GWAS and Quantitative genomics

Workshop on Polygenic Adaption
ICTS, Bangalore
6 – 17 May 2024

Bruce Walsh
University of Arizona
jbwalsh@arizona.edu

Overview

- GWAS methodology
- Overview of GWAS results
- More on LD (D' vs r^2)
  - Common marker-rare causal allele mismatch
- eSNPs and regSNP
  - Cis vs trans
  - Other regSNPs
  - Mediation
- GWAS, Architectures, and selection

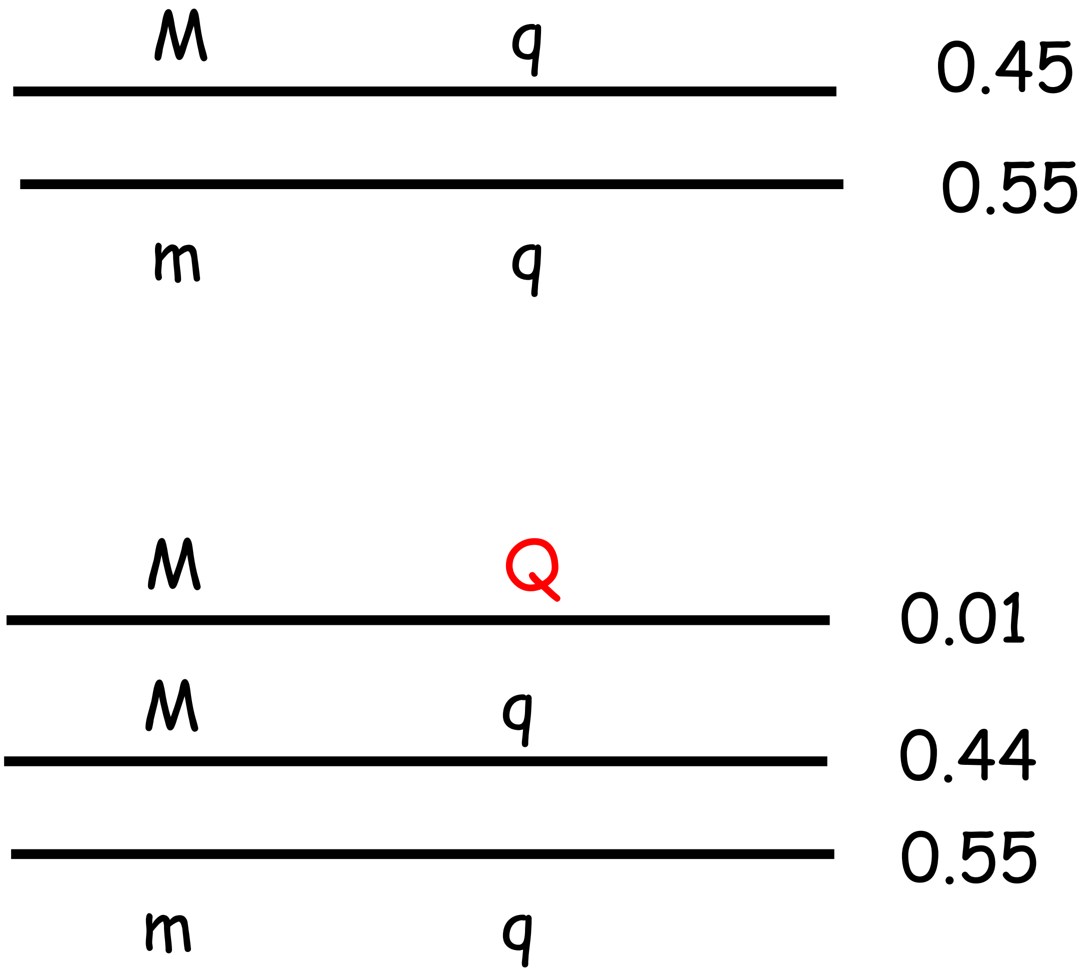
Background, Additional reading

- WVL (Walsh, Visscher, Lynch) 2024.
  - Chapter 20: GWAS
  - Chapter 21: Quantitative Genomics
  - Appendix 2: Mediation analysis

Part I:
Association Mapping

- Association mapping uses a set of very dense markers in a set of (largely) unrelated individuals
- Requires population level LD
- Allows for very fine mapping (1-20 kB)

