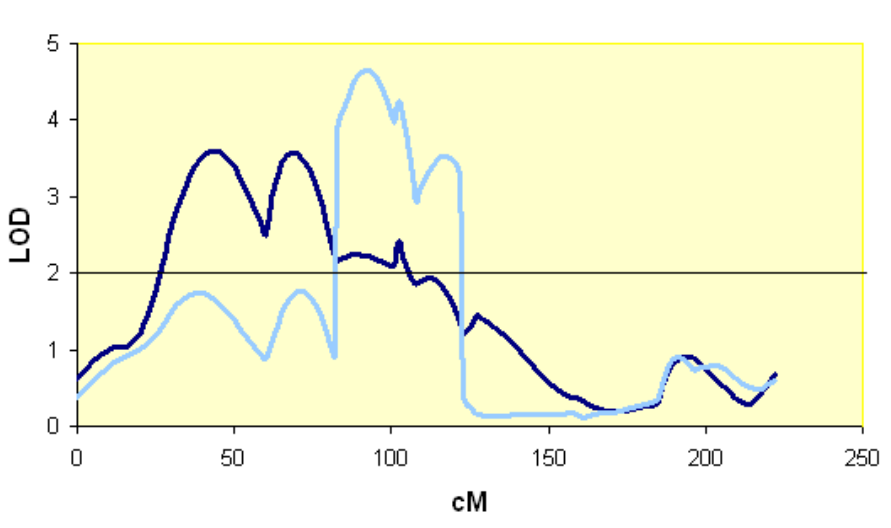
Quantitative trait loci, genome wide association mapping en het vinden van genen.



Plant Ecology and Phytochemistry
Institute of Biology Leiden (IBL)
Leiden University

Natural variation presents one of the fundamental challenges of modern biology. Soon, the genome sequences of thousands of individuals will be known for each of several species. But how does the genotypic variation that will be observed among these individuals translate into phenotypic variation?

Nordborg and Weigel, Nature 2008

Phenotypic variation is mostly quantitative variation

Next generation sequencing

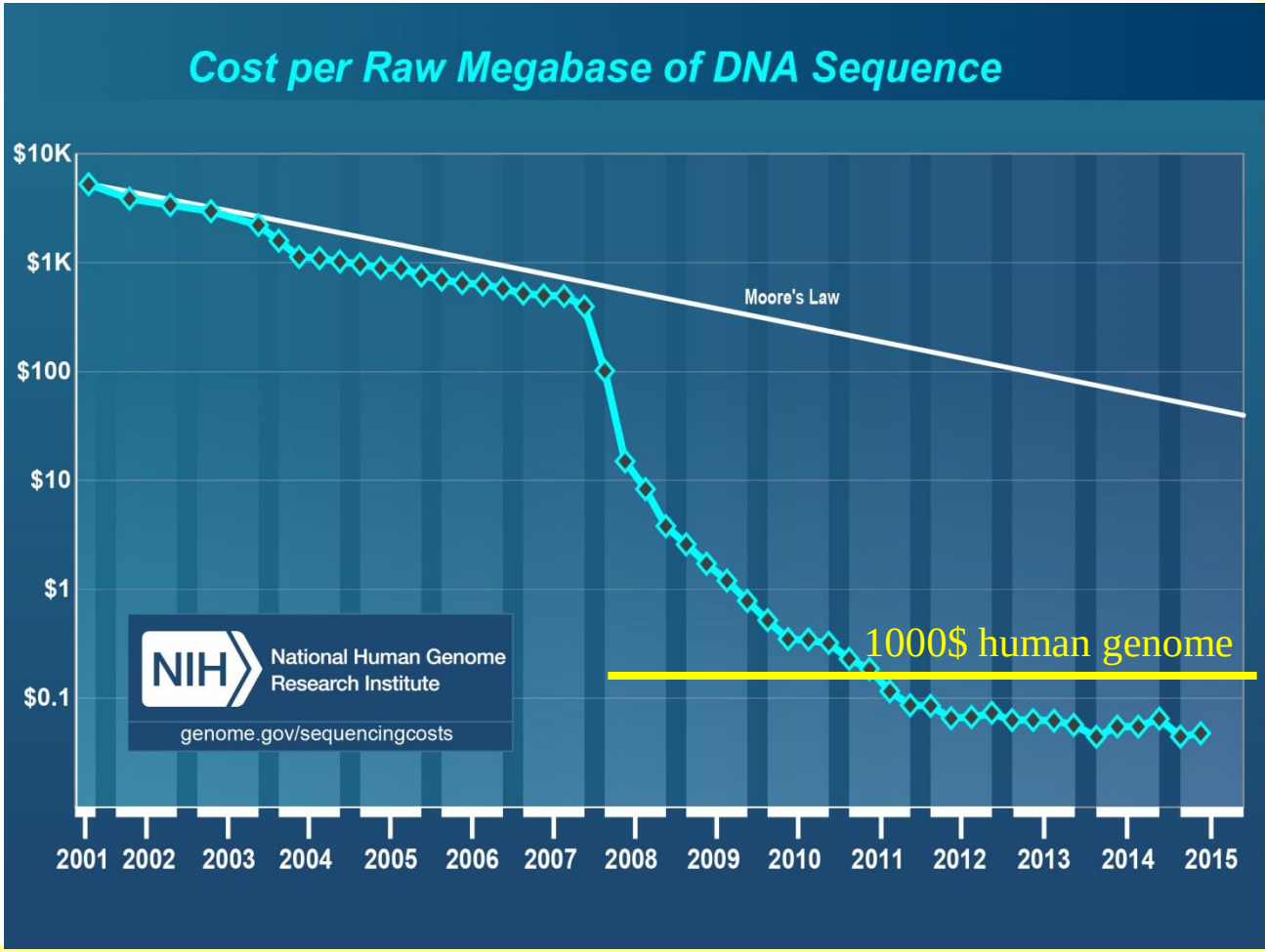
For more than 25 years sequencing traditionally has been carried out with the Sanger method.

Since 2005 (454) and 2007 (Illumina) next generation sequencing machines were launched.

What are the new solutions?	Read length(bp)	Output 2017
454 GS FLX (Roche)	600	
Illumina	**2 x 150**	**1500 Gb**
Ion Torrent	400	
Pac Bio	8000 - 30000	
Minion/Promethion	**8000 – 40000**	**6400 Gb**

Next generation sequencing

Figure 4.5: The rapidly declining cost of DNA sequencing.
The line labeled "Moore's Law" describes the long-term trend whereby the cost of computing halves every two years. Taken from **www.genome.gov/sequencingcosts/**.



The Modern Synthesis

Phenotypic variation is mostly quantitative variation

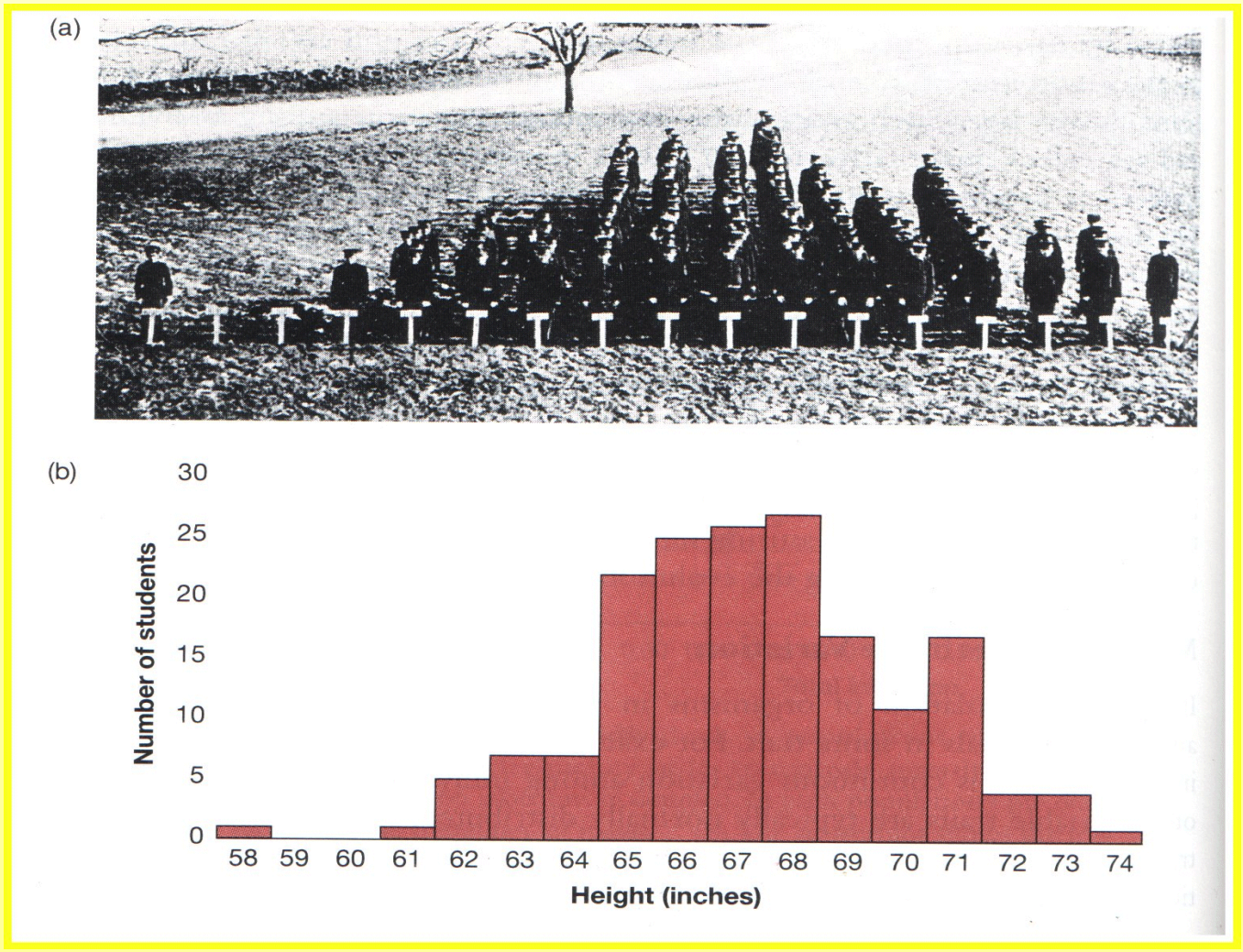
The Modern Synthesis

The incorporation of Mendelian Genetics into Evolutionary Theory and Quantitative Genetics.

