GWAS in structured populations

Magnus Nordborg
Gregor Mendel Institute

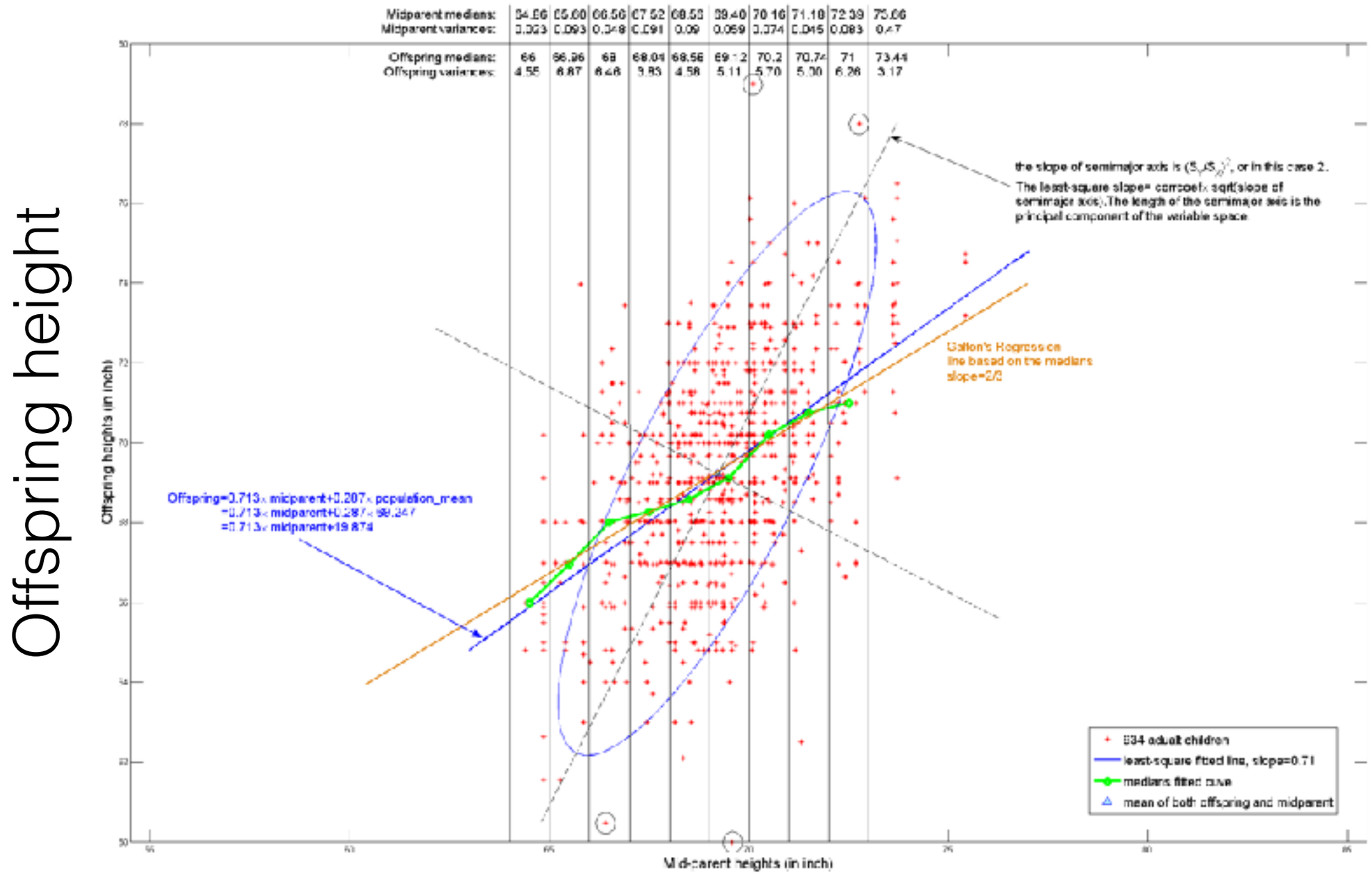
GREGOR MENDEL INSTITUTE
OF MOLECULAR PLANT BIOLOGY

ÖAW
Austrian Academy
of Sciences

The agenda

- How does genetic variation translate (“map”) into phenotypic information, and how is this translation affected by the environment?
- How do we identify (i.e.,“map”) the causal variants?
- Can we predict phenotypes?

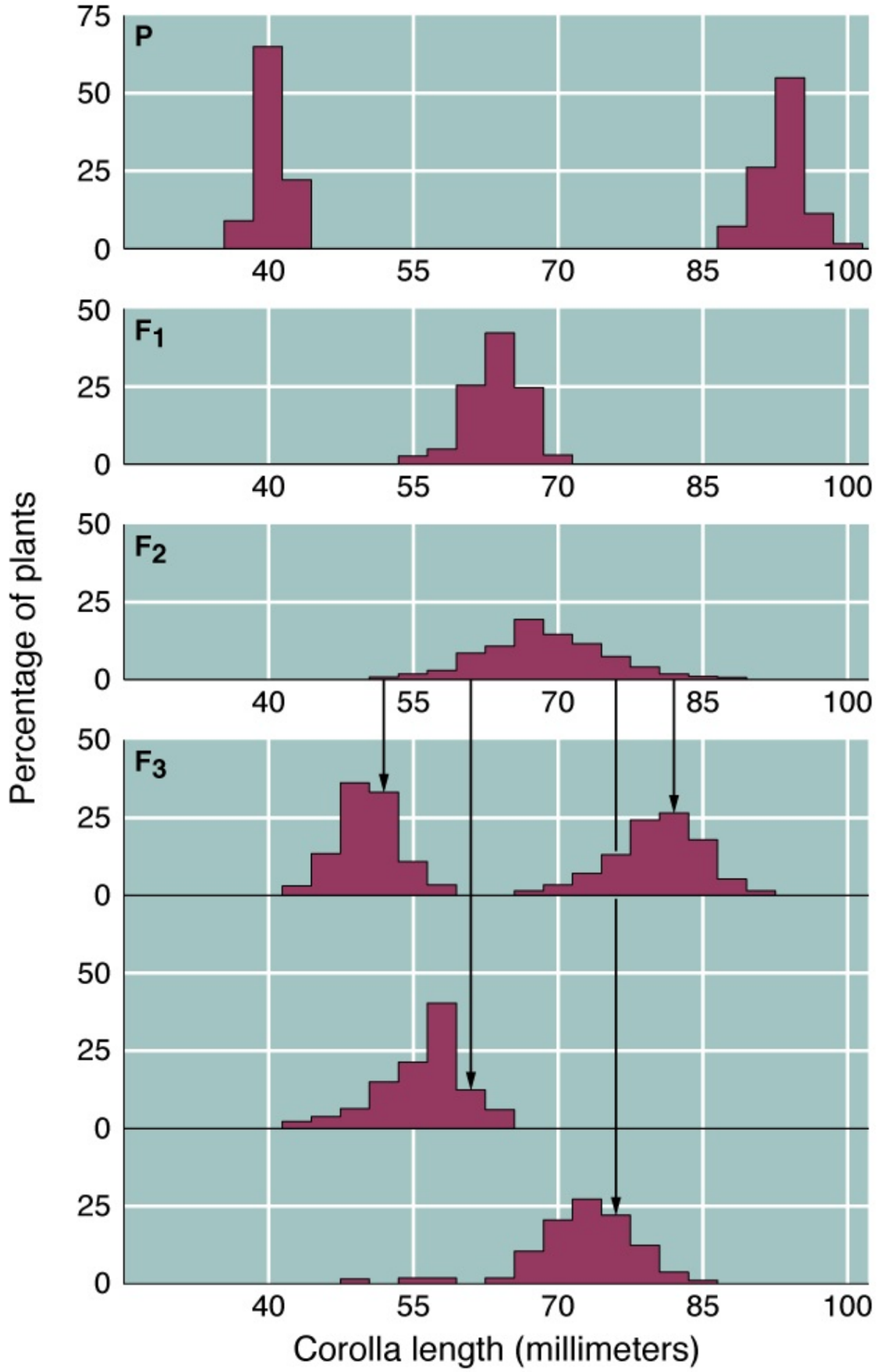
Like begets like
(Sir Francis Galton et al.)



Offspring height

Mid-parent height

How many genes?



The simplest model possible (Fisher 1918)

Genotype	AA	Aa	aa
Phenotypic value	*a*	*d*	*-a*
Frequency	p^2	$2pq$	q^2

population mean:

$$M = ap^2 + 2dpq - aq^2$$
$$= a(p - q) + 2dpq$$

Dominance and epistasis

- Dominance reflects deviation from additivity between alleles at a locus

- Epistasis reflects deviation from additivity between loci

Average effect

- The *average effect of an allele* is the mean deviation from the population mean of individuals with the allele and everything else chosen randomly

- The *average effect of a gene substitution* is the expected effect of substituting one allele for the other

Variance decomposition

- The phenotypic variance can be decomposed into the genetic and environmental part:

$$V_P = V_G + V_E$$

- But all this assumes lack of interactions:

$$V_P = V_A + V_D + V_I + V_E + \mathrm{COV}_{GE} + \ldots$$

Heritability

- Heritability if the proportion of the phenotypic variance that is due to genetic factors

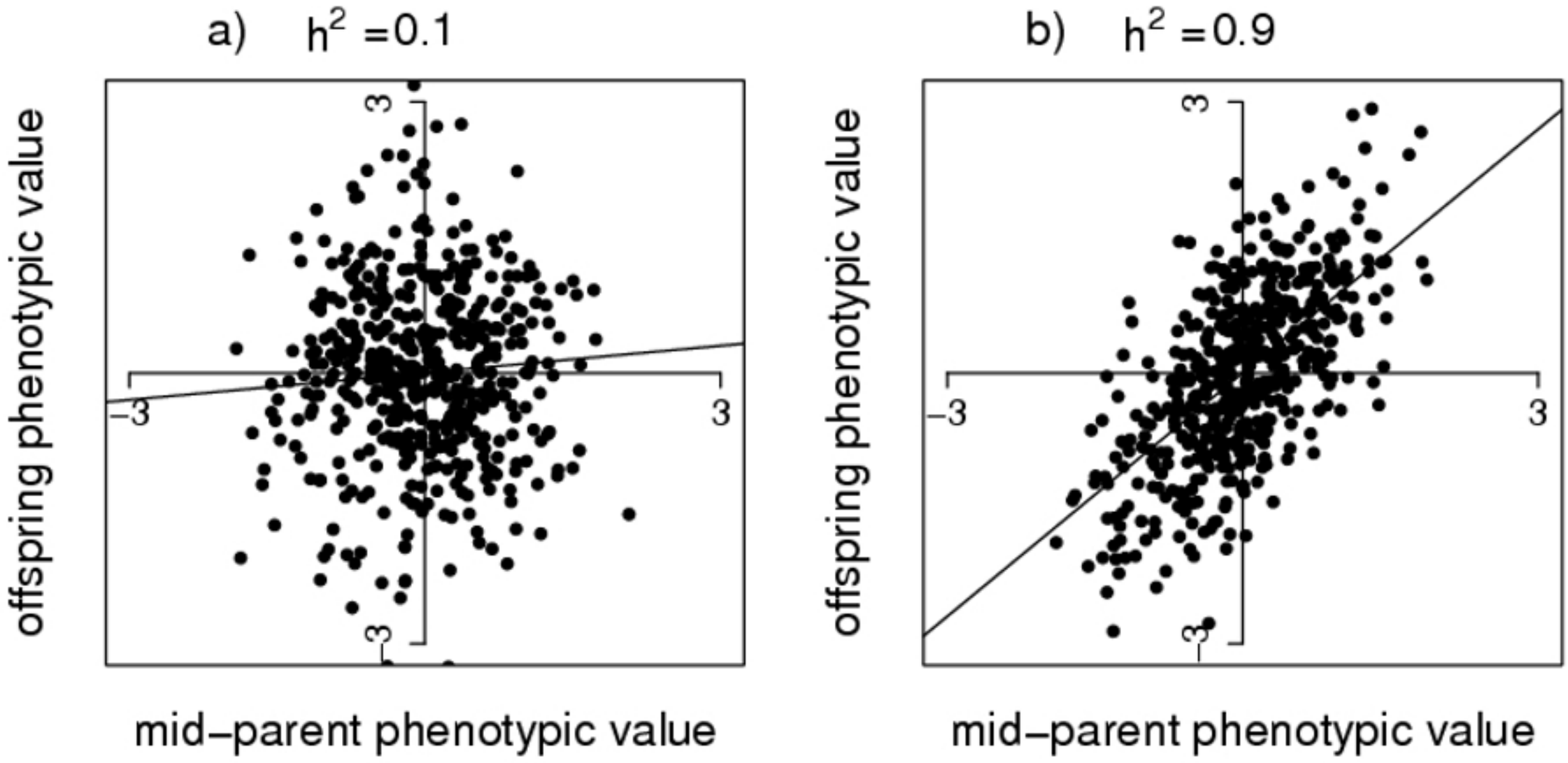
- Broad-sense:

$$H^2 = V_G/V_P$$

- Narrow-sense:

$$h^2 = V_A/V_P$$

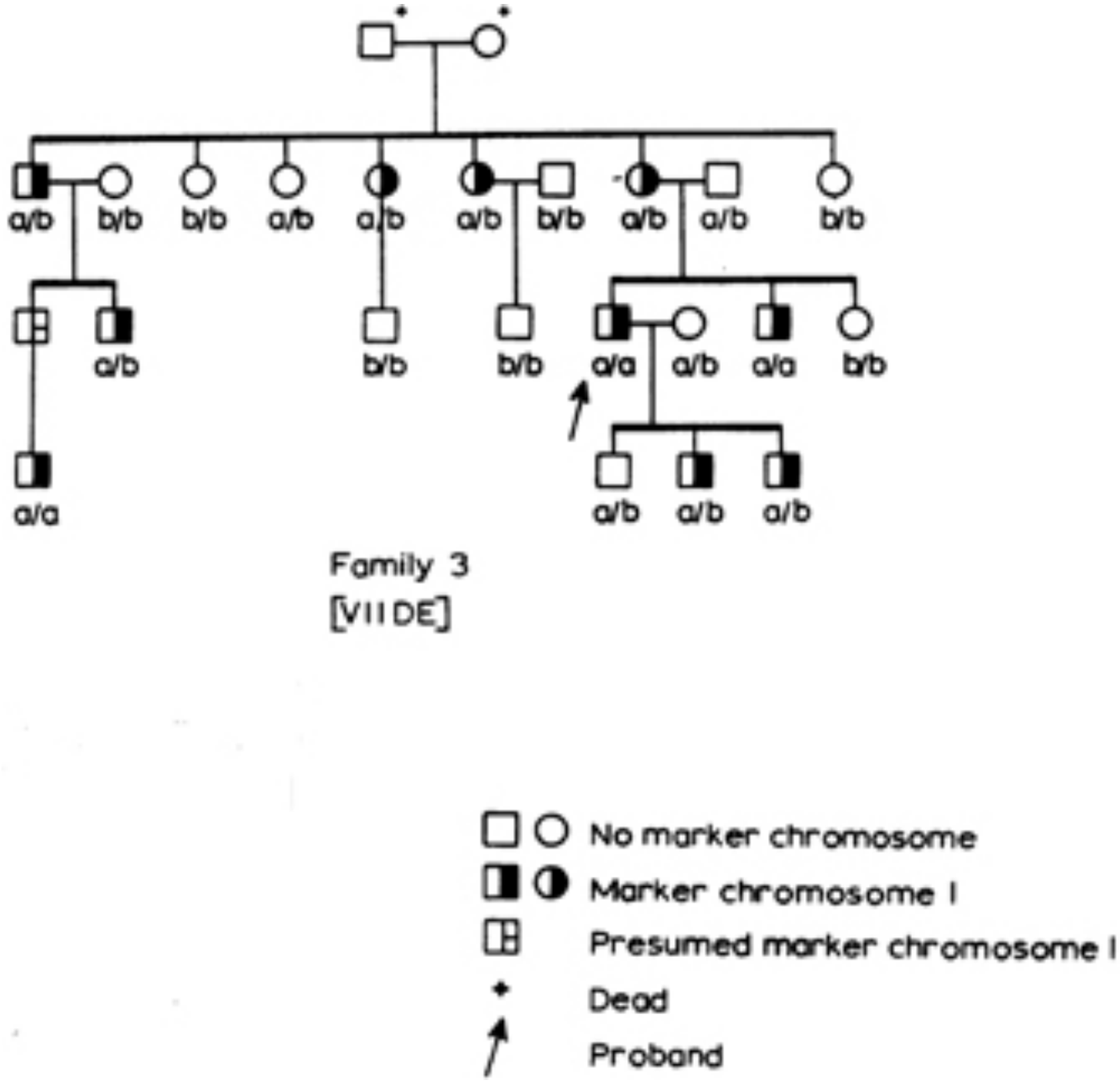
Typically estimated from parent-offspring regression



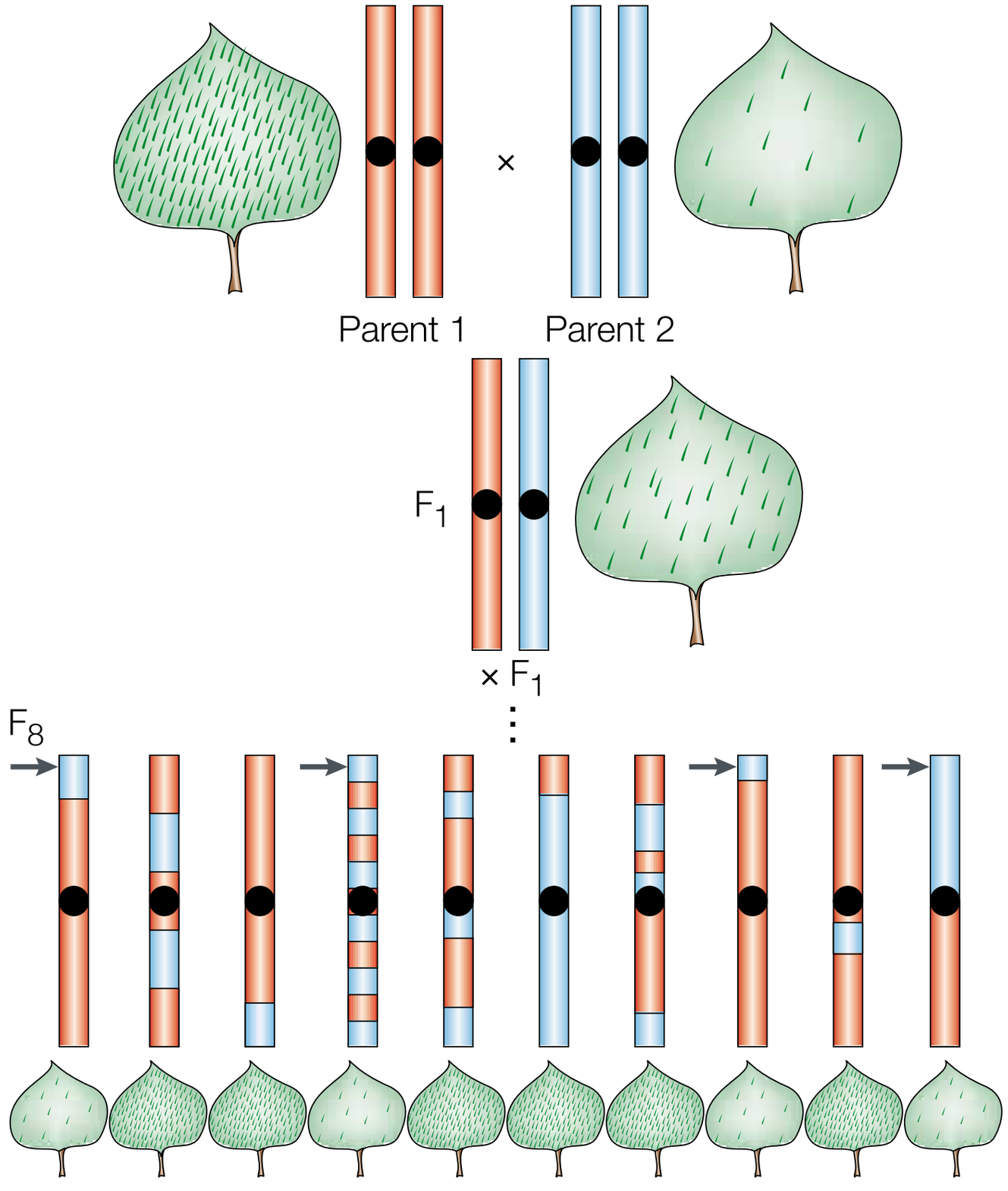
…but beware of environmental correlations

How do we map genes?

Linkage mapping

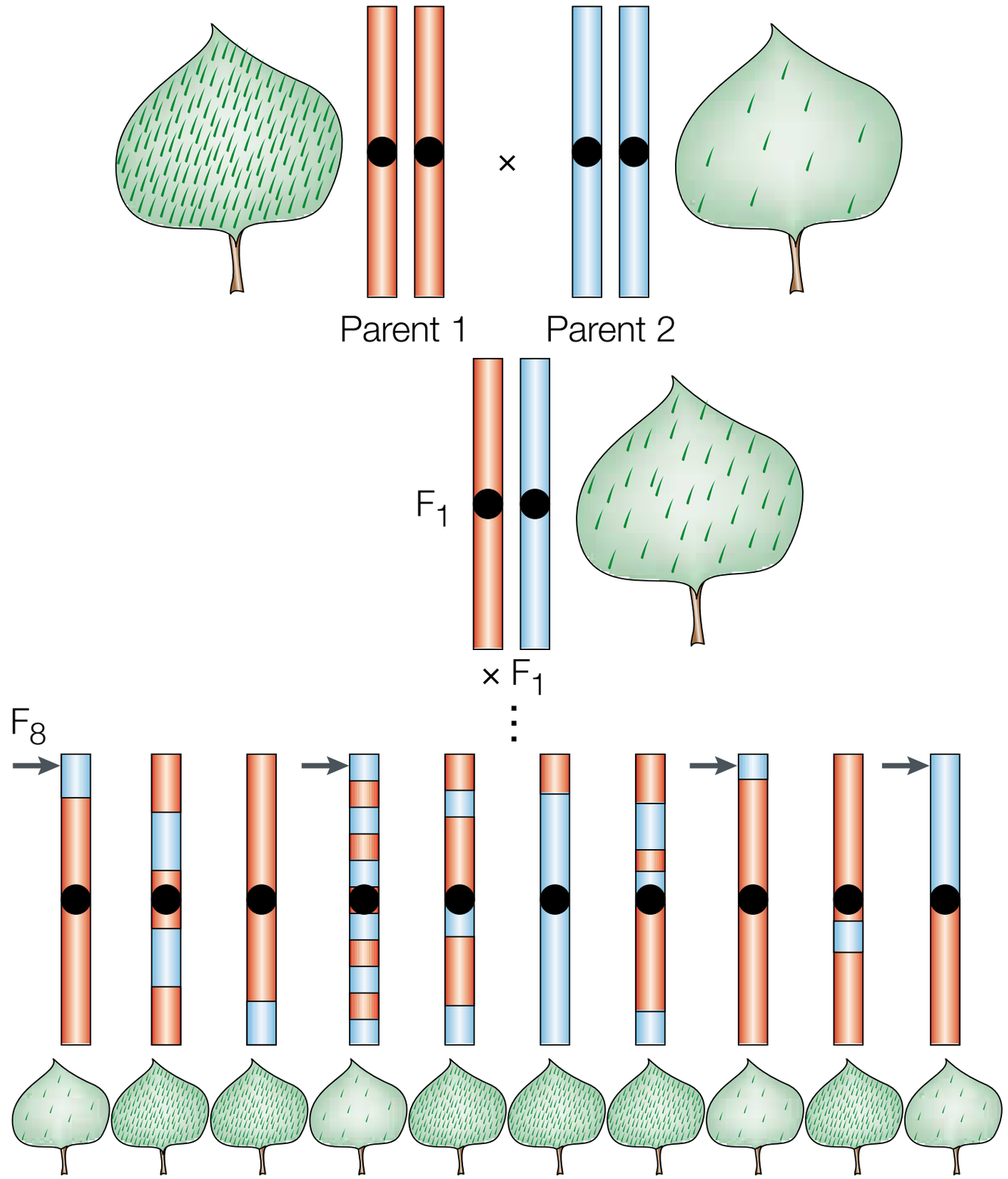


Linkage mapping



Linkage mapping

***Extremely* low resolution!**
(limited by the number of informative meioses in pedigree or cross)



So why do it?

- *Historically, resolution was limited by markers!*
- High power because causative alleles are segregating at high frequencies — always works (if sample size is large enough)
- Controlled environment (experimental crosses)

…if sample size is large enough…

- Remember Fisher's model?