Mutation generates LD



Mutation Q only found on M-bearing chromosomes (haplotypes)

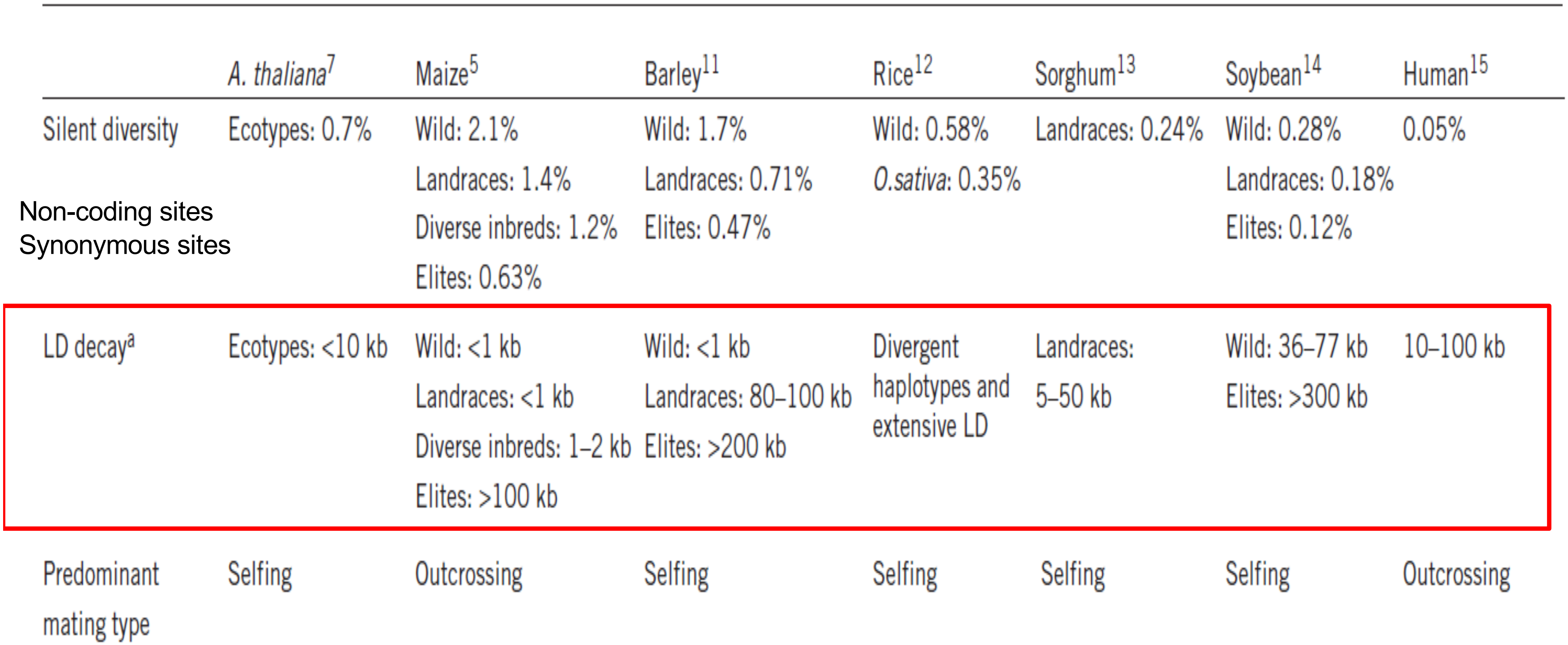
LD: Linkage disequilibrium

D(AB) = freq(AB) - freq(A)*freq(B).
LD = 0 if A and B are independent. If LD not zero, correlation between A and B in the population

If a marker and QTL are linked, then the marker and QTL alleles are in LD in close relatives, generating a marker-trait association.