What is the underlying genetics of quantitative traits, like weight of an animal, date of first flowering, etc

Quantitative traits

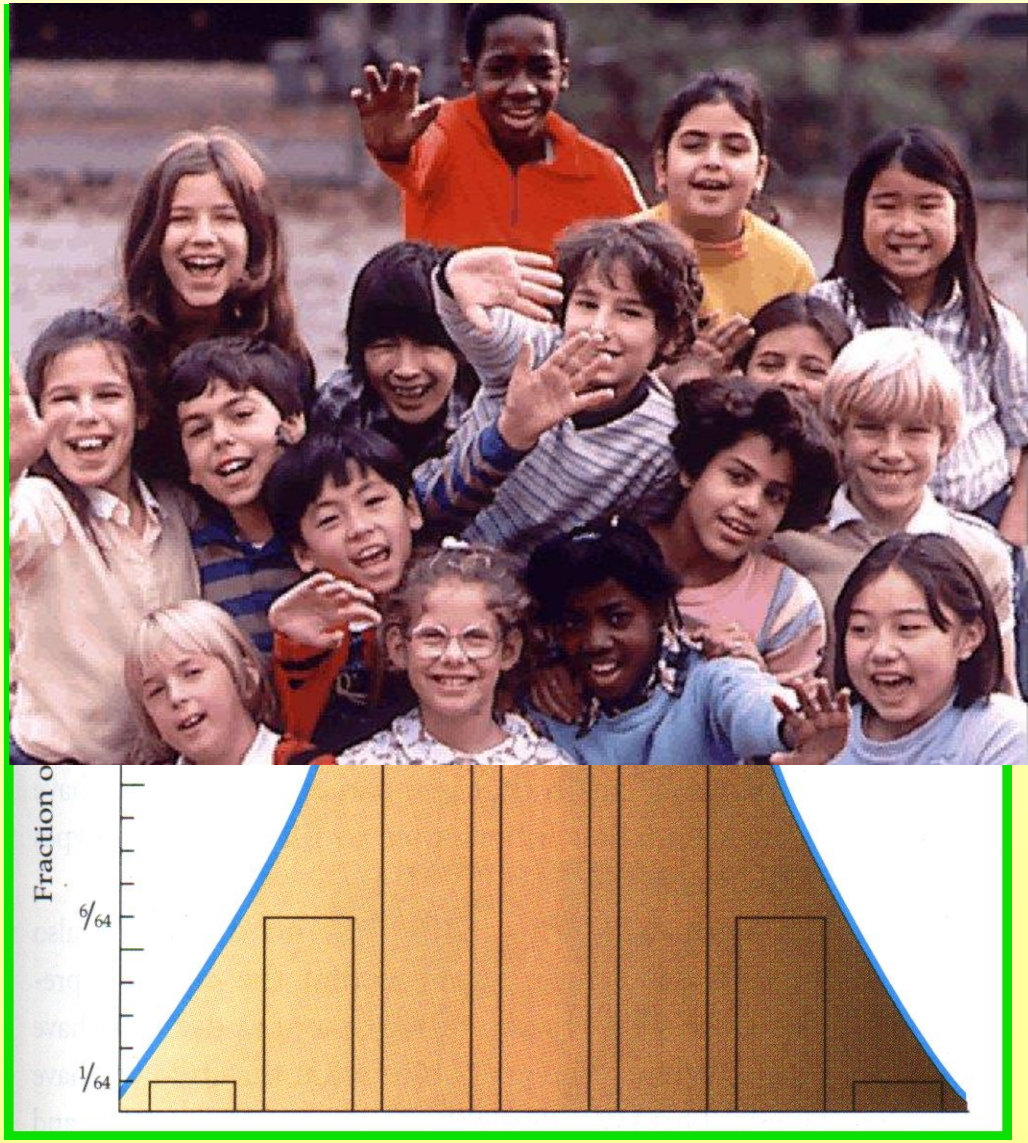
Characters or traits that are of degree not of kind



Quantitative traits

Mostly, but not necessarily, determined by variation at many loci (1 - 20)

Polygenic inheritance



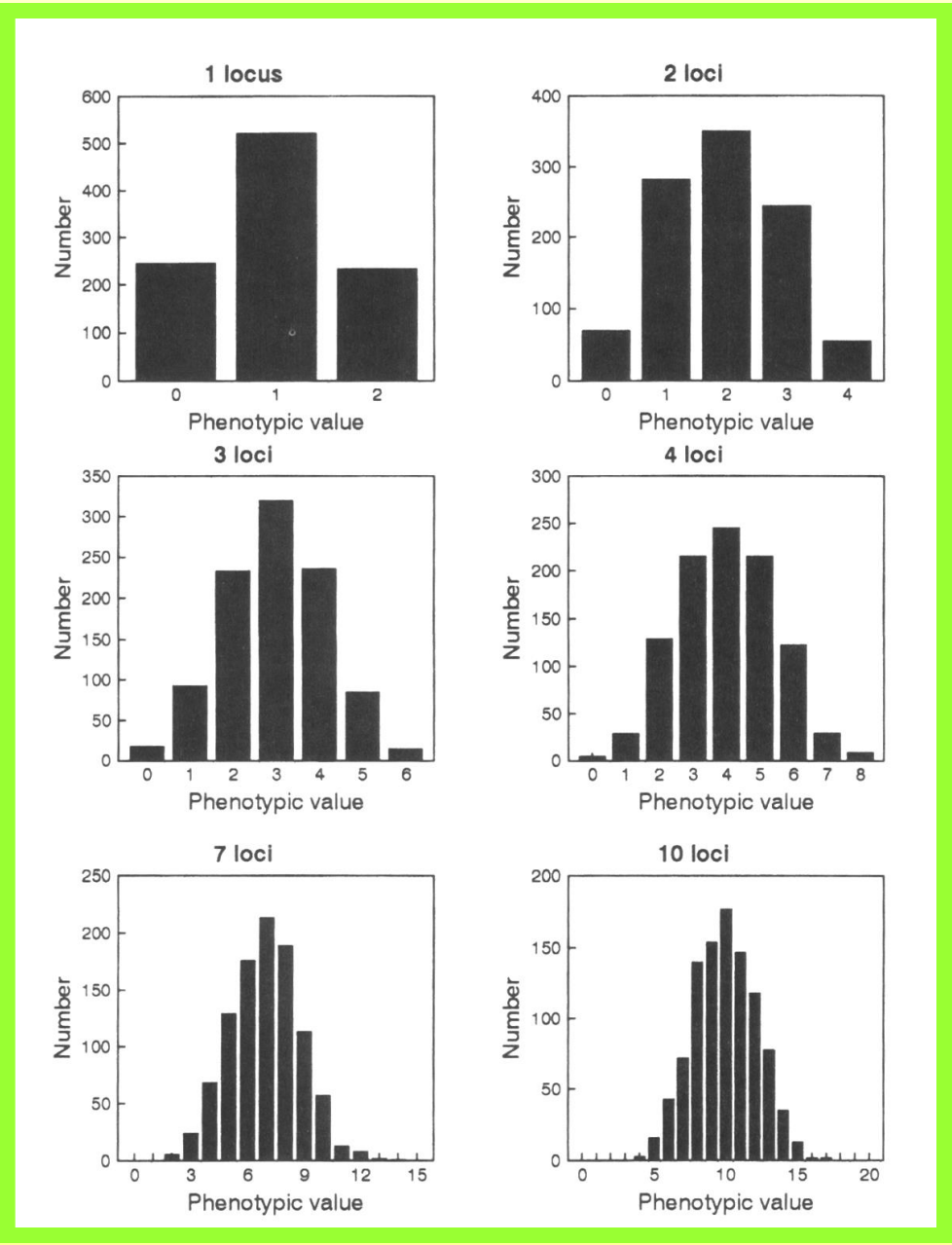
Quantitative traits

Individual genes can not be identified by their segregation

Hence, Mendelian analysis does not apply

Quantitative Genetic traits are sensitive to external and internal pertubations

$$P = G + E$$



QTL mapping

With Quantitative Trait Locus (QTL) mapping we try to find the genes that underlie Quantitative Genetic variation and determine their properties

Find underlying loci

Determine number of loci

Additive and dominance effects

Epistatic and pleiotropic effects

Study constraints

Marker assisted breeding

QTL mapping

Based on linkage disequilibrium (LD) between alleles at a marker locus at which genotypes can be distinguished unambiguously, and alleles at the linked QTL

Took off in early 1980 because of development molecular markers

QTL mapping

Requirements

Genetic variation for the quantitative trait within or between populations or strains

Linkage map of polymorphic (molecular) marker loci that covers the whole genome

Marker loci should be: eg SNPs

- (highly) polymorfic
- abundant
- preferably co-dominant

Genetic mapping

A genetic map is a network of connected genetic markers, (ideally) covering the whole genome of an organism.

How is a genetic map made?