- A quantitative trait can be highly heritable yet have the property that no single polymorphism has large enough effect to be mapped or studied (using realistic sample sizes)

- In fact, in the early days of QTL mapping, many quantitative geneticists thought the entire idea of mapping genes ridiculous…

Genome-wide association studies (GWAS)

The Human Genome Project meant markers were no longer limiting…

Prospects for whole-genome linkage disequilibrium mapping of common disease genes

Leonid Kruglyak

Recently, attention has focused on the use of whole-genome linkage disequilibrium (LD) studies to map common disease genes. Such studies would employ a dense map of single nucleotide polymorphisms (SNPs) to detect association between a marker and disease. Construction of SNP maps is currently underway. An essential issue yet to be settled is the required marker density of such maps. Here, I use population simulations to estimate the extent of LD surrounding common gene variants in the general human population as well as in isolated populations. Two main conclusions emerge from these investigations. First, a useful level of LD is unlikely to extend beyond an average distance of roughly 3 kb in the general population, which implies that approximately 500,000 SNPs will be required for whole-genome studies. Second, the extent of LD is similar in isolated populations unless the founding bottleneck is very narrow or the frequency of the variant is low (<5%).

The promise

- Almost gene-level resolution (kb instead of Mb)
- No need for pedigrees!
- But does it work?

The HapMap Project

Vol 447|7 June 2007|**doi:10.1038/nature05911**

nature

ARTICLES

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

There is increasing evidence that genome-wide association (GWA) studies represent a powerful approach to the identification of genes involved in common human diseases. We describe a joint GWA study (using the Affymetrix GeneChip 500K Mapping Array Set) undertaken in the British population, which has examined ~2,000 individuals for each of 7 major diseases and a shared set of ~3,000 controls. Case-control comparisons identified 24 independent association signals at $P < 5 \times 10^{-7}$: 1 in bipolar disorder, 1 in coronary artery disease, 9 in Crohn's disease, 3 in rheumatoid arthritis, 7 in type 1 diabetes and 3 in type 2 diabetes. On the basis of prior findings and replication studies thus-far completed, almost all of these signals reflect genuine susceptibility effects. We observed association at many previously identified loci, and found compelling evidence that some loci confer risk for more than one of the diseases studied. Across all diseases, we identified a large number of further signals (including 58 loci with single-point *P* values between 10^{-5} and 5×10^{-7}) likely to yield additional susceptibility loci. The importance of appropriately large samples was confirmed by the modest effect sizes

Bipolar disorder

Coronary artery disease

Crohn's disease

Hypertension

Rheumatoid arthritis

Type 1 diabetes

Type 2 diabetes

$-\log_{10}(P)$

Chromosome

nature genetics

The GWAS debacle

Genome-wide association analysis identifies 20 loci that influence adult height

Michael N Weedon[1,2,23], Hana Lango[1,2,23], Cecilia M Lindgren[3,4], Chris Wallace[5], David M Evans[6], Massimo Mangino[7], Rachel M Freathy[1,2], John R B Perry[1,2], Suzanne Stevens[7], Alistair S Hall[8], Nilesh J Samani[7], Beverly Shields[2], Inga Prokopenko[3,4], Martin Farrall[9], Anna Dominiczak[10], Diabetes Genetics Initiative[21], The Wellcome Trust Case Control Consortium[21], Toby Johnson[11–13], Sven Bergmann[11,12], Jacques S Beckmann[11,14], Peter Vollenweider[15], Dawn M Waterworth[16], Vincent Mooser[16], Colin N A Palmer[17], Andrew D Morris[18], Willem H Ouwehand[19,20], Cambridge GEM Consortium[22], Mark Caulfield[5], Patricia B Munroe[5], Andrew T Hattersley[1,2], Mark I McCarthy[3,4] & Timothy M Frayling[1,2]

Adult height is a model polygenic trait, but there has been limited success in identifying the genes underlying its normal variation. To identify genetic variants influencing adult human height, we used genome-wide association data from 13,665 individuals and genotyped 39 variants in an additional 16,482 samples. We identified 20 variants associated with adult height ($P < 5 \times 10^{-7}$, with 10 reaching $P < 1 \times 10^{-10}$). Combined, the 20 SNPs explain ~3% of height variation, with a ~5 cm difference between the 6.2% of people with 17 or fewer 'tall' alleles compared to the 5.5% with 27 or more 'tall' alleles. The loci we identified implicate genes in Hedgehog signaling (*IHH, HHIP, PTCH1*), extracellular matrix (*EFEMP1, ADAMTSL3, ACAN*) and cancer (*CDK6, HMGA2, DLEU7*) pathways, and provide new insights into human growth and developmental processes. Finally, our results provide insights into the genetic architecture of a classic quantitative trait.

The GWAS debacle

nature genetics

Genome-wide associatio
influence adult height

Michael N Weedon[1,2,23], Hana Lango[1,2,23],
Massimo Mangino[7], Rachel M Freathy[1,2], J
Nilesh J Samani[7], Beverly Shields[2], Inga Prok
Initiative[21], The Wellcome Trust Case Control Co
Jacques S Beckmann[11,14], Peter Vollenweider[15], Dawn
Andrew D Morris[18], Willem H Ouwehand[19,20], Ca
Patricia B Munroe[5], Andrew T Hattersley[1,2], M arthy[3,4] & Timothy M Frayling[1,2]

Adult height is a model polygenic trait, but there has been limited success in identifying the genes underlying its normal variation. To identify genetic variants influencing adult human height, we used genome-wide association data from 13,665 individuals and genotyped 39 variants in an additional 16,482 samples. We identified 20 variants associated with adult height ($P < 5 \times 10^{-7}$, with 10 reaching $P < 1 \times 10^{-10}$). Combined, the 20 SNPs explain ~3% of height variation, with a ~5 cm difference between the 6.2% of people with 17 or fewer 'tall' alleles compared to the 5.5% with 27 or more 'tall' alleles. The loci we identified implicate genes in Hedgehog signaling (*IHH, HHIP, PTCH1*), extracellular matrix (*EFEMP1, ADAMTSL3, ACAN*) and cancer (*CDK6, HMGA2, DLEU7*) pathways, and provide new insights into human growth and developmental processes. Finally, our results provide insights into the genetic architecture of a classic quantitative trait.

Yengo et al. (2018, bioRxiv, 10.1101/274654):

- ~700,000 samples
- 3,290 SNPs
- ~24.6% of the variance

...height has a heritability of 80%!

Vol 461|8 October 2009|**doi:10.1038/nature08494**

nature

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively small increments in risk, and explain only a small proportion of familial clustering, leading many to question how the remaining, 'missing' heritability can be explained. Here we examine potential sources of missing heritability and propose research strategies, including and extending beyond current genome-wide association approaches, to illuminate the genetics of complex diseases and enhance its potential to enable effective disease prevention or treatment.

NATURE|Vol 456|6 November 2008

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

If you want to predict how tall your children

Even though these genome-wide association studies (GWAS) turned up dozens of variants, contribute to a variety of traits and common diseases. But even when dozens of genes have

ARKINS

nature genetics

ANALYSIS

...of tiny effect...

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of complex traits in human populations. Where is the remaining heritability? We estimated the proportion of variance for human height explained by 294,831 SNPs genotyped on 3,925 unrelated individuals using a linear model analysis, and validated the estimation method with simulations based on the observed genotype data. We show that 45% of variance can be explained by considering all SNPs simultaneously. Thus, most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests. We provide evidence that the remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, exacerbated by causal variants having lower minor allele frequency than the SNPs explored to date.

of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium (LD) with the SNPs that have been genotyped. Lack of complete LD might, for instance, occur if causal variants have lower minor allele frequency (MAF) than genotyped SNPs. Here we test these two hypotheses and estimate the contribution of each to the heritability of height in humans as a model complex trait.

Height in humans is a classical quantitative trait, easy to measure and studied for well over a century as a model for investigating the genetic basis of complex traits[9,10]. The heritability of height has been estimated to be ~0.8 (refs. 9,11–13). Rare mutations that cause extreme short or tall stature have been found[14,15], but these do not explain much of the variation in the general population. Recent GWASs on tens of thousands of individuals have detected ~50 variants that are associated with height in the population, but these in total account for only ~5% of phenotypic variance[16–19].

Data from a GWAS that are collected to detect statistical associations

A random-effects model is used to estimate variance components

The phenotype, *y*, is written

$$y = u + \epsilon$$

where

$$u \sim N(0, \sigma_g K)$$

are random deviates, with covariance determined by the kinship matrix, *K*, and

$$\epsilon \sim N(0, \sigma_e)$$

is noise.

In other words…

- The SNPs that are significantly associated with height (in the marginal sense) jointly explain a tiny fraction of the variance

- The joint effect of all SNPs (or even a reasonable subset), when used to estimate kinship (relatedness), explain a much greater fraction of the variance

Could have told you so!

R. A. Fisher, "The Correlation Between Relatives on the Supposition of Mendelian Inheritance", *Transactions of the Royal Society of Edinburgh*, 52:399-433, 1918

“The world’s most expensive test of the mutation-selection balance hypothesis”

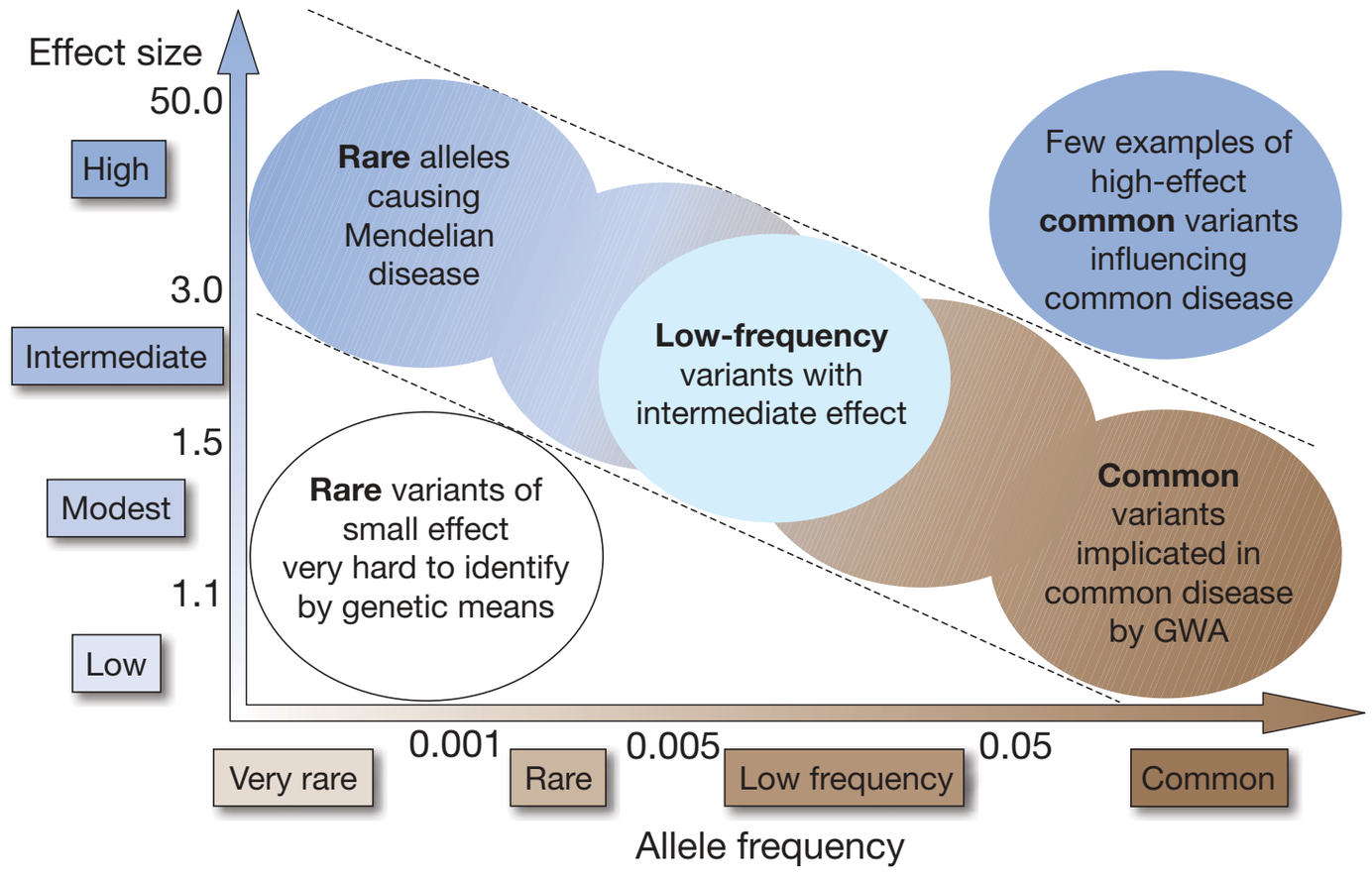
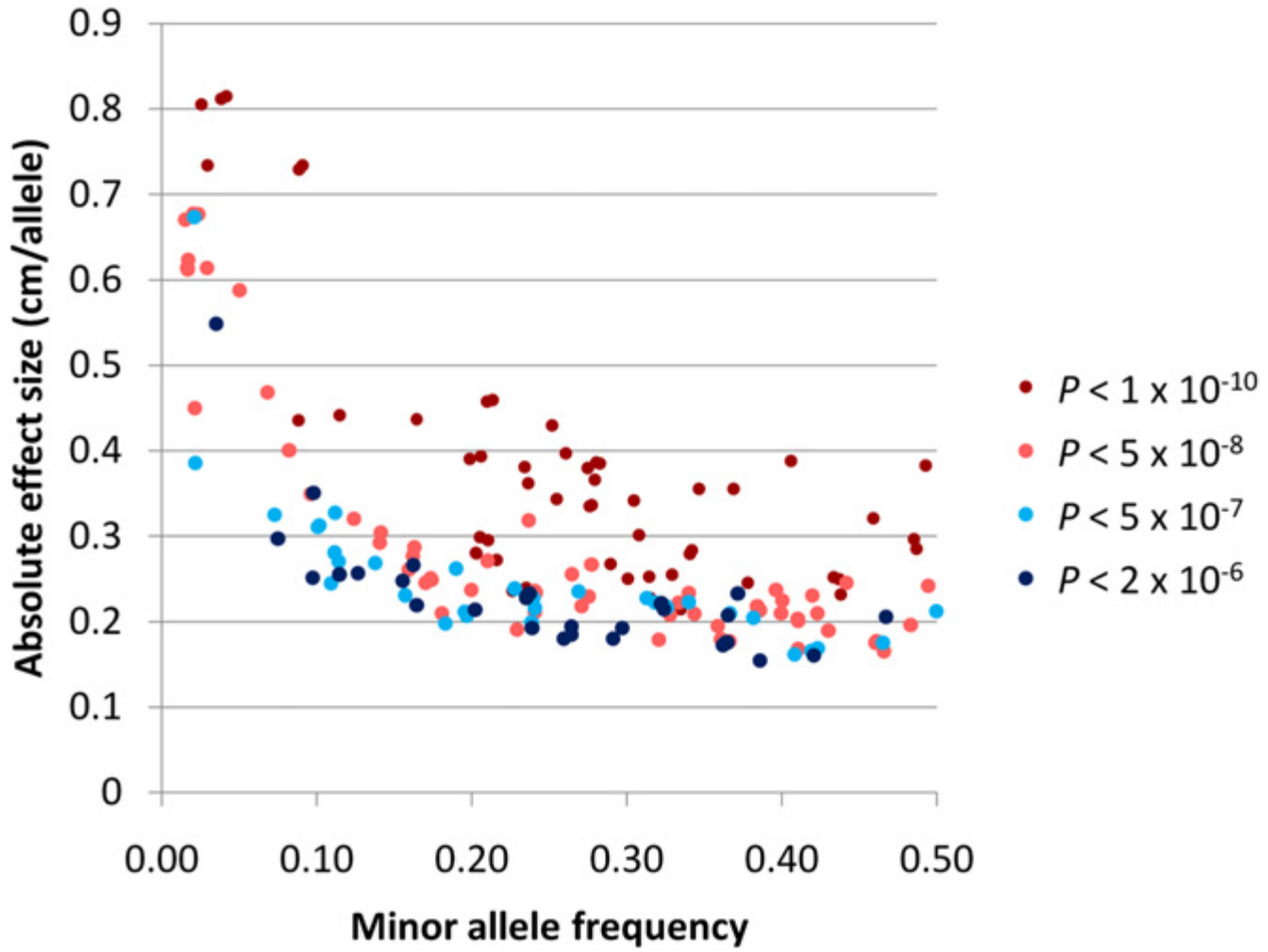


Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

“The world’s most expensive test of the mutation-selection balance hypothesis”



from Lanktree *et al.* 2011, *Amer. J. Hum. Genet.*

Figure 2. The Effect Size of Identified Height-Associated Genetic Variants as a Function of Minor Allele Frequency
Each point is colored by the strength of association observed in the phase III meta-analysis.

There were Cassandras…

How many diseases does it take to map a gene with SNPs?

Kenneth M. Weiss[1] & Joseph D. Terwilliger[2]

"They all talked at once, their voices insistent and contradictory and impatient, making of unreality a possibility, then a probability, then an incontrovertible fact, as people will when their desires become words."
—W. Faulkner, *The Sound and the Fury*, 1929

Through rose-coloured glasses darkly

There are more than a few parallels between the California gold rush and today's frenetic drive towards linkage disequilibrium (LD) mapping based on single-nucleotide polymorphisms (SNPs). This is fuelled by a faith that the genetic determinants of complex traits

A SNP is not the same as a disease-predisposing allele

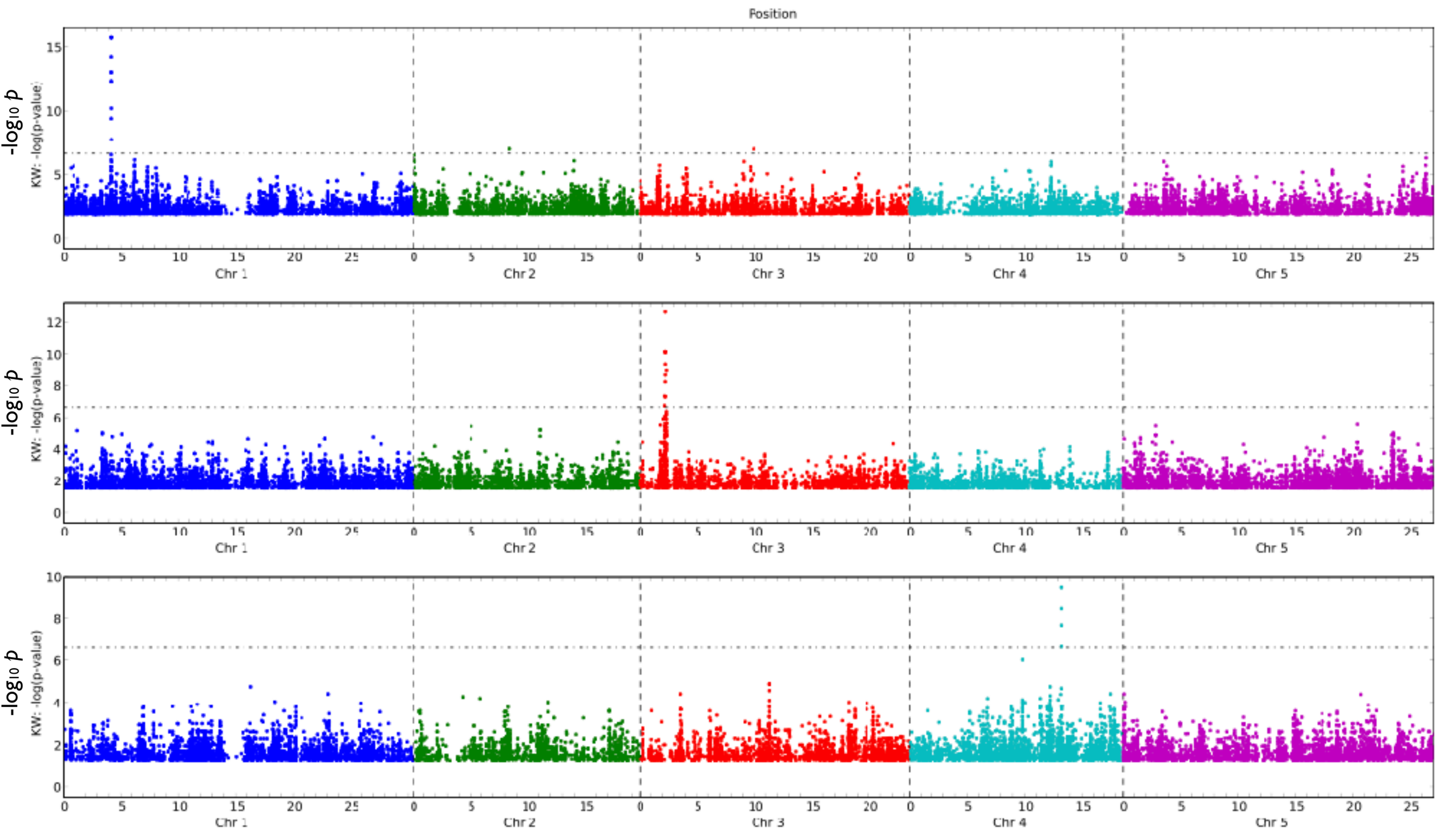
A linkage or LD analysis would test the null hypothesis that alleles of some gene (G_p) that influence some phenotype (Ph) are inherited independently of alleles at some specific chromosomal position (G_X; Box 1, equation 1). In this case, the only

Population structure confounding in GWAS

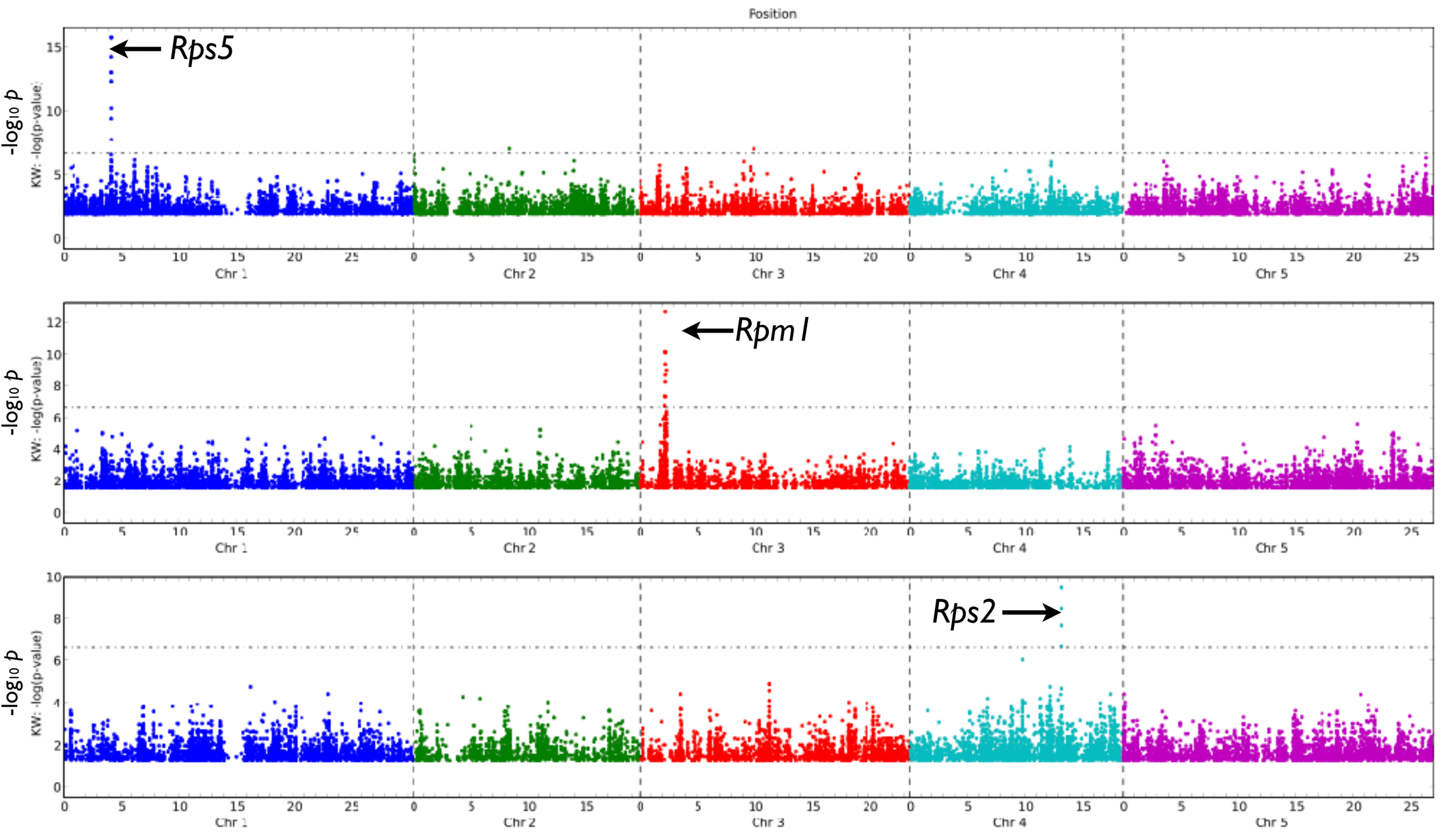
…meanwhile, people doing GWAS in model organisms* had run into a different problem…