The decay of D: $r^2(t) = (1\text{-}c)^t\, r^2(0)$
here c is the recombination rate. Tightly-linked genes (small c) initially in LD can retain LD for long periods of time

LD range in major crops



	A. thaliana[7]	Maize[5]	Barley[11]	Rice[12]	Sorghum[13]	Soybean[14]	Human[15]
Silent diversity	Ecotypes: 0.7%	Wild: 2.1% Landraces: 1.4% Diverse inbreds: 1.2% Elites: 0.63%	Wild: 1.7% Landraces: 0.71% Elites: 0.47%	Wild: 0.58% *O.sativa*: 0.35%	Landraces: 0.24%	Wild: 0.28% Landraces: 0.18% Elites: 0.12%	0.05%
LD decay[a]	Ecotypes: <10 kb	Wild: <1 kb Landraces: <1 kb Diverse inbreds: 1–2 kb Elites: >100 kb	Wild: <1 kb Landraces: 80–100 kb Elites: >200 kb	Divergent haplotypes and extensive LD	Landraces: 5–50 kb	Wild: 36–77 kb Elites: >300 kb	10–100 kb
Predominant mating type	Selfing	Outcrossing	Selfing	Selfing	Selfing	Selfing	Outcrossing

Non-coding sites
Synonymous sites