1) Genotype all the offspring of a specified cross for a large number of genetic markers

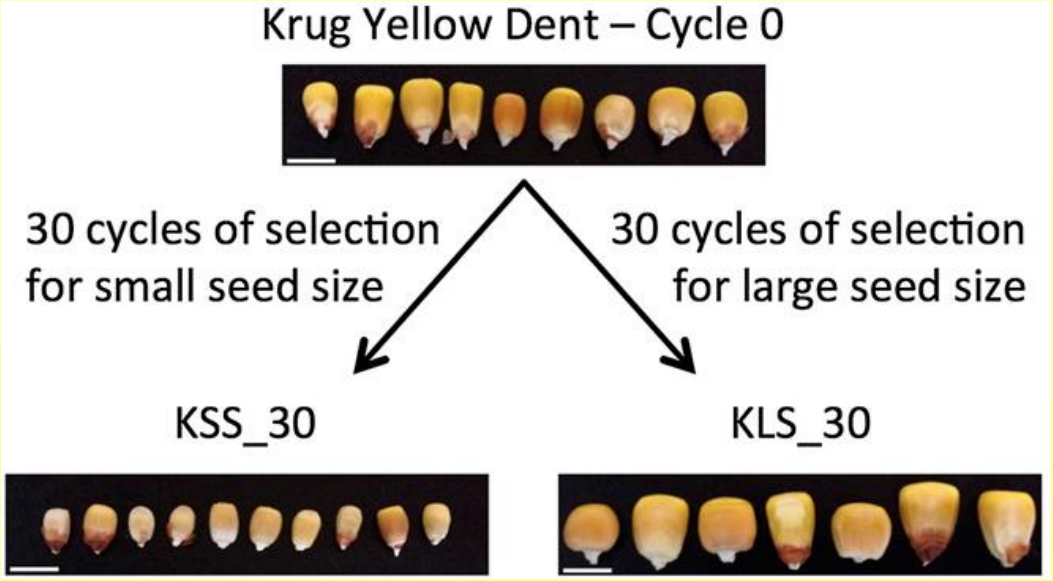
2) Calculate the linkage between all genetic marker pairs

3) Group the markers according to linkage and number of chromosomes

Genetic mapping

Use crosses are that are homozygous at QTL and marker loci

All alleles that *decrease* trait in one parental line and all alleles that *increase* trait in the other parental line



Hirsch et al. 2014

Divergent artificial selection on seed size in *Zea mays*.

Genetic mapping

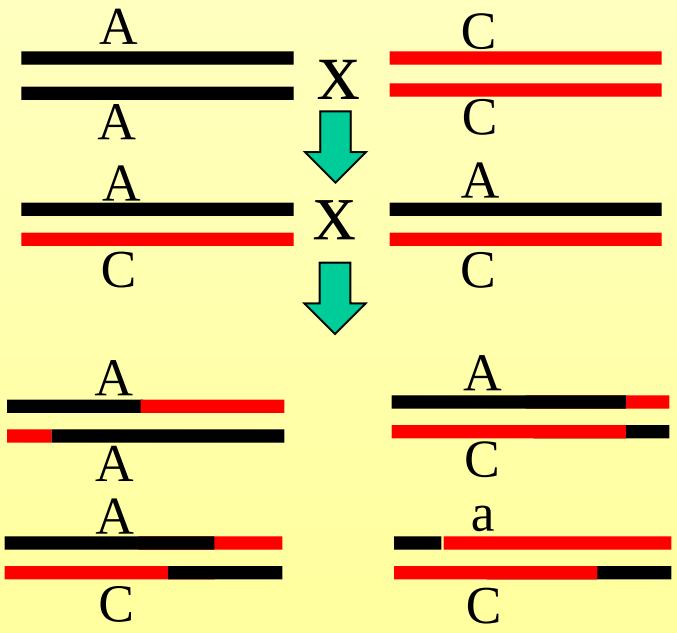
Types of crosses for making genetic maps

Several crossing schemes are possible. An F2 cross is often used. Purpose is to get a segregating populations for genetic markers and the QTL

P_A x P_B

F_1 x F_1

F_2



AC x AC ==> **F_2** : 25% AA, 50% AC and 25% CC

Genetic mapping

Estimating genetic linkage in a backcross

A cross ($\frac{A\ C}{G\ T}$ x $\frac{A\ C}{A\ C}$) yielded the following outcome:

	Observed		Expected (no linkage)
$\frac{A\ C}{A\ C}$	97	(Non-recombinants: Nr)	60
$\frac{A\ C}{G\ T}$	88		60
$\frac{A\ T}{A\ C}$	23	(Recombinants: R)	60
$\frac{G\ C}{A\ C}$	32		60

Chi-square: $(nNr- nR)^2/n = (97+88-23-32)^2/240 = 70.4$ ($p< 0.001$)

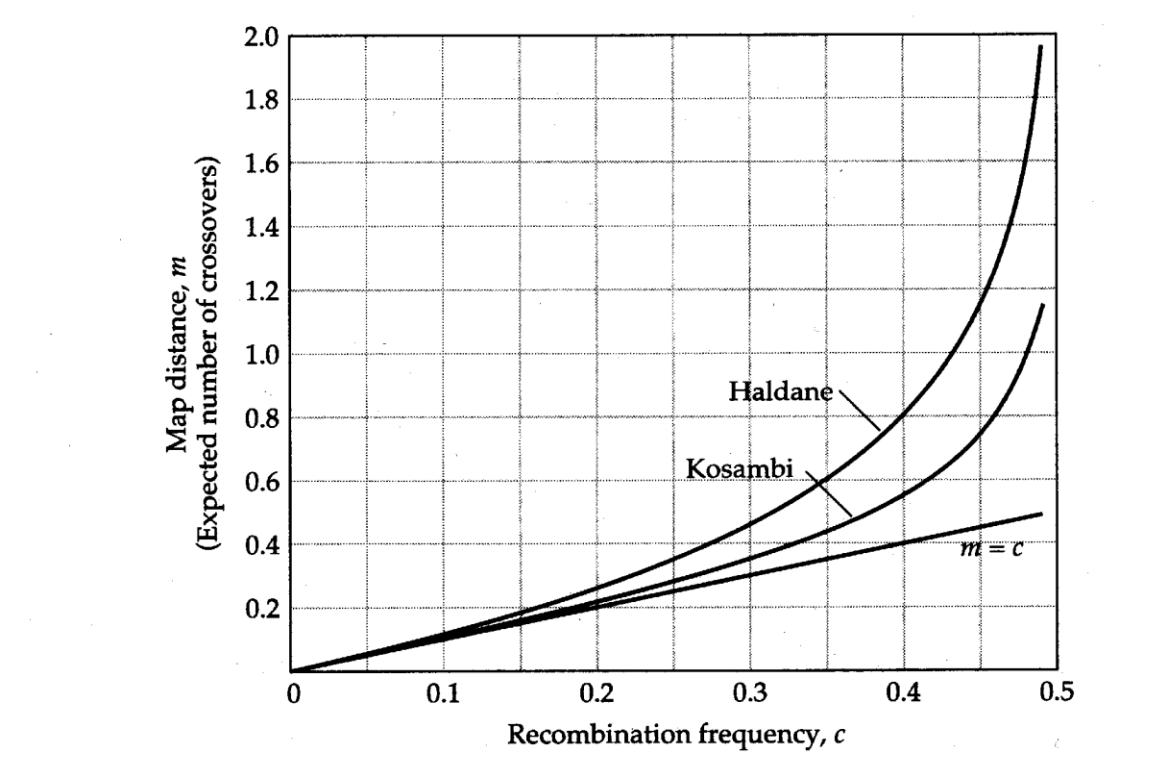
$r = (23+32)/240 = 0.229$

Genetic mapping

Map function

Kosambi (1944) and Haldane's mapping (including interference)

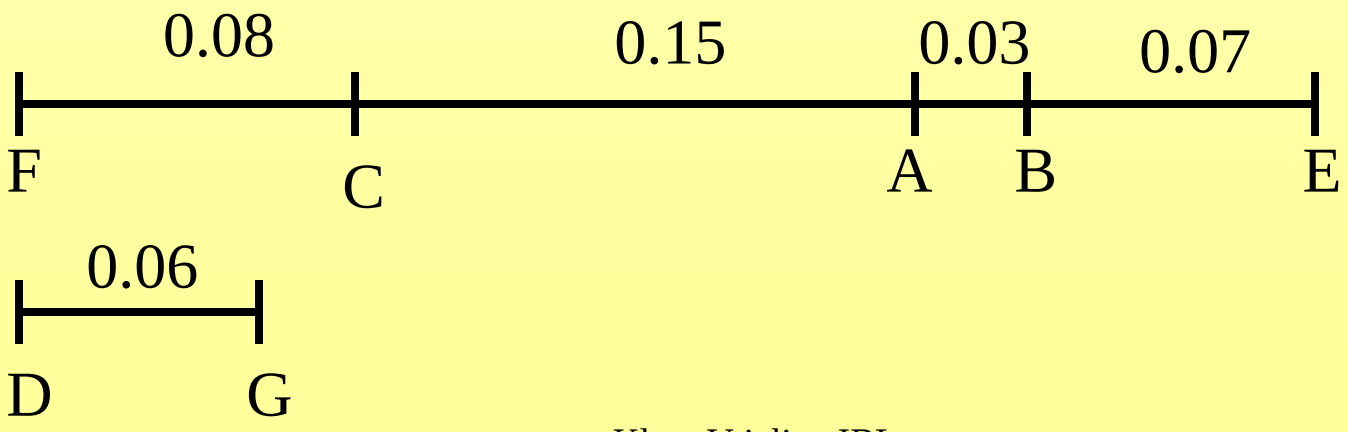
x (in cM) = 0.25 ln[(1+2r)/(1-2r)]