*maize, mouse, *Arabidopsis*

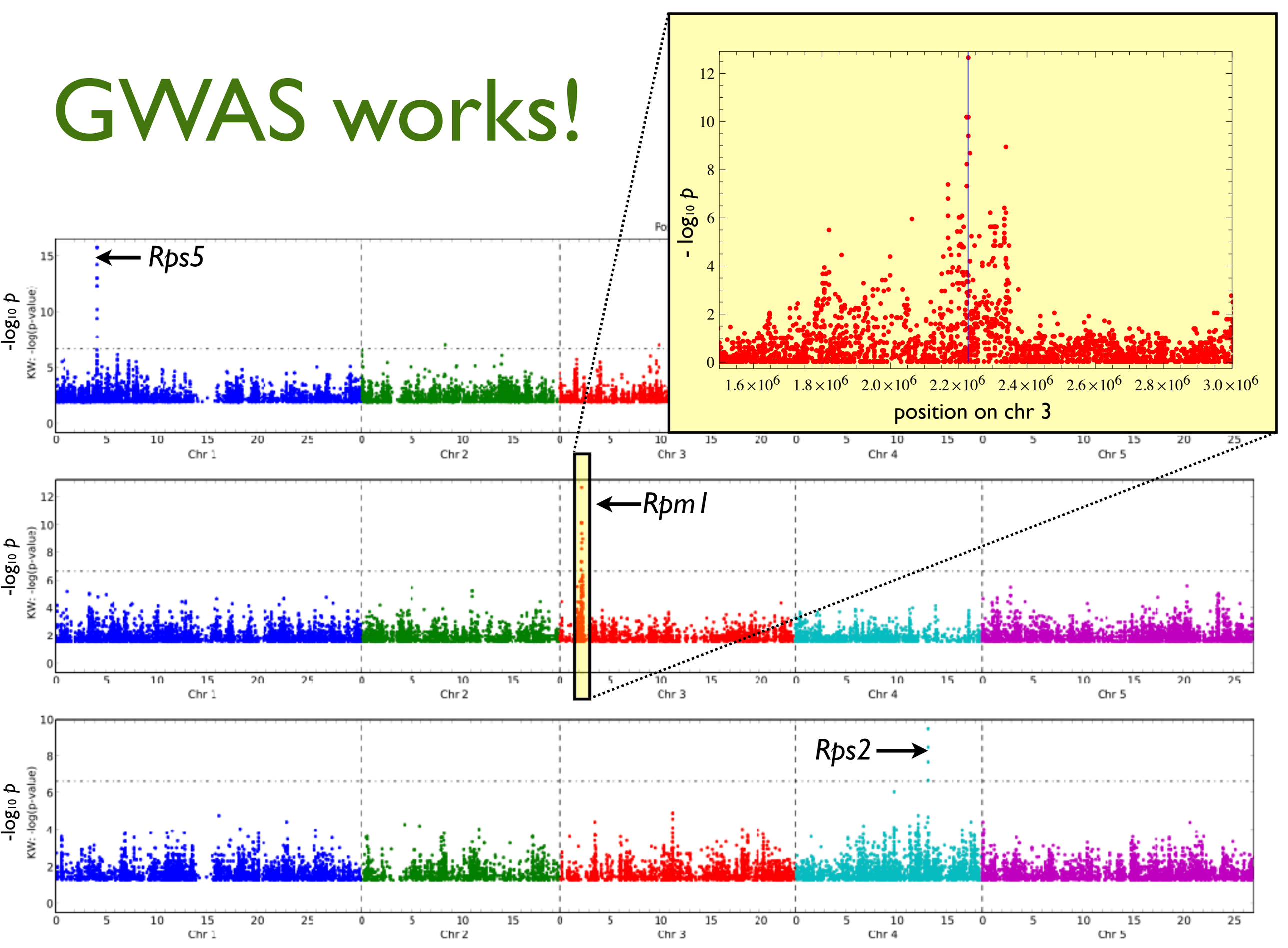
GWAS works!



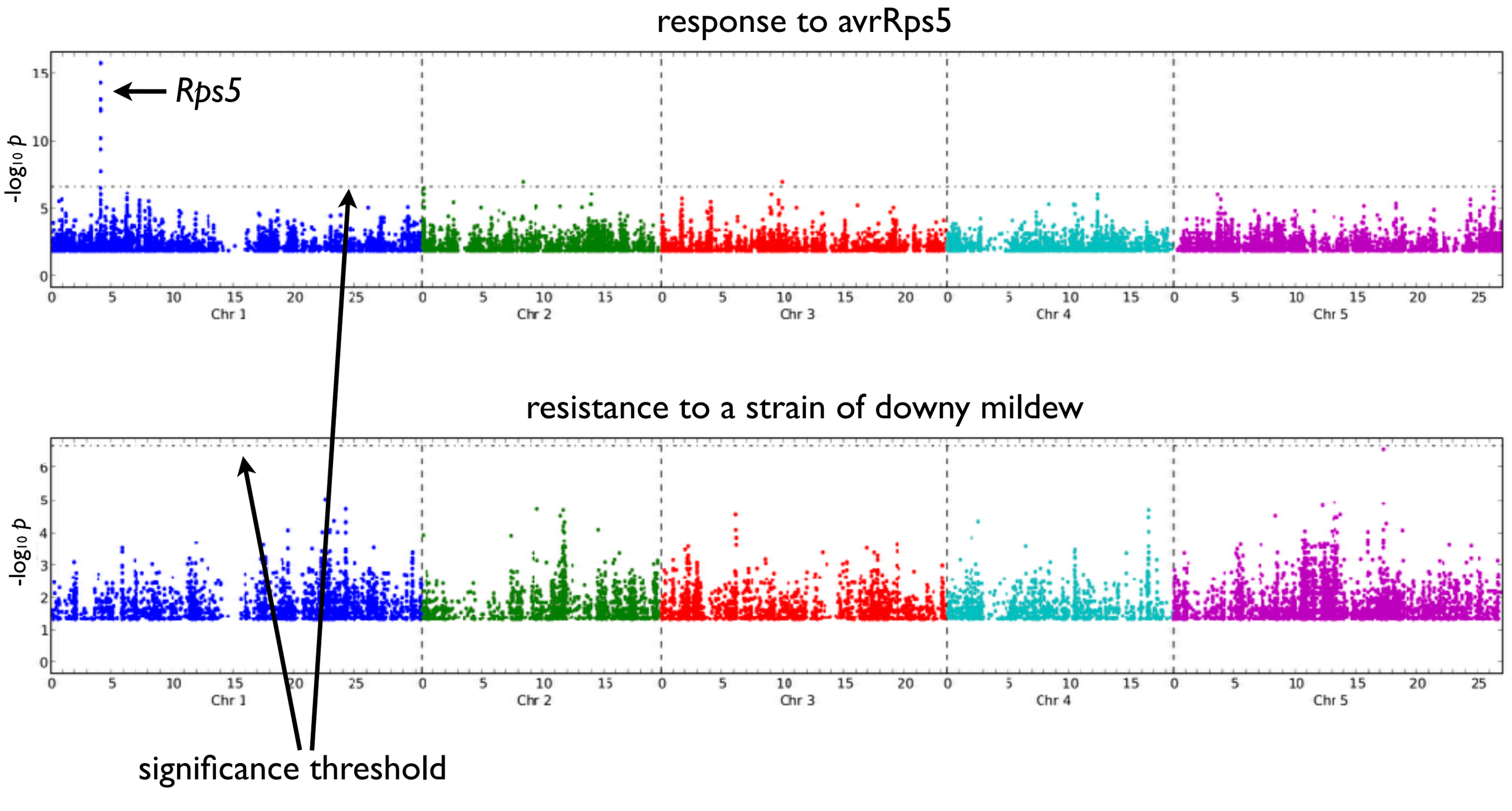
GWAS works!


Position
-log10 p
KW: -log(p-value)
Rps5
Rpm1
Rps2
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5

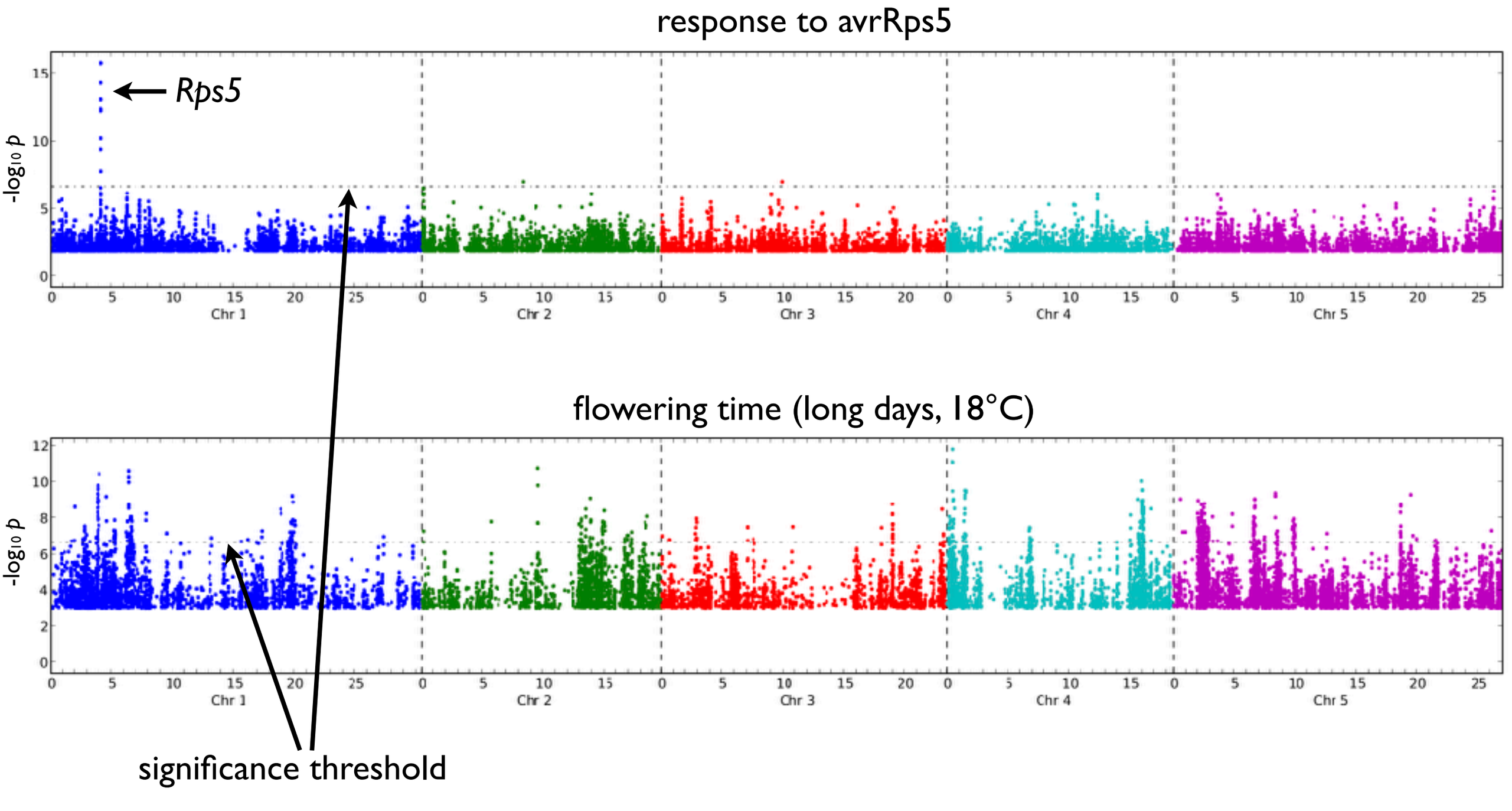
GWAS works!