Buckler and Gore 2007. Nat. Genet. 39:1056-1057

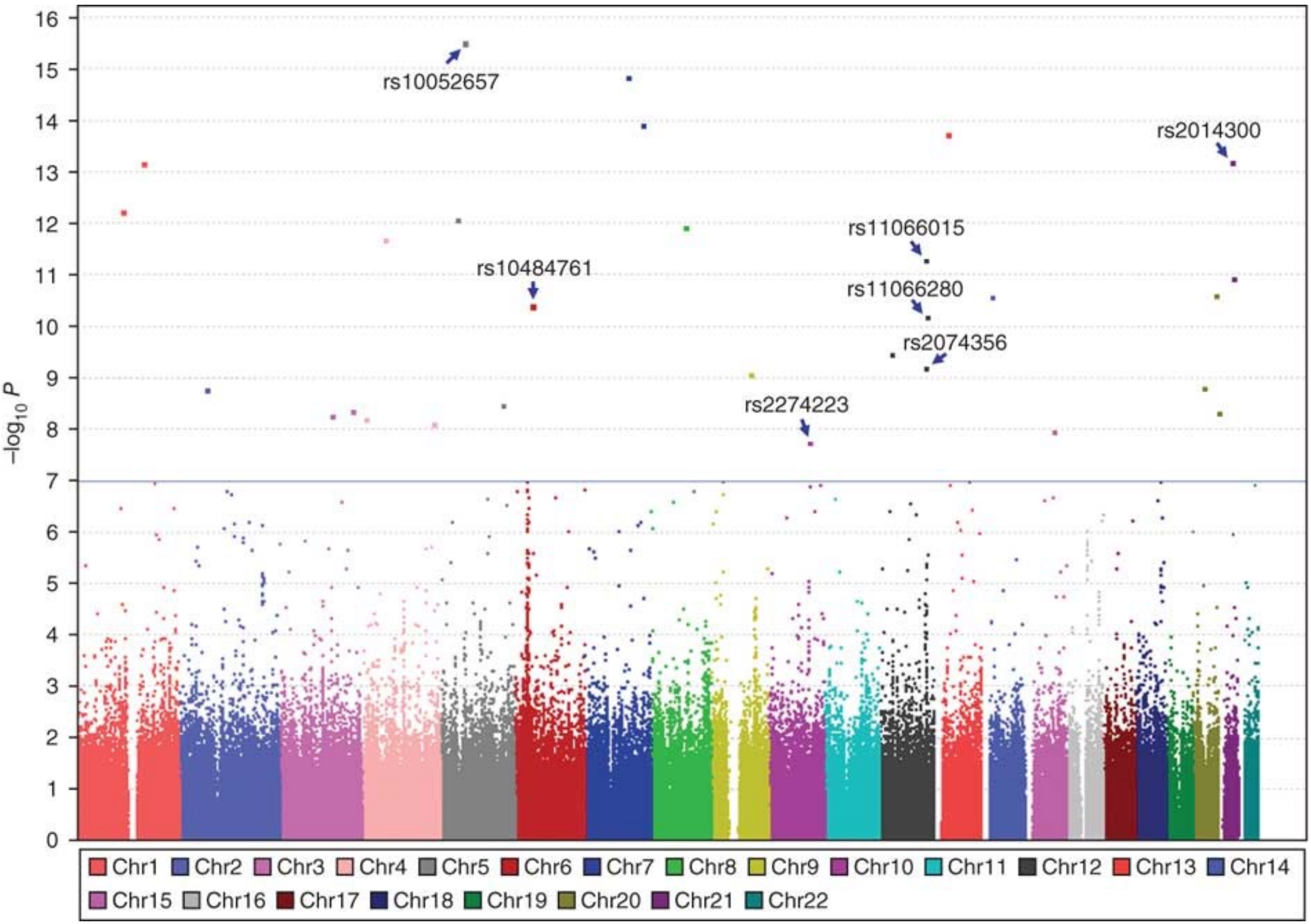
Association mapping

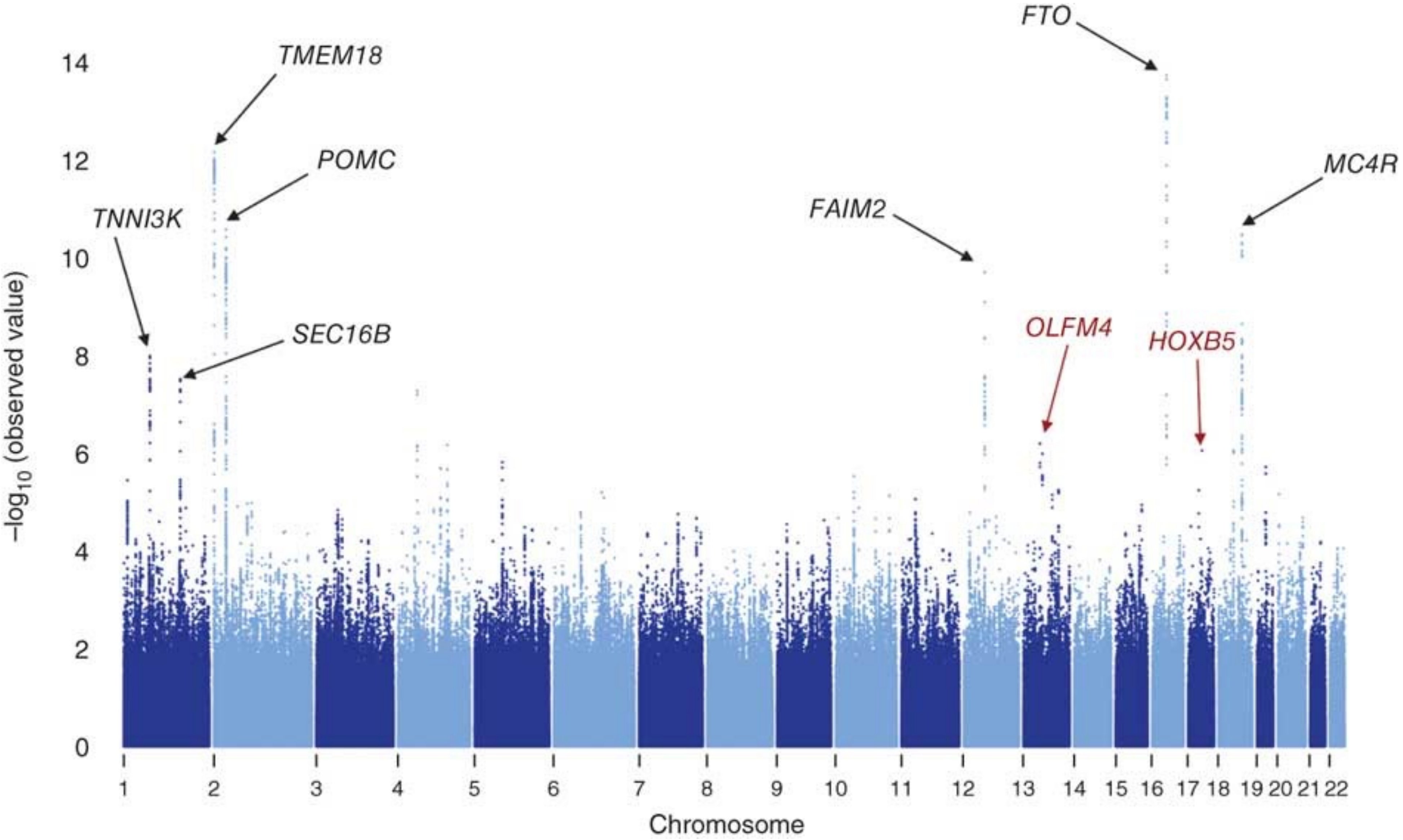
- Marker-trait associations within a population of unrelated individuals
- Very high marker density (~ 100s of markers/cM) required
  - Marker density no less than the average track length of linkage disequilibrium (LD)
- Relies on very slow breakdown of initial LD generated by a new mutation near a marker to generate marker-trait associations
  - LD decays very quickly unless very tight linkage
  - Hence, resolution on the scale of LD in the population(s) being studied (1 ~ 40 kB)
- Widely used since early 2000's. Mainstay of human genetics, strong inroads in breeding, evolutionary genetics
- Power a function of the genetic variance of a QTL, not its mean effects, $\mathrm{Var(m)} = r^2\ \mathrm{Var(QTL)} = r^2\ 2a^2p(1\text{-}p)$