Genetic mapping

Calculate linkage and cM for all possible combinations of segregating molecular markers eg:

A-B = 0.03	B-C = 0.18	C-E = 0.25	E-F = NL
A-C = 0.15	B-D = NL	C-F = 0.08	E-G = NL
A-D = NL	B-E = 0.07	C-G = NL	G-F = NL
A-E = 0.10	B-F = 0.26	D-E = NL	
A-F = 0.23	B-G = NL	D-F = NL	
A-G= NL	C-D = NL	D-G = 0.06	



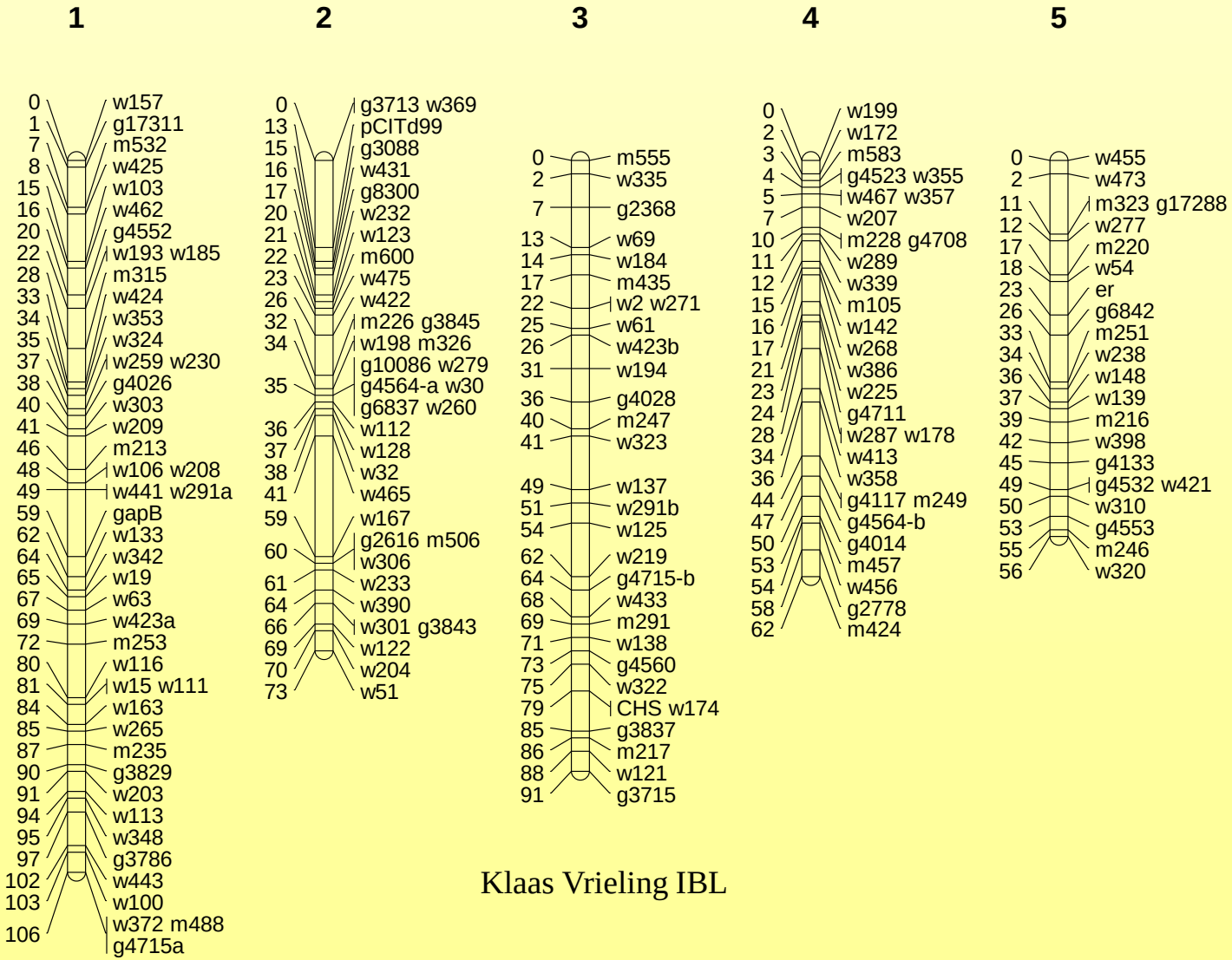
Genetic mapping

Calculate r (or cM) for all possible combinations of segregating molecular markers.

After grouping as many linkage groups are obtained as chromosomes!

Genetic mapping

Arabidopsis RIL's (F_8) 101 individuals 178 loci



Klaas Vrieling IBL

QTL mapping

Analysis: - individuals are scored for their genotype at the marker locus

- individuals are scored for their phenotype for the trait

- difference in mean phenotype among marker genotype classes indicates that a QTL is linked to the marker

QTL mapping

SNP marker 22

	Genotype	Traitvalue
53 individuals	AA	55,4
118 individuals	AC	53,7
` 64 indivduals	CC	56,1

SNP marker 211

	Genotype	Traitvalue
48 individuals	CC	29,4
132 individuals	GC	25,7
` 55 indivduals	GG	110,1

QTL mapping

Centi-Morgans do not reflect a physical distance in basepairs.

	Kb/cM
Yeast	6
Neurospora	80
Arabidopsis	140
Drosophila	700
Tomato	510
Human	1110
Corn	2140

QTL mapping

One QTL still consists of a region covering 5000.000 bp up to 100.000.000 bp.

One QTL can contain multiple genes of the trait under study (1 QTL does not necessarily indicate 1 gene)

So by sequencing it is still an effort to find the underlying genes

QTL mapping *Barbarea*

Barbarea vulgaris (Watercress, Barbarakruid)

Plutella xylostella

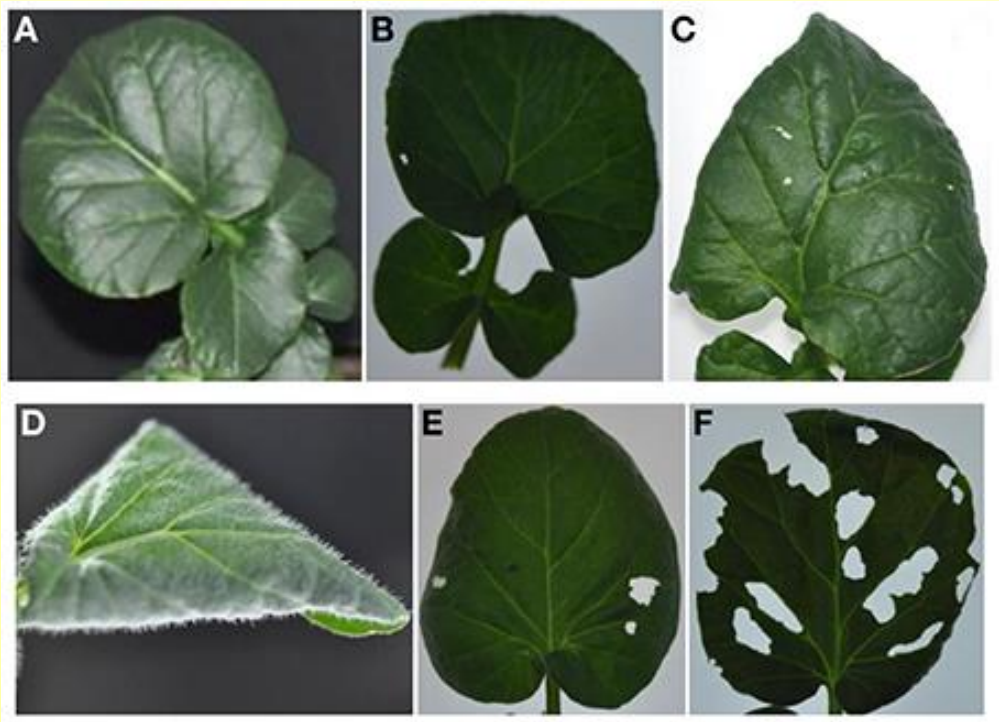
Phyllotreta nemorum

Is a biennial plant that occurs on open, sunny, moist, rich soils.

It is attacked by *Plutella xylostella* (koolmot) and *Phyllotreta nemorum* (grote gestreepte aardvlo)

Contains saponins and glucosinolates as anti-herbivored defence

QTL mapping *Barbarea*



G-type

Glabrous (smooth)

Resistant to flea beetle

Glucobarbarin (glucosinolate)

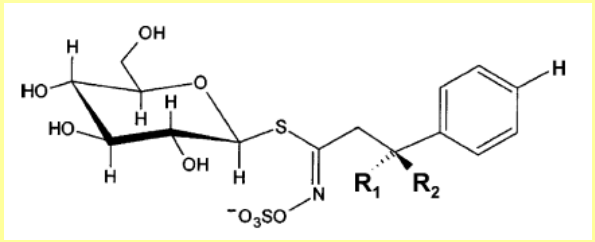
Hederagenin cellobioside (saponin)

P – type

Pubescent (fine haired)

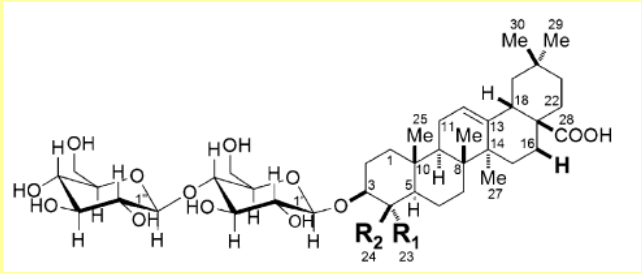
Susceptible to flea beetle

Contains glucosibarin (glucosinolate)



Glucosibarin

Phyllotreta nemorum



Hederagenin cellobioside

QTL mapping *Barbarea*

Do hairs play a role in the resistance against the herbivores?