Rps5
Rpm1
Rps2
-log10 p
KW: -log(p-value)
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
position on chr 3

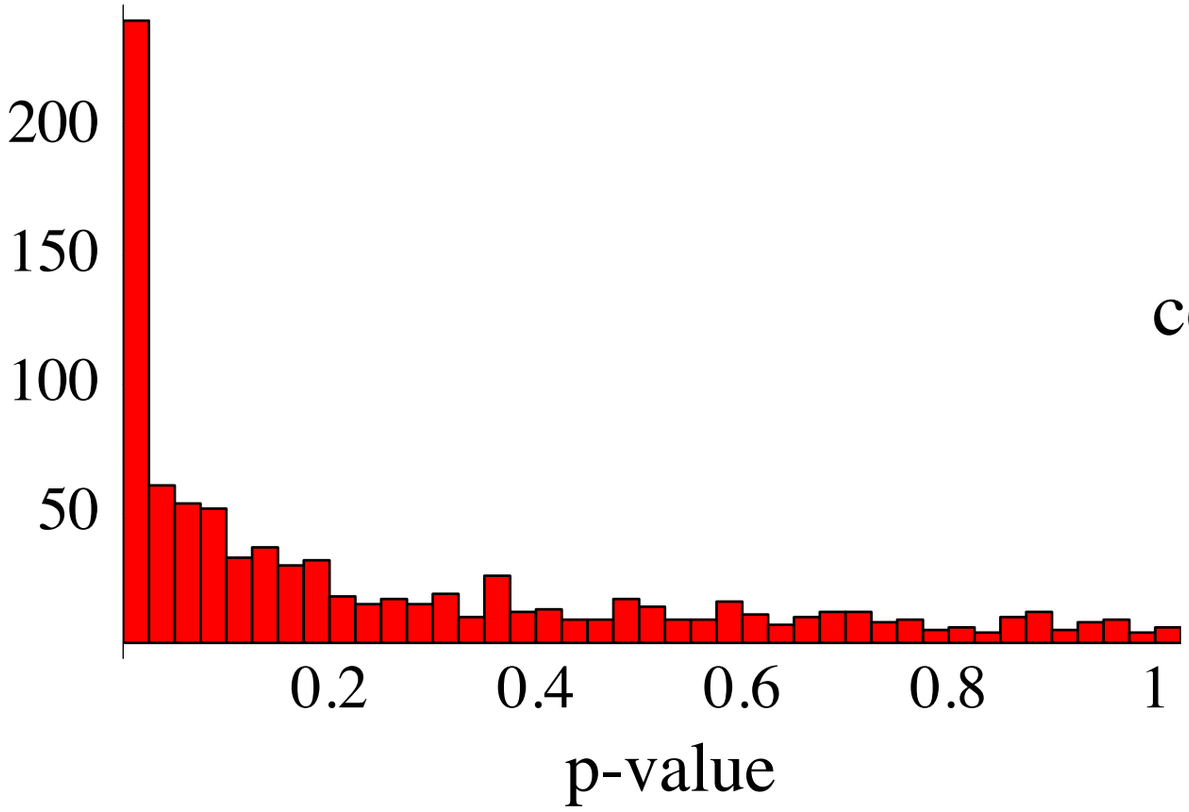
...except when it doesn't...

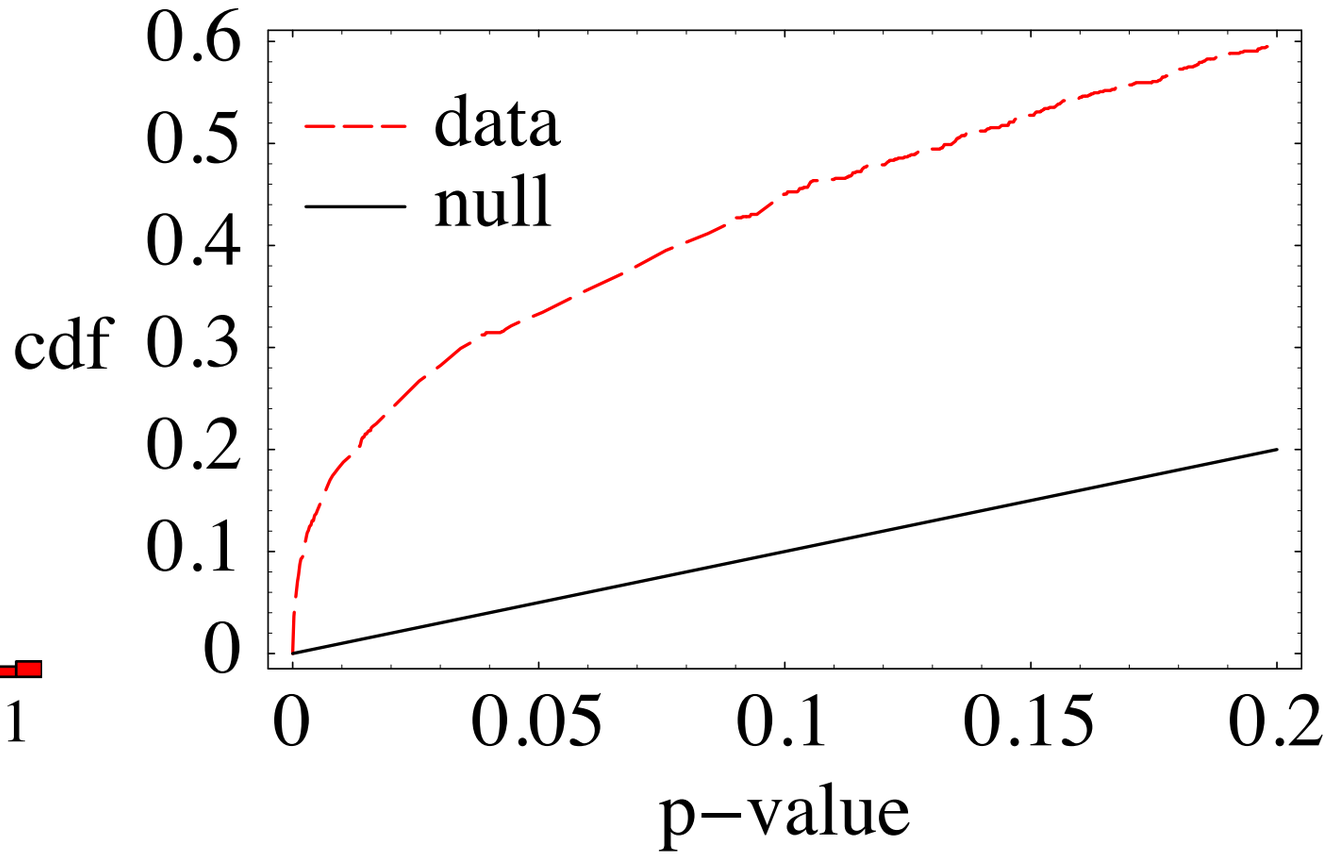


...and then there is the matter of confounding...



...because of "population structure"!





Alleles are not independent!

Dealing with "population structure"

- The vast excess of highly significant associations is due to the genetic background
- We handled this using the same mixed model used for human height above:

$$Y = X\beta + u + \epsilon, \; u \sim N(0, \sigma_g K), \; \epsilon \sim N(0, \sigma_e I)$$

We understand this now...

COMMENT

The nature of confounding in genome-wide association studies

Bjarni J. Vilhjálmsson[1,2] and Magnus Nordborg[3,4]

The authors argue that population structure per se is not a problem in genome-wide association studies — the true sources are the environment and the genetic background, and the latter is greatly underappreciated. They conclude that mixed models effectively address this issue.

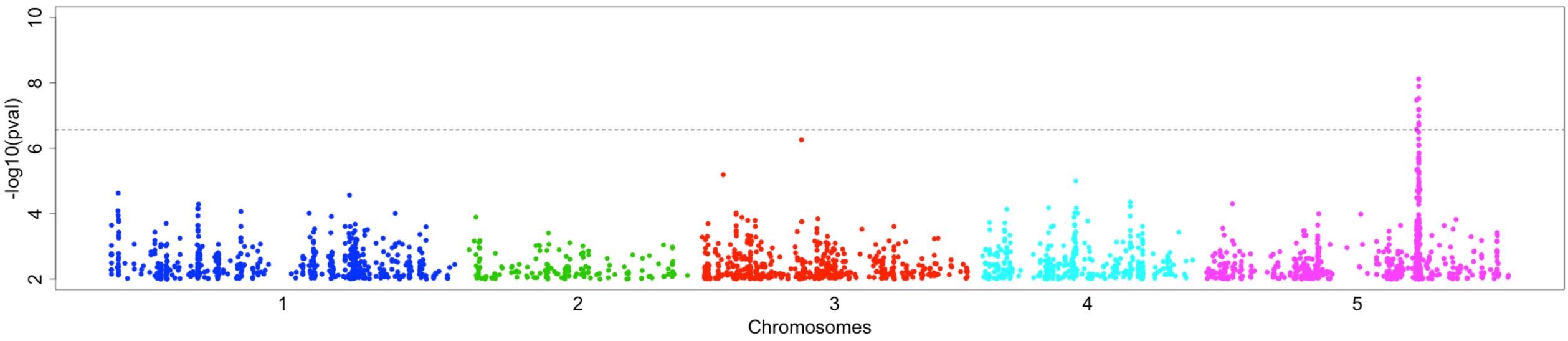
> “population structure is not the fundamental source of the problem, and removing it is not the solution”

Thanks to dramatically decreasing genotyping and sequencing costs, genome-wide association studies (GWASs) are becoming the default method for studying the genetics of natural variation. The increasing number and diversity of GWASs will require appropriate statistical analysis methods. The most basic problem is assessing the significance of an association in the light of confounding effects that may cause spurious associations.

The aspect of this problem that has received the most attention is the danger of false positives in structured populations. If the study population is a mixture of populations that differ with respect to allele frequencies as well as the trait of interest, spurious correlations in ‘unrelated’ individuals. Variation in relatedness is a basic property of natural populations, as is correlation between causative loci. This issue is familiar to quantitative geneticists[5] but has not been widely appreciated in other fields. It is important for GWASs and will become crucial as sample sizes increase.

To demonstrate this, let us return to the chopstick example but fast-forward to the era of millions of SNPs. Genetic differentiation between East Asians and other populations means that vast numbers of markers in addition to *HLA-A1* would be associated with chopstick skill. These markers would also be correlated with *HLA-A1*, with each other and with any trait (genetic or not) that