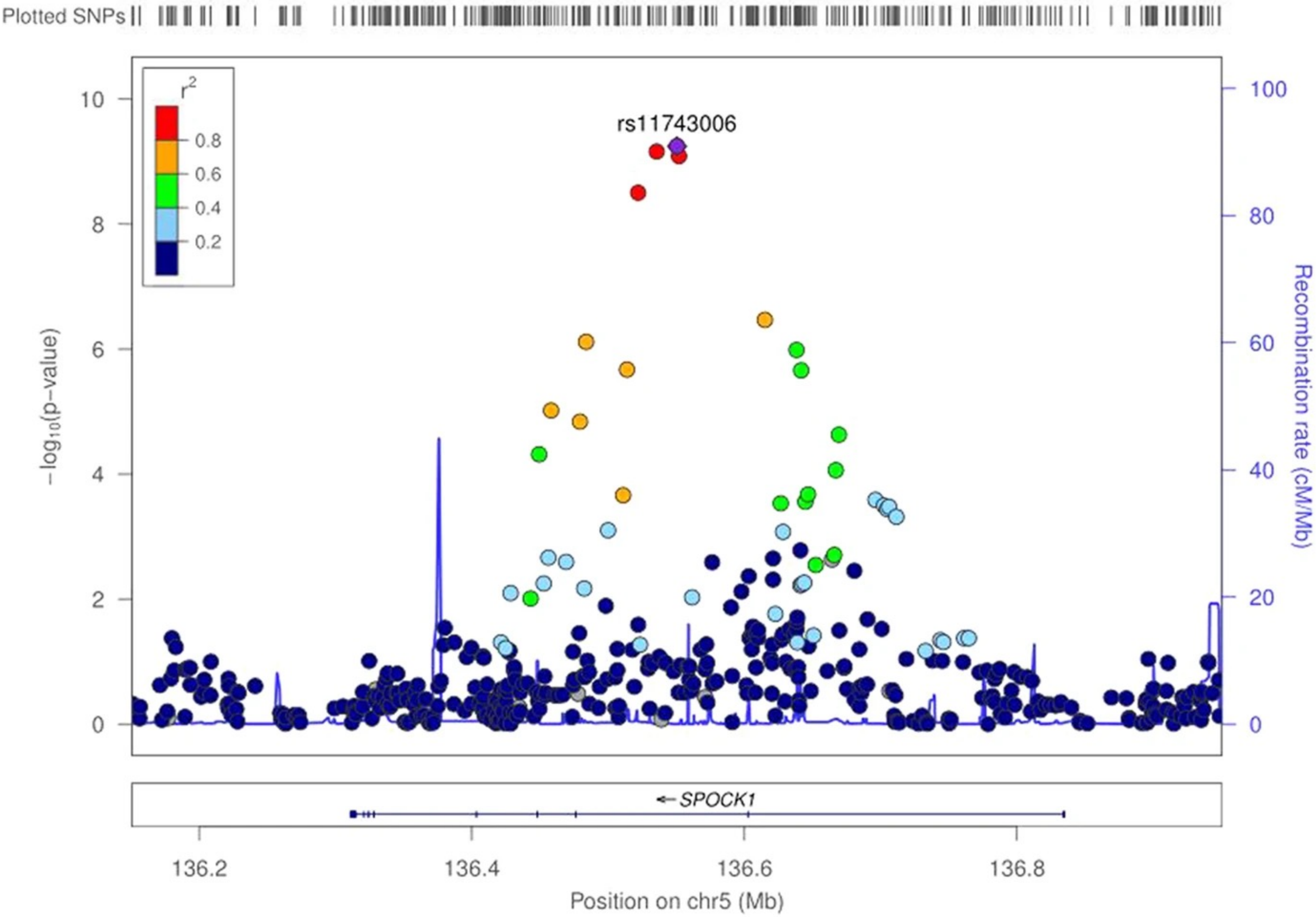
Manhattan plots

- The results for a Genome-wide Association study (or GWAS) are typically displayed using a Manhattan plot.
  - At each SNP, -ln(p), the negative log of the p value for a significant marker-trait association is plotted. Values above a threshold indicate significant effects
  - Threshold set by Bonferroni-style multiple comparisons correction
  - With n markers, an overall false-positive rate of p requires each marker be tested using p/n.
  - With n = 10^6 SNPs, p must exceed 0.01/10^6 or 10^{-8} to have a control of 1% of a false-positive

PRO
KOS


16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
$-\log_{10} P$
rs10052657
rs2014300
rs11066015
rs10484761
rs11066280
rs2074356
rs2274223
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22


FTO
TMEM18
POMC
TNNI3K
MC4R
FAIM2
SEC16B
OLFM4
HOXB5
14
12
10
8
6
4
2
0
–log10 (observed value)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome


Plotted SNPs
r²
0.8
0.6
0.4
0.2
rs11743006
−log₁₀(p−value)
0
2
4
6
8
10
Recombination rate (cM/Mb)
0
20
40
60
80
100
SPOCK1
136.2
136.4
136.6
136.8
Position on chr5 (Mb)

Linkage vs. Association

The distinction between linkage and association is subtle, yet critical

Marker allele M is associated with the trait if Cov(M,y) is not 0

While such associations can arise via linkage, they can also arise via population structure.

Thus, association DOES NOT imply linkage, and linkage is not sufficient for association

When population being sampled actually consists of several distinct subpopulations we have lumped together, marker alleles may provide information as to which group an individual belongs. If there are other risk factors in a group, this can create a false association btw marker and trait

Example. The Gm marker was thought (for biological reasons) to be an excellent candidate gene for diabetes in the high-risk population of Pima Indians in the American Southwest. Initially a very strong association was observed:

Gm^+	Total	% with diabetes
Present	293	8%
Absent	4,627	29%

Gm^+	Total	% with diabetes
Present	293	8%
Absent	4,627	29%