Do glucosinolates play a role in the resistance against herbivores?

Do saponins play a role in the resistance against herbivores?

How many QTLs can we find for each trait?

QTL mapping *Barbarea*

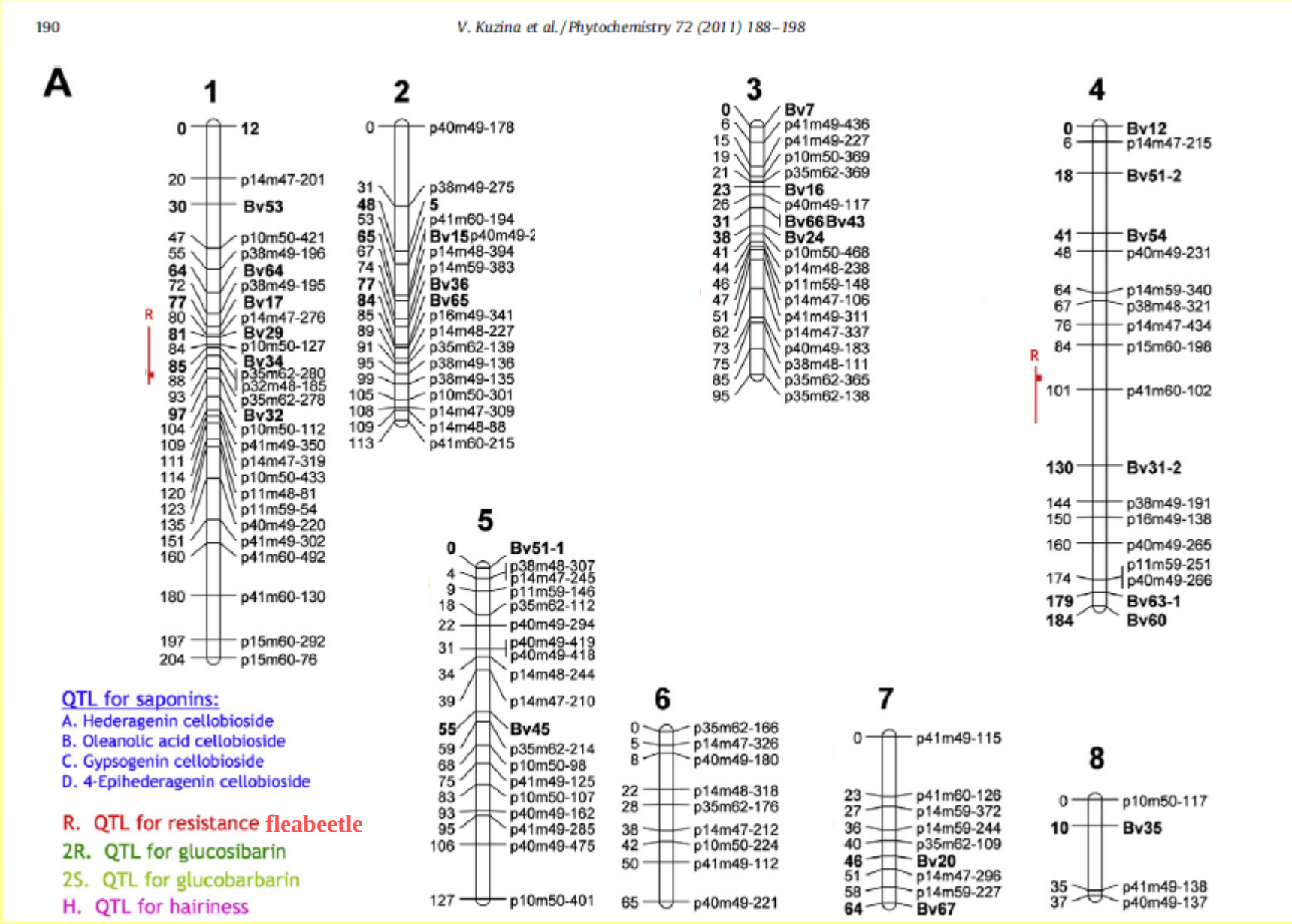
A cross was made between a P and G type parent of *B. vulgaris*, two F1s were crossed and 129 F2 genotypes were raised in a greenhouse.

From these 129 F2 genotypes it was determined whether they were resistant against the flea beetle.

Glucosinolate and saponin concentrations were measured as well as the hairiness of the plants

All 129 plants were genotyped with 99 AFLP and 27 SSR markers.

QTL mapping *Barbarea*



Kuzina et al. 2012

QTL mapping *Barbarea*

Two QTL s are found and the explained variance excludes that there are more QTL of larger effect. Assuming a two gene model with the P type being resistant (AABB) and the G type being susceptible (aabb) to the fleabeetle

Genotype	Larvae survival
AABB	0.04 ± 0.02
AAbb	0.75 ± 0.06
AABb	0.26 ± 0.04
aaBB	0.26 ± 0.08
aabb	0.96 ± 0.02
aaBb	0.74 ± 0.04
AaBB	0.11 ± 0.04
Aabb	0.89 ± 0.03
AaBb	0.50 ± 0.05

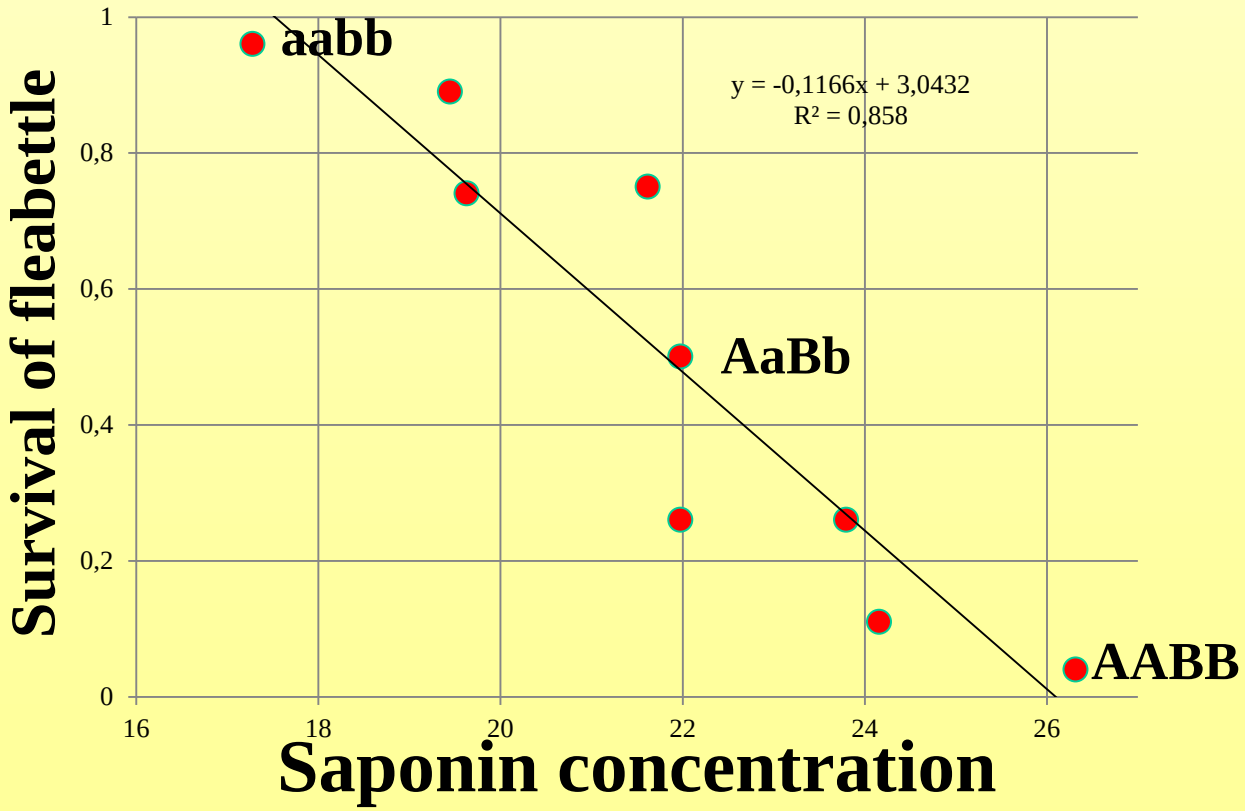
Effect of each A allele is a 0,15 loss in survival

Effect of each B allele is a 0.36 loss in survival

Kuzina et al. 2012

QTL mapping *Barbarea*

Concentration of saponins is negatively correlated with flee beetle survival



Kuzina et al. 2012

QTL mapping *Barbarea*

So what are the underlying genes?

The genome of *Barbarea* is not sequenced so it was compared to the genome of *Arabidopsis thaliana* assuming synteny.
But *A. thaliana* does not contain the particular saponins.

From other research they Kuzina et al. came to the conclusion that genes from oxidosqualene cyclases and P450's families and perhaps transcription factors are the causal factors.