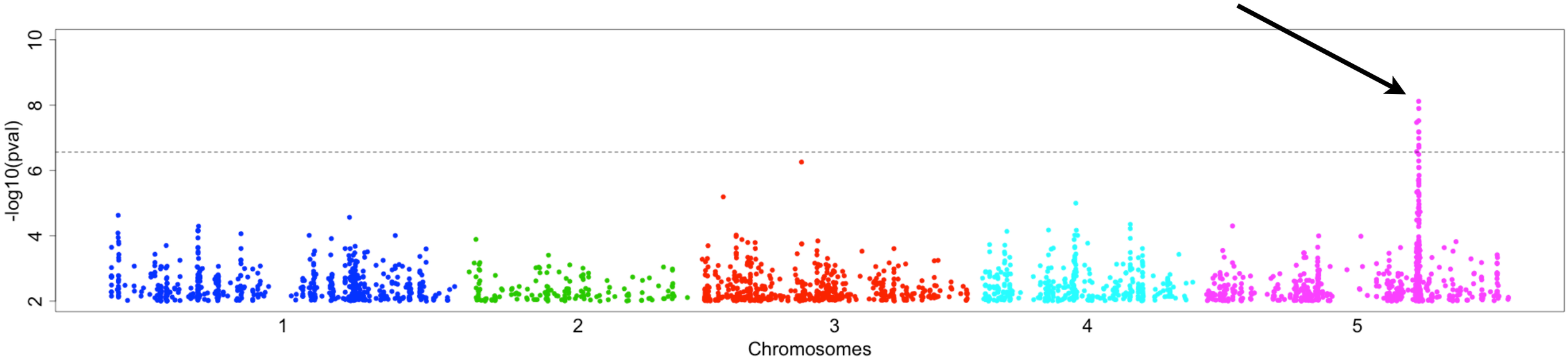
The problem of fine mapping



GWAS for germination rate after 21 days storage in Swedish lines

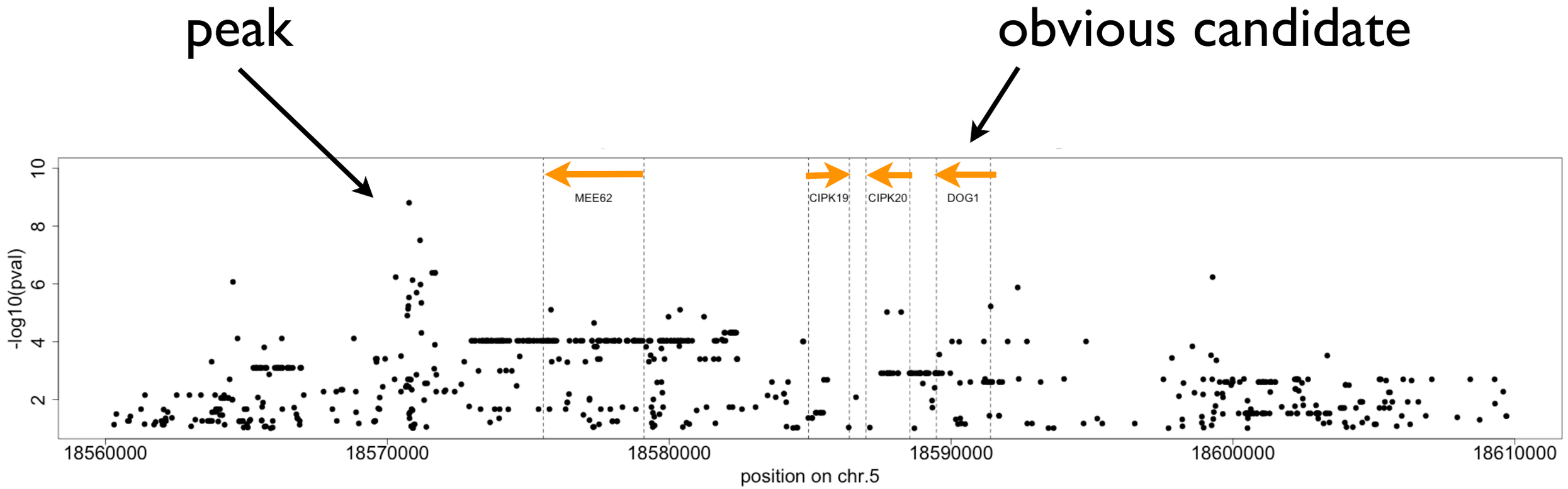
The problem of fine mapping

Peak includes *Delay Of Germination 1*

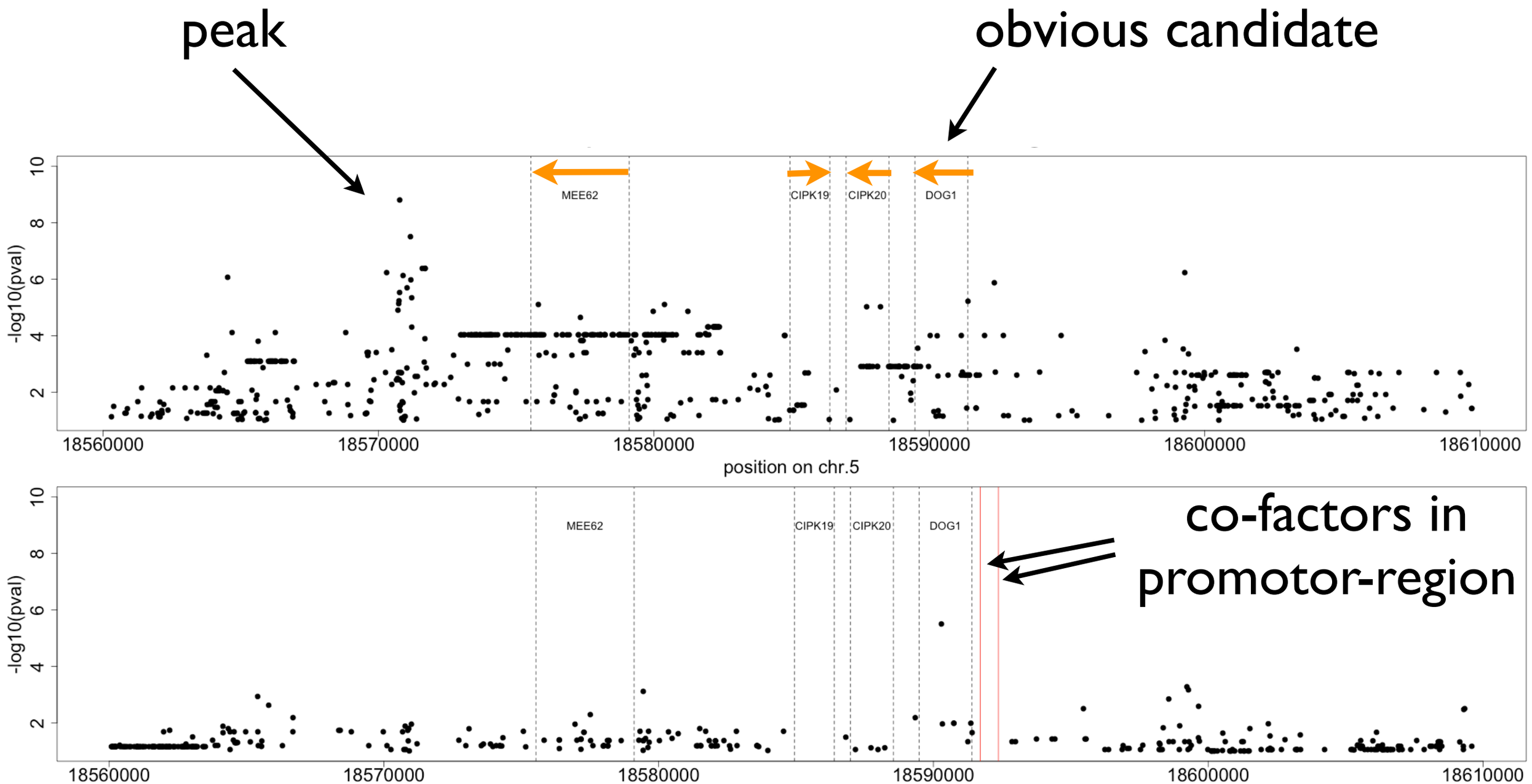


GWAS for germination rate after 21 days storage in Swedish lines

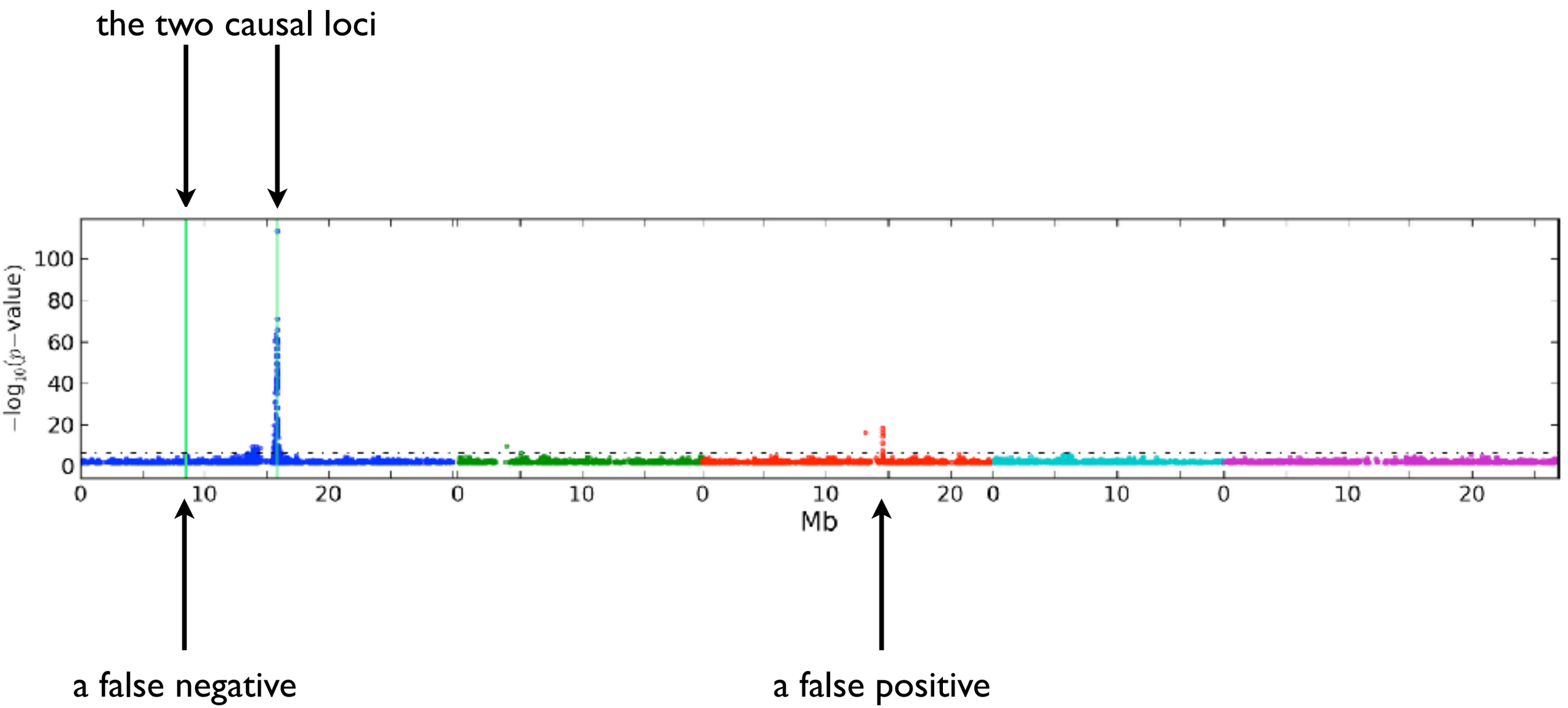
..but peak is far from the obvious candidate...



Including two co-factors eliminates spurious peak



A simulation example…



Some obvious extensions

- GWAS with multiple loci (stepwise regression [Segura *et al.*, *Nature Genet.* 2012])
- Fully Bayesian estimation of the additive effect of all polymorphism in the genome (Erbe *et al.*, *J. Dairy Sci.* 2012)

Nature Reviews Genetics | AOP, published online 30 March 2010; doi:10.1038/nrg2788

COMPLEX TRAITS

Plants are not humans

Advances in genomics have made it possible to use genome-wide association (GWA) studies to explore the genetic basis of common traits in humans, leading to advances in biological knowledge and technical expertise. The power of the GWA approach has now been extended successfully to *Arabidopsis thaliana*, in which results are proving to be strikingly different from those obtained in humans.

It is hoped that GWA studies in plants will shed light on the natural genetic variation that underlies adaptive and commercially important traits. Plant geneticists are aided in this task by the existence of inbred lines of *A. thaliana*, which can be phenotyped easily for many traits in different environments. Atwell *et al.* conducted their GWA study on almost 200 such lines by analysing associations between a genome-wide set of 250,000 SNPs and 107 traits across four categories: flowering, development, defence and element concentration.

The first striking finding was that, compared with humans, *A. thaliana* has a high degree of population structure (see also Further Reading and Websites below). This could potentially lead to a problematic number of spurious associations; however, the authors show that some of these false positives can be eliminated by using a mixed-model approach (which can correct for fine-scale structure). For example, previously known variants were easily identified — including those for disease resistance and for variation in sodium levels. However, association peaks were not always very sharp. In one case, a peak spanned 500 kb (encompassing 100 genes), showing that complex genetic architecture involving multiple correlated causative variants can be a further confounding factor — one that has previously not been recognized in humans.

Second, there is a higher probability than would be expected by chance that the SNPs that are strongly associated will include candidate SNPs (SNPs that are expected to be connected to a trait based on functional knowledge).

A third difference is that the number of samples used in the plant GWA study was very low; the study used a maximum of 192 lines, whereas most human GWA studies use thousands of individuals. This low sample size would seem to preclude identifying any association at all were it not for two additional advantages offered by *A. thaliana*: the existence of intermediate-frequency alleles of large effect (explaining up to 20% of the variation for several traits, compared with the preponderance of minor-effect alleles in humans) and the ability to control for environmental noise in the inbred lines. Both of these features 'amplify' the association signal, allowing true associations to stand out.

The GWA approach has some intrinsic limitations — for example, the ability only to identify alleles that are common. However, falling genotyping and sequencing costs and a growing list of new phenotypes means that such studies should have a bright future for characterizing the natural genetic variants that underlie quantitative traits in *A. thaliana* and other organisms.

Tanita Casci

ORIGINAL RESEARCH PAPER Atwell, S., Huang, Y. S., Vilhjálmsson, B. J. & Willems, G. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 24 Mar 2010 (doi:10.1038/nature08800)
FURTHER READING Platt, A. *et al.* The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet.* **6**, e1000843 (2010)
WEBSITES
1001 Genomes Project: http://www.1001genomes.org
Genome-wide association studies in *Arabidopsis thaliana*: http://arabidopsis.usc.edu

Nature Reviews Genetics | AOP, published online 30 March 2010; doi:10.1038/nrg2788

COMPLEX TRAITS

Adaptation is different from disease

Advances in genomics have made it possible to use genome-wide association (GWA) studies to explore the genetic basis of common traits in humans, leading to advances in biological knowledge and technical expertise. The power of the GWA approach has now been extended successfully to *Arabidopsis thaliana*, in which results are proving to be strikingly different from those obtained in humans.

It is hoped that GWA studies in plants will shed light on the natural genetic variation that underlies adaptive and commercially important traits. Plant geneticists are aided in this task by the existence of inbred lines of *A. thaliana*, which can be phenotyped easily for many traits in different environments. Atwell *et al.* conducted their GWA study on almost 200 such lines by analysing associations between a genome-wide set of 250,000 SNPs and 107 traits across four categories: flowering, development, defence and element concentration.