Problem: freq(Gm^+) in Caucasians (lower-risk diabetes Population) is 67%, Gm^+ rare in full-blooded Pima

The association was re-examined in a population of Pima that were 7/8th (or more) full heritage:

Gm^+	Total	% with diabetes
Present	17	59%
Absent	1,764	60%

Indicator (0 / 1) Variable
for SNP genotype k. Typically
k = 3, i.e. AA, Aa aa

$$y = \mu + \sum_{k=1}^{n} \beta_k \, M_k + \sum_{j=1}^{m} \gamma_j \, b_j + e$$

Significant β indicates
marker-trait association

m unlinked markers that
vary across subpopulations.
b_j = marker genotype indicator
variable

SNP marker
under consideration

Variations on this theme (eigenstrat) --- use all of the marker information to extract a set of significant PCs, which are then included in the model as cofactors

Structure plus Kinship Methods

Association mapping in plants offer occurs by first taking a large collection of lines, some closely related, others more distantly related. Thus, in addition to this collection being a series of subpopulations (derivatives from a number of founding lines), there can also be additional structure within each subpopulation (groups of more closely related lines within any particular lineage).

$$Y = X\beta + Sa + Qv + Zu + e$$

Fixed effects in blue, random effects in red

This is a mixed-model approach. The program TASSEL runs this model.

Q-K method

$$Y = X\beta + Sa + Qv + Zu + e$$

β = vector of fixed effects

a = SNP effects

v = vector of subpopulation effects (STRUCTURE)
Q_{ij} = Prob(individual i in group j). Determined from STRUCTURE output

u = shared polygenic effects due to kinship.
Cov(u) = var(A)*A, where the relationship matrix A estimated from marker data matrix K, also called a GRM – a genomic relationship matrix

Which markers to include in K?

- Best approach is to leave out the marker being tested (and any in LD with it) when construction the genomic relationship matrix
  - LOCO approach – leave out one chromosome (which the tested marker is linked to)
- Best approach seems to be to use most of the markers
- Other mixed-model approaches along these lines

Part II:
Overview of GWAS data

Take-home message from GWAS