QTL mapping *Barbarea*

Conclusions

Two loci are found that confer resistance to the flee beetle

These two loci are coupled with loci coding for saponin concentration

Hairiness and Glucobarbarin are not involved in resistance against the flea beetle

Markers could be used for selecting for resistance against the flea beetle

Resistance could not be coupled to underlying genes

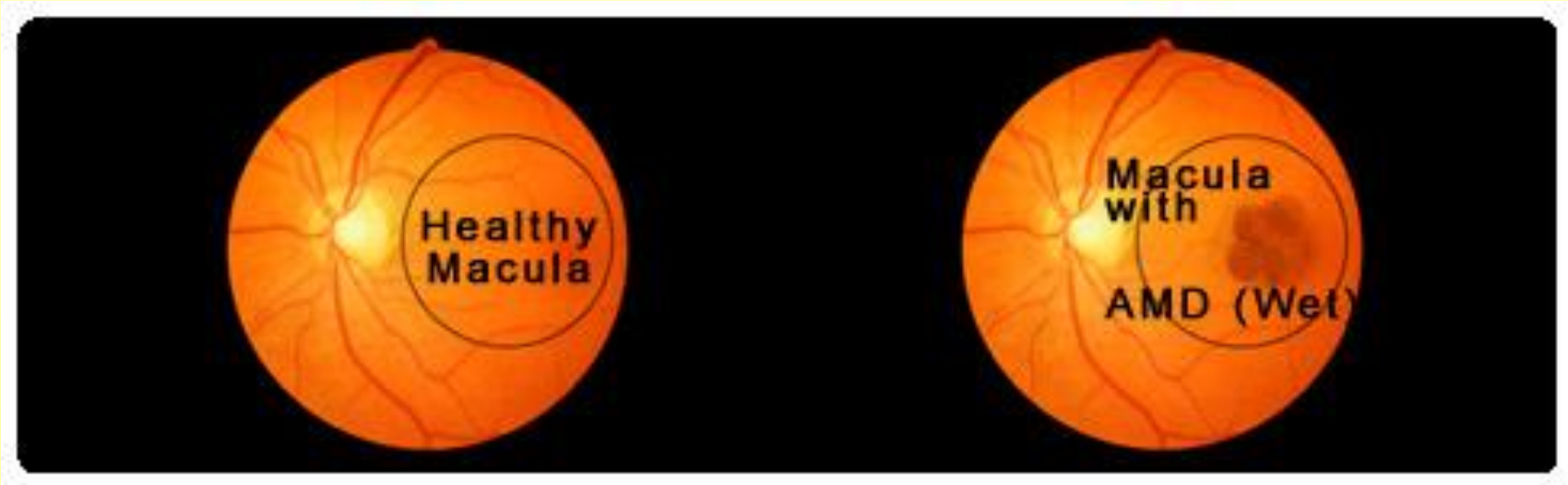
GWAS (Genome wide association mapping)

Coupling of molecular variants to (quantitative) traits, like weight of an animal, date of first flowering, simultaneously in many unrelated individuals.

First used in humans to detect underlying loci of diseases

GWAS AMD

Age-related macular degeneration (AMD) is the leading cause of blindness in the developed world.

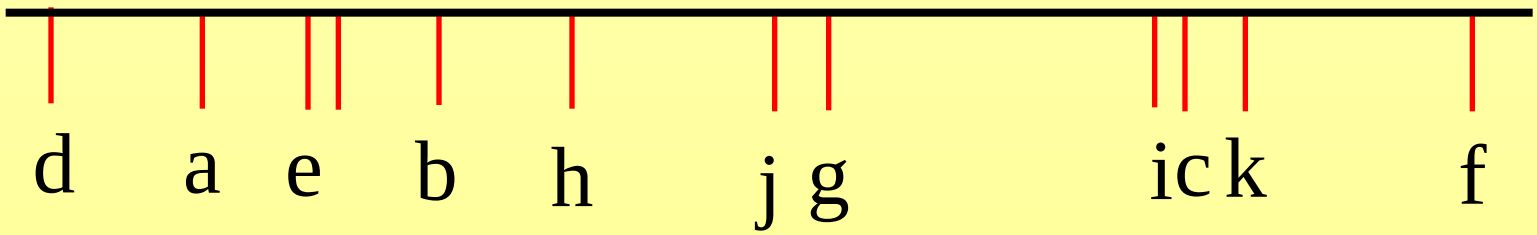


AMD is characterized by progressive destruction of the retina's central region (macula), causing central field visual loss

GWAS AMD

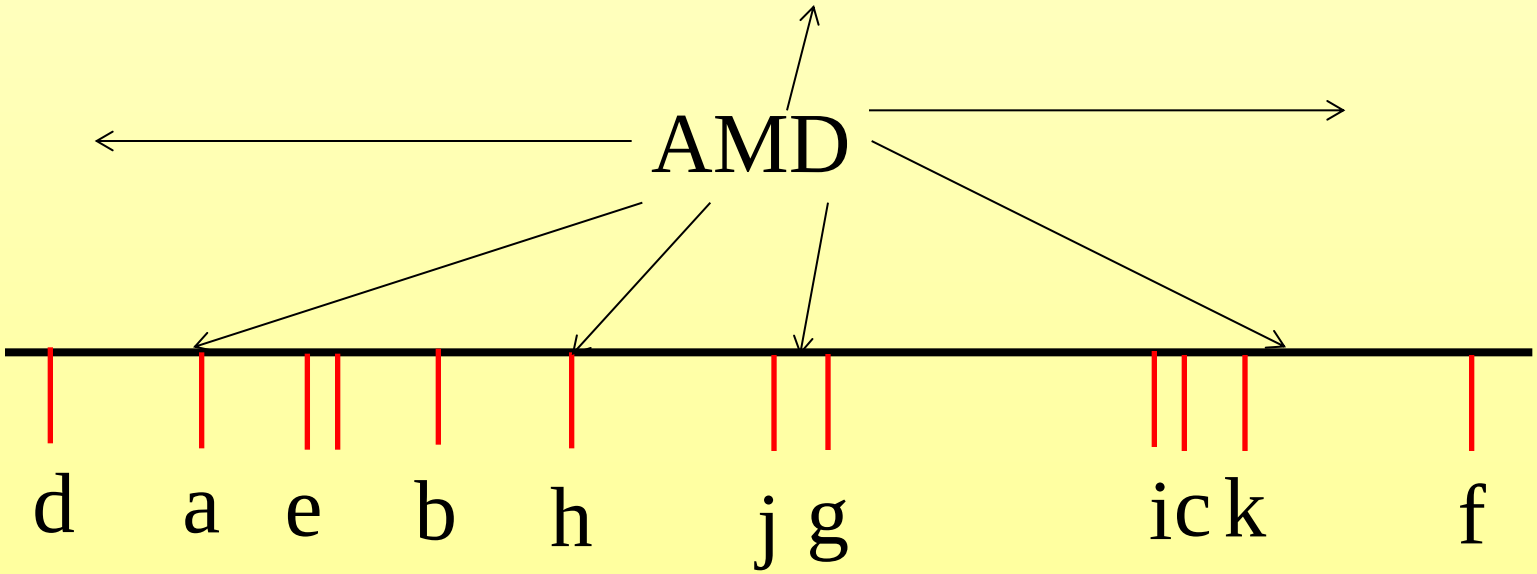
Klein et al. 2005 genotyped 96 individuals with AMD and 50 control individuals without AMD for 116,204 SNPs covering the whole human genome.

For all SNP loci in humans we know the position on the human genome.



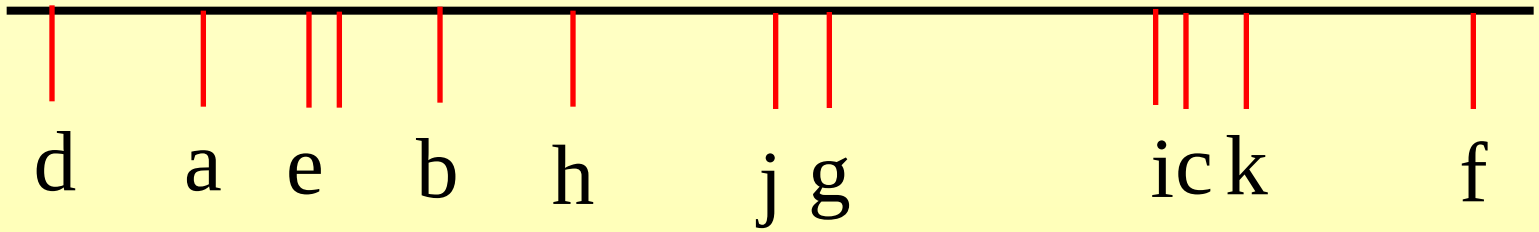
GWAS AMD

Where on the genome are the gene(s) causing AMD located??



GWAS **AMD**

So test the SNPs one by one:



		AMD			
		With		Without	
SNP d	A	45	45,3	24	23,6
	G	51	50,6	26	26,4

GWAS AMD

		AMD			
		With		Without	
SNP k45	C	2	45,3	48	23,6
	G	95	50,6	1	26,4

So SNP K45 is associated with AMD

Notice that the example is simplified because we deal with genotypes CC, CG and GG (and not only with allele C and G)

GWAS AMD

96 individuals with AMD and 50 control subjects without AMD were genotyped for 116,204 SNPs.

Manhattan plot

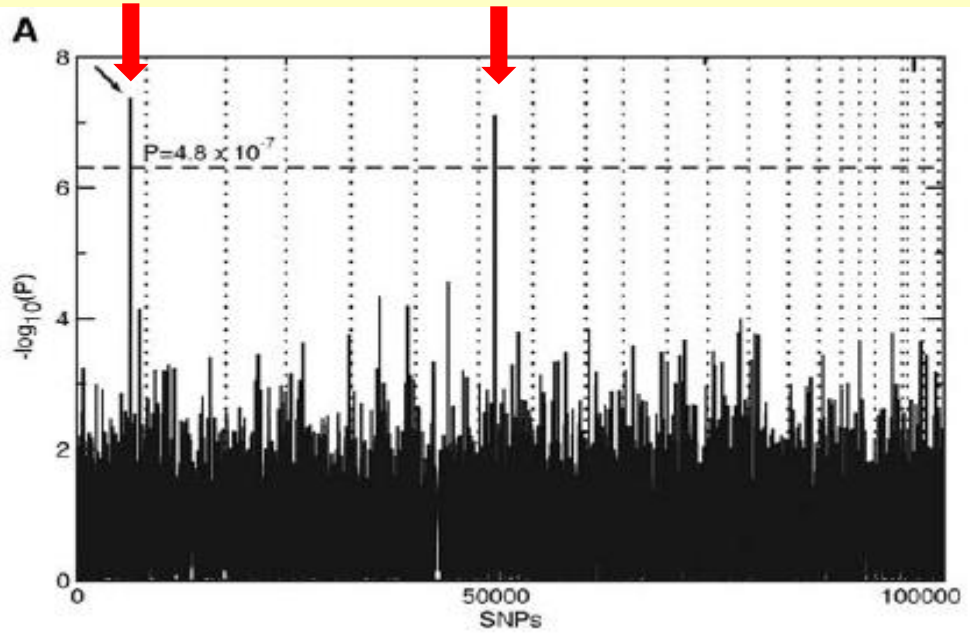
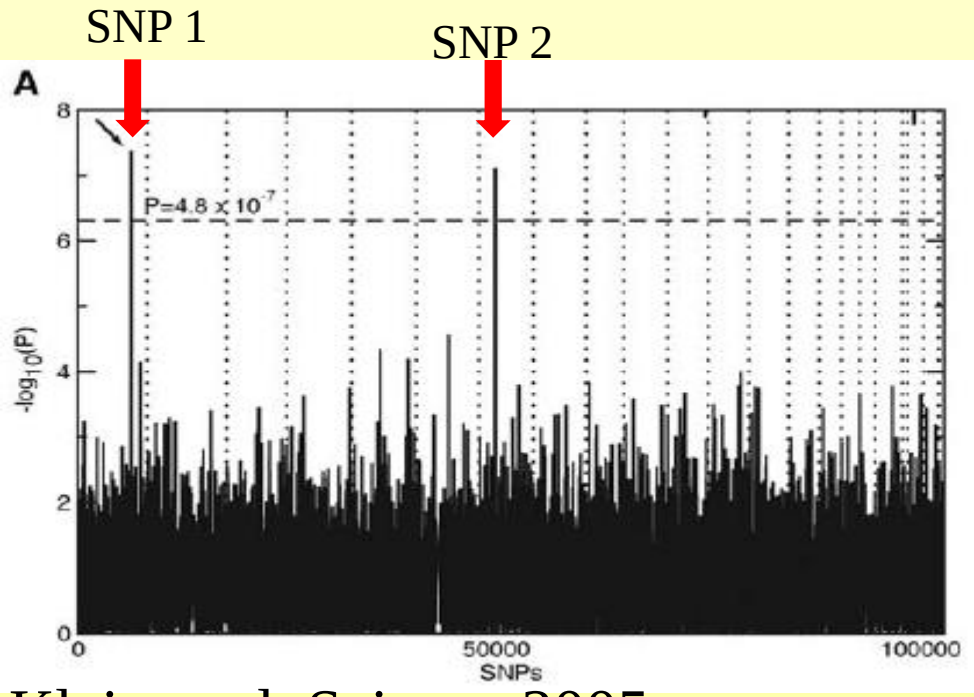


Fig. 1. (A) *P* values of genome-wide association scan for genes that affect the risk of developing AMD. $-\log_{10}(p)$ is plotted for each SNP in chromosomal order. The spacing between SNPs on the plot is uniform and does not reflect distances between SNPs on the chromosomes. The dotted horizontal line shows the cutoff for $P = 0.05$ after Bonferroni correction. The vertical dotted lines show chromosomal boundaries. The arrow indicates the peak for SNP rs380390, the most significant association, which was studied further. (B) Variation in genotype frequencies between cases and controls.

P level was corrected for the number of statistical tests

Klein et al. Science 2005

GWAS AMD



Klein et al. Science 2005

Genotyping failed for 21 individuals in SNP 2. Resequencing of these 21 individuals for SNP 2 and retesting showed that SNP2 became not significant

SNP 1 lies in an intron of the gene for complement factor H (CFH), a gene related to the immune system.

The CFH gene was resequenced to find a mutation strongly correlated with AMD. CFH is a key regulator of the complement system of innate immunity.

GWAS AMD

Conclusions

One QTL was detected for AMD

QTL referred to an intron in the CHF gene

Subsequent sequencing identified a SNP in the gene leading to amino acid changes and that correlated highly with AMD

GWAS Grain shape traits in rice

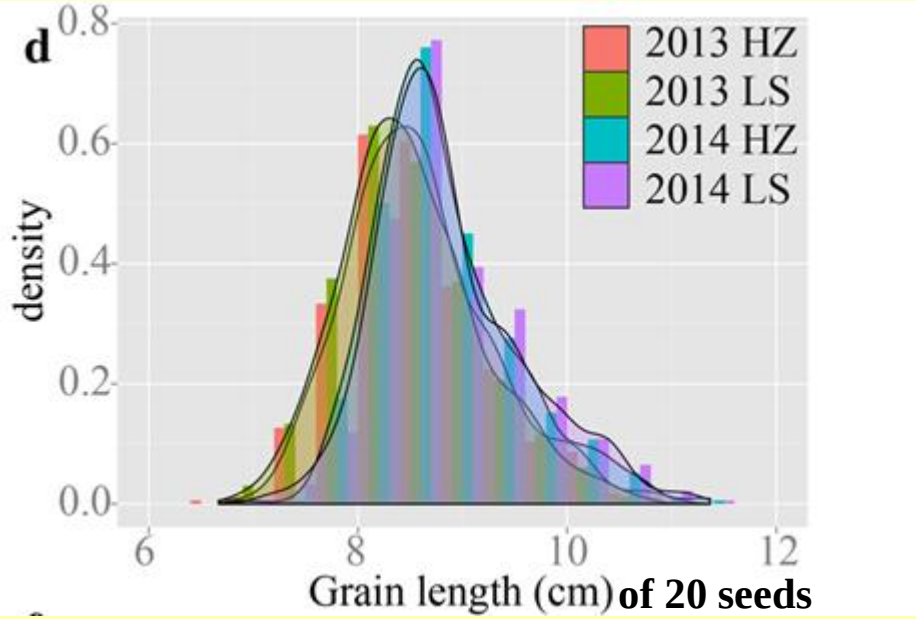
Variation in length of rice grain

Feng et al. 2016

Variation in width of rice grain

Grain shape is a key determinant in grain yield in rice

GWAS Grain shape traits in rice



469 worldwide rice accessions were grown on 2 different locations and 2 years in China and length of the grains was measured. Normal distribution suggesting multiple QTLs

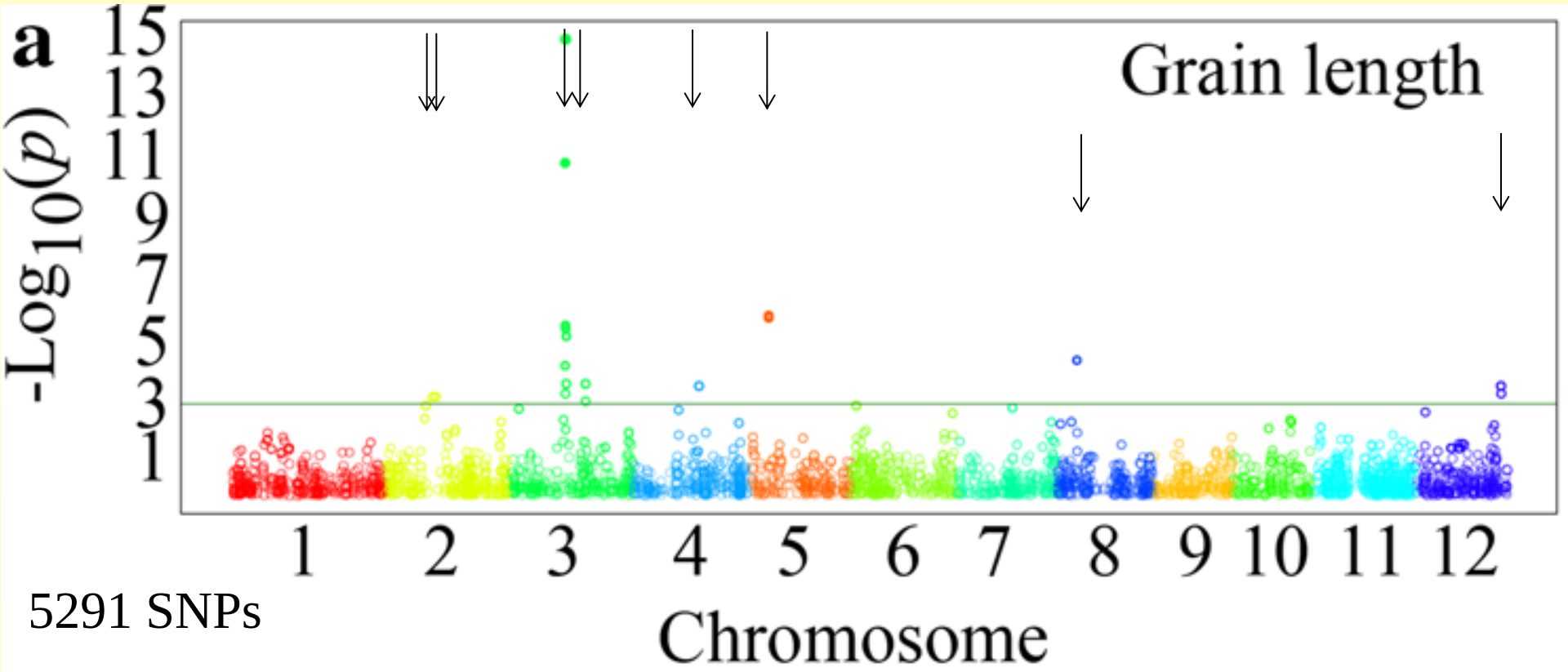
Trait	Environment	Mean ± SE	Range
GL (cm)	2013, HZ	8.58 ± 0.03	6.67–10.92
	2013, LS	8.51 ± 0.03	6.83–10.67
	2014, HZ	8.81 ± 0.03	6.92–11.37
	2014, LS	8.86 ± 0.03	7.11–11.20

Broad sense heritability is 79%

Feng et al. 2016

GWAS Grain shape traits in rice

Feng et al. 2016



5291 SNPs

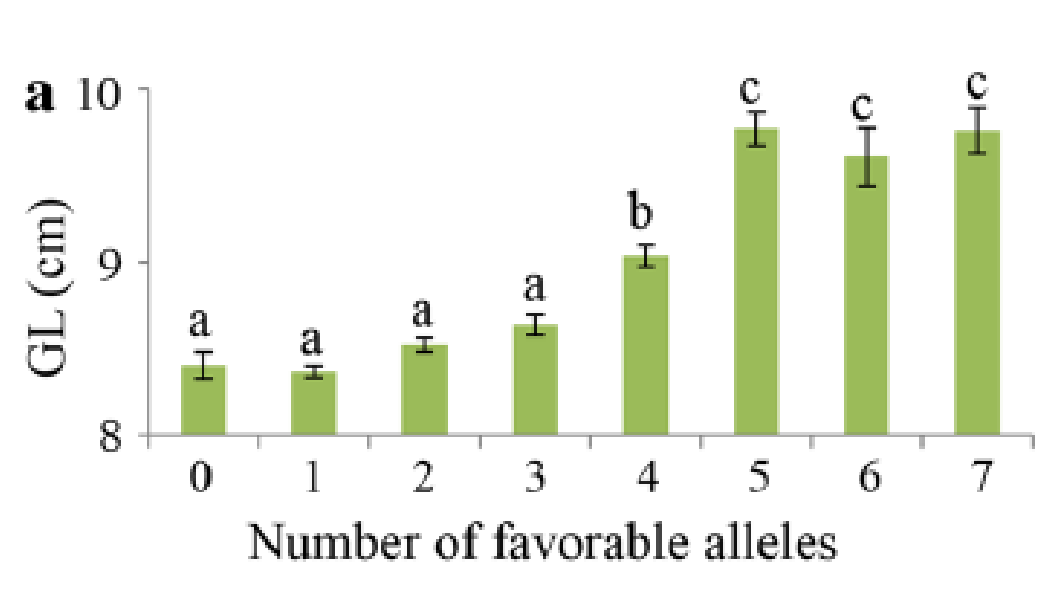
Trait	Marker	Chr.	Position	Allele[a]	MAF	*P* value	R^2 (%)[b]	Known QTL
GL	seq-rs918	2	13760905	G/A	0.17	6.68E−04	50.23	
	seq-rs919	2	14722011	T/G	0.17	6.68E−04		
	seq-rs1614	3	16939138	T/C	0.25	1.87E−15		*GS3*
	seq-rs1697[c]	3	22579680	G/T	0.21	2.63E−04		*qGL3a*
	seq-rs2123	4	19733128	C/T	0.05	2.89E−04		*qGL4b*
	gwseq-rs14[d]	5	5377176	T/C	0.21	1.91E−06		*qSW5*
	seq-rs3750	8	5964991	C/T	0.06	4.35E−05		
	seq-rs5920	12	25594275	G/A	0.36	2.91E−04		*qGL12*

GWAS Grain shape traits in rice

Trait	Marker	Chr.	Position	Allele[a]	No. genes in 100kB
GL	seq-rs918	2	13760905	G/A	8
	seq-rs919	2	14722011	T/G	6
	seq-rs1614	3	16939138	T/C	0
	seq-rs1697[c]	3	22579680	G/T	16
	seq-rs2123	4	19733128	C/T	18
	gwseq-rs14[d]	5	5377176	T/C	0
	seq-rs3750	8	5964991	C/T	18
	seq-rs5920	12	25594275	G/A	33

Feng et al. 2016

GWAS Grain shape traits in rice

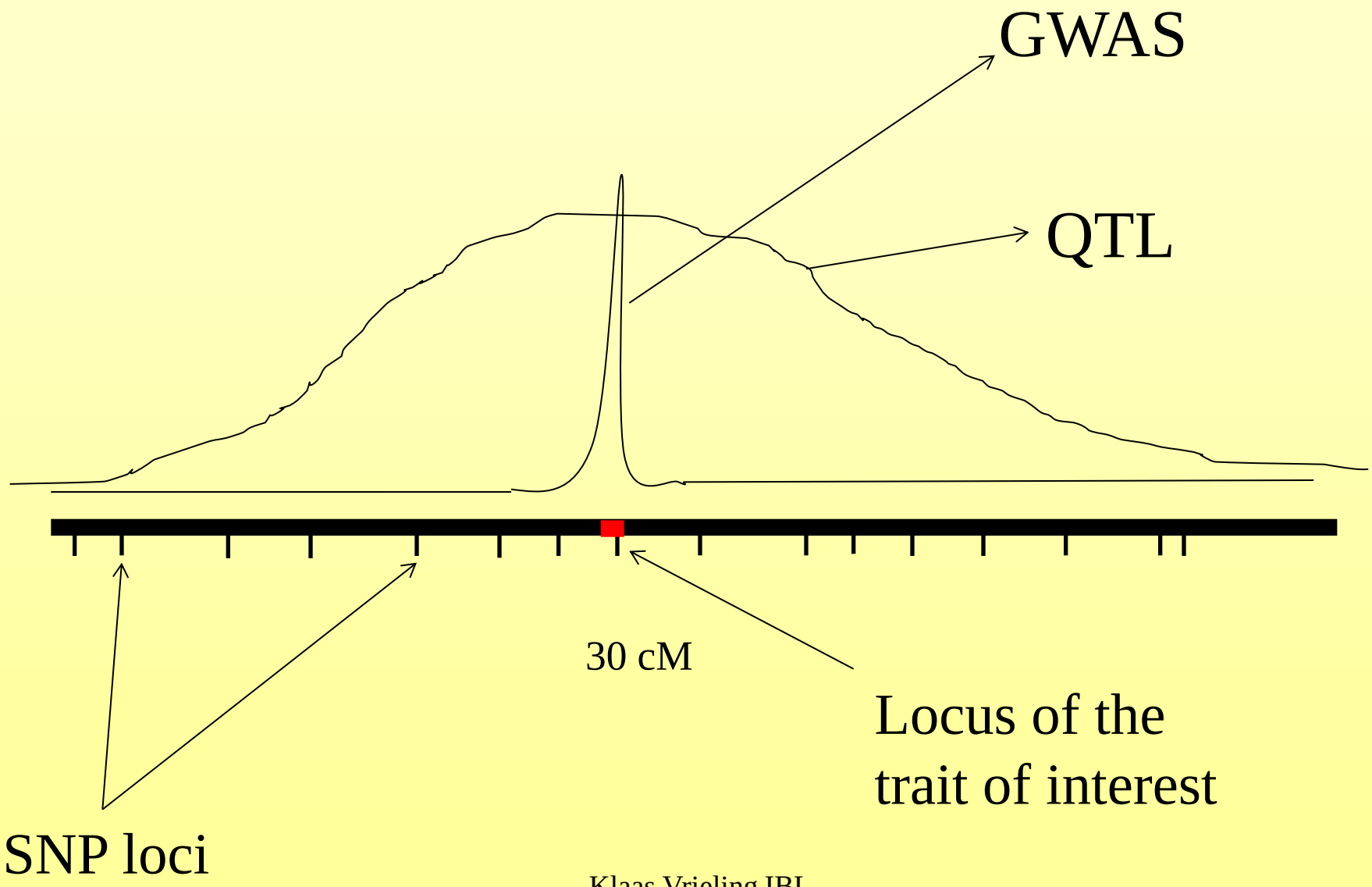


If all favorable alleles are in one rice variety length can be increased by 20%.

Breeders can use the SNP alleles to select the right offspring for breeding larger rice grains (marker assisted breeding)

Feng et al. 2016

Comparing GWAS and QTL



Comparing GWAS and QTL

QTL analysis	GWAS
Related individuals: progeny of a cross	Un related individuals: different accessions
Only loci in the particular cross can be detected	A broader range of the genetic variation is sampled
Circa 100-200 (SNP) markers	>5000 SNP markers
100-500 individuals	100-500 accessions
Used for maker assisted breeding	Used for maker assisted breeding

Comparing GWAS and QTL

QTL analysis

Genomic region targeted circa 20 cM (5.000.000- 20.000.000)

Finding sequence of underlying gene(s) difficult

GWAS

Genomic region targeted circa (10.000 - 200.000 bp)

Finding sequence of underlying gene(s) is feasible because also genomic information is available

Effect of the gene has to be validated by knock-out and or upregulation.

Conclusion

• With a QTL analysis and GWAS a heritable trait can be dissected into several loci.

•A rough indication can be obtained where the underlying gene(s) are located on the genome for QTLs.

•GWAS gives a more precise location on the (sequenced) genome

•Finding the underlying gene is possible with GWAS

•QTL analysis and GWAS can be used for solving evolutionary questions about constraints

•QTL analysis and GWAS can be used for marker assisted breeding