The first striking finding was that, compared with humans, *A. thaliana* has a high degree of population structure (see also Further Reading and Websites below). This could potentially lead to a problematic number of spurious associations; however, the authors show that some of these false positives can be eliminated by using a mixed-model approach (which can correct for fine-scale structure). For example, previously known variants were easily identified — including those for disease resistance and for variation in sodium levels. However, association peaks were not always very sharp. In one case, a peak spanned 500 kb (encompassing 100 genes), showing that complex genetic architecture involving multiple correlated causative variants can be a further confounding factor — one that has previously not been recognized in humans.

Second, there is a higher probability than would be expected by chance that the SNPs that are strongly associated will include candidate SNPs (SNPs that are expected to be connected to a trait based on functional knowledge).

A third difference is that the number of samples used in the plant GWA study was very low; the study used a maximum of 192 lines, whereas most human GWA studies use thousands of individuals. This low sample size would seem to preclude identifying any association at all were it not for two additional advantages offered by *A. thaliana*: the existence of intermediate-frequency alleles of large effect (explaining up to 20% of the variation for several traits, compared with the preponderance of minor-effect alleles in humans) and the ability to control for environmental noise in the inbred lines. Both of these features 'amplify' the association signal, allowing true associations to stand out.

The GWA approach has some intrinsic limitations — for example, the ability only to identify alleles that are common. However, falling genotyping and sequencing costs and a growing list of new phenotypes means that such studies should have a bright future for characterizing the natural genetic variants that underlie quantitative traits in *A. thaliana* and other organisms.

Tanita Casci

ORIGINAL RESEARCH PAPER Atwell, S., Huang, Y. S., Vilhjálmsson, B. J. & Willems, G. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 24 Mar 2010 (doi:10.1038/nature08800)
FURTHER READING Platt, A. *et al.* The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet.* **6**, e1000843 (2010)
WEBSITES
1001 Genomes Project: http://www.1001genomes.org
Genome-wide association studies in *Arabidopsis thaliana*: http://arabidopsis.usc.edu

What about skin color?

OPEN ACCESS Freely available online

PLOS | GENETICS

Genetic Architecture of Skin and Eye Color in an African-European Admixed Population

Sandra Beleza[1,2]*, Nicholas A. Johnson[3], Sophie I. Candille[1], Devin M. Absher[4], Marc A. Coram[5], Jailson Lopes[2,6,7], Joana Campos[2], Isabel Inês Araújo[7], Tovi M. Anderson[1], Bjarni J. Vilhjálmsson[8], Magnus Nordborg[8], António Correia e Silva[7], Mark D. Shriver[9], Jorge Rocha[2,6,10], Gregory S. Barsh[1,4]*, Hua Tang[1]

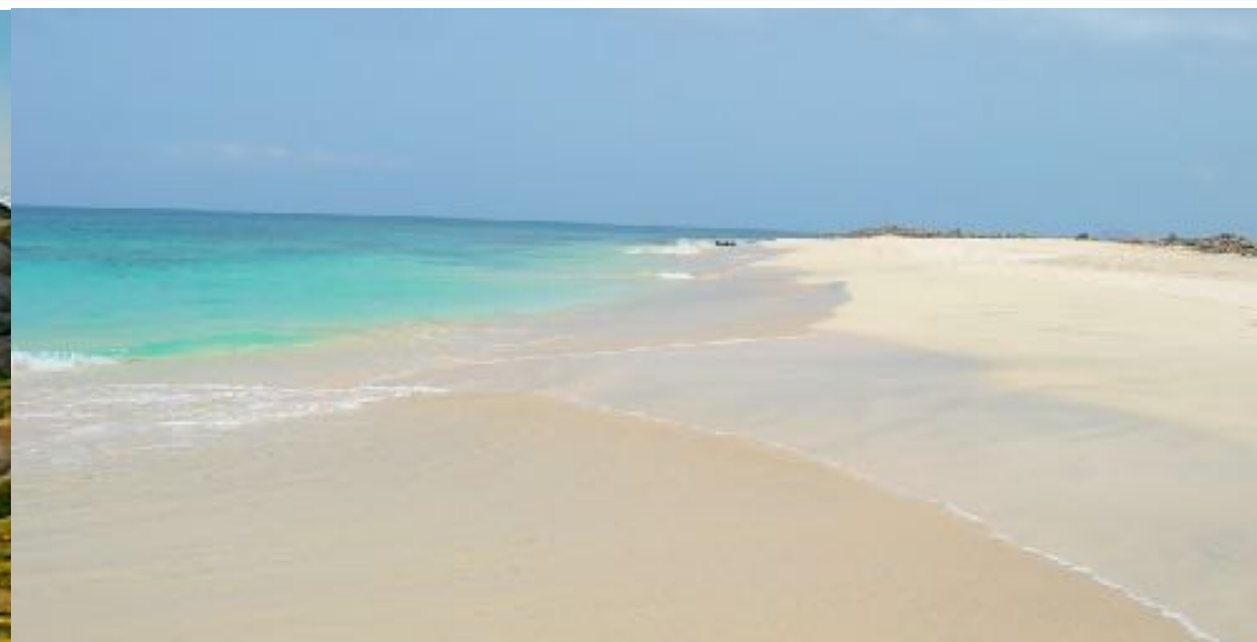
1 Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America, **2** Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal, **3** Department of Statistics, Stanford University, Stanford, California, United States of America, **4** HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, United States of America, **5** Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California, United States of America, **6** Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO), Vairão, Portugal, **7** Universidade de Cabo Verde (Uni-CV), Praia, Santiago, Cabo Verde, **8** Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria, **9** Department of Anthropology, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **10** Departamento de Biologia, Faculdade de Ciências da Universidade do Porto, Porto, Portugal

Abstract

Variation in human skin and eye color is substantial and especially apparent in admixed populations, yet the underlying genetic architecture is poorly understood because most genome-wide studies are based on individuals of European ancestry. We study pigmentary variation in 699 individuals from Cape Verde, where extensive West African/European admixture has given rise to a broad range in trait values and genomic ancestry proportions. We develop and apply a new approach for measuring eye color, and identify two major loci (*HERC2*[*OCA2*] $P = 2.3 \times 10^{-62}$, *SLC24A5* $P = 9.6 \times 10^{-9}$) that account for both blue versus brown eye color and varying intensities of brown eye color. We identify four major loci (*SLC24A5* $P = 5.4 \times 10^{-27}$, *TYR* $P = 1.1 \times 10^{-9}$, *APBA2*[OCA2] $P = 1.5 \times 10^{-8}$, *SLC45A2* $P = 6 \times 10^{-9}$) for skin color that together account for 35% of the total variance, but the genetic component with the largest effect (~44%) is average genomic ancestry. Our results suggest that adjacent cis-acting regulatory loci for *OCA2* explain the relationship between skin and eye color, and point to an underlying genetic architecture in which several genes of moderate effect act together with many

Cabo Verde

Colonized by Portuguese 1460
Major center for African slave trade
Independence from Portugal 1975



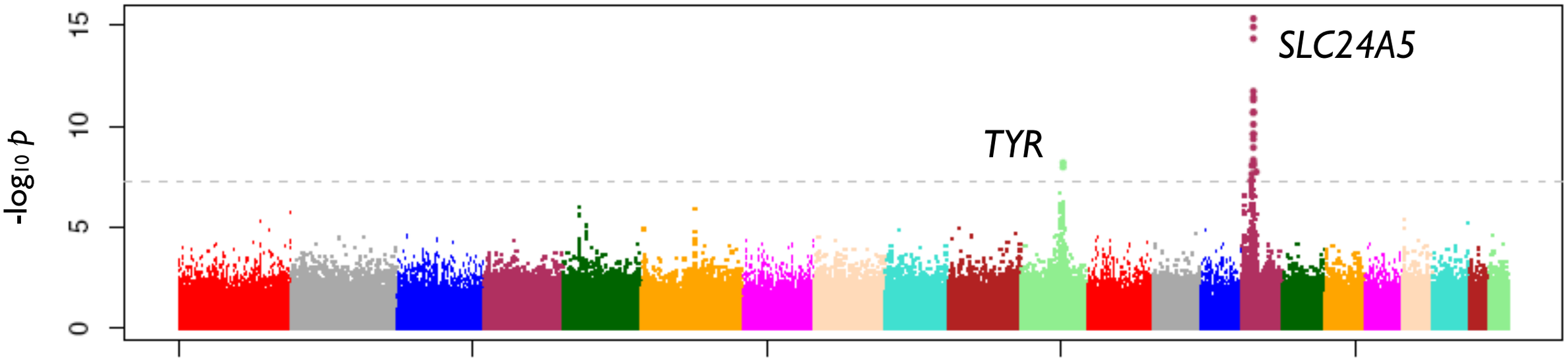




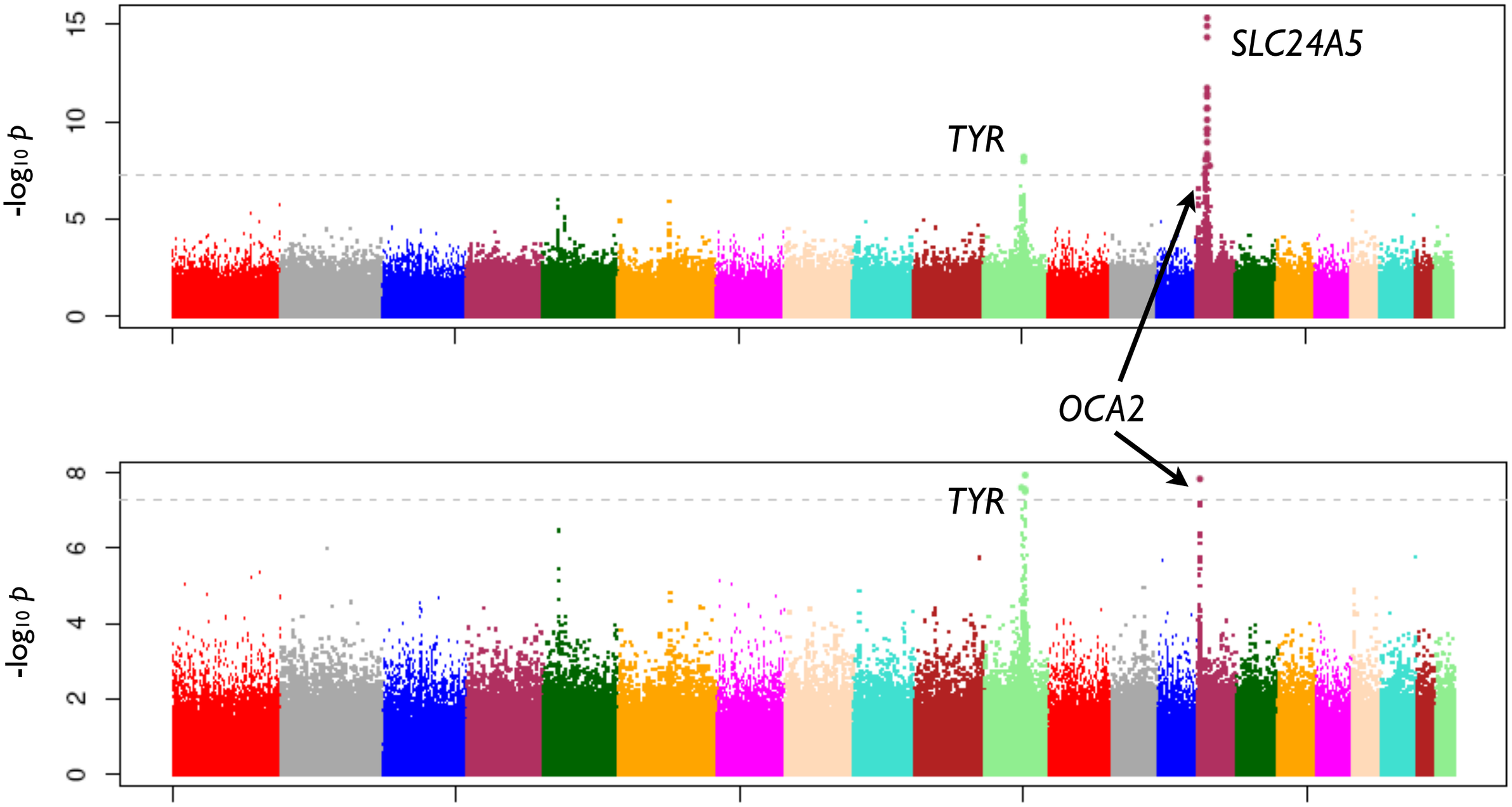





GWAS of skin reflectance



GWAS of skin reflectance



Genome-wide markers explain much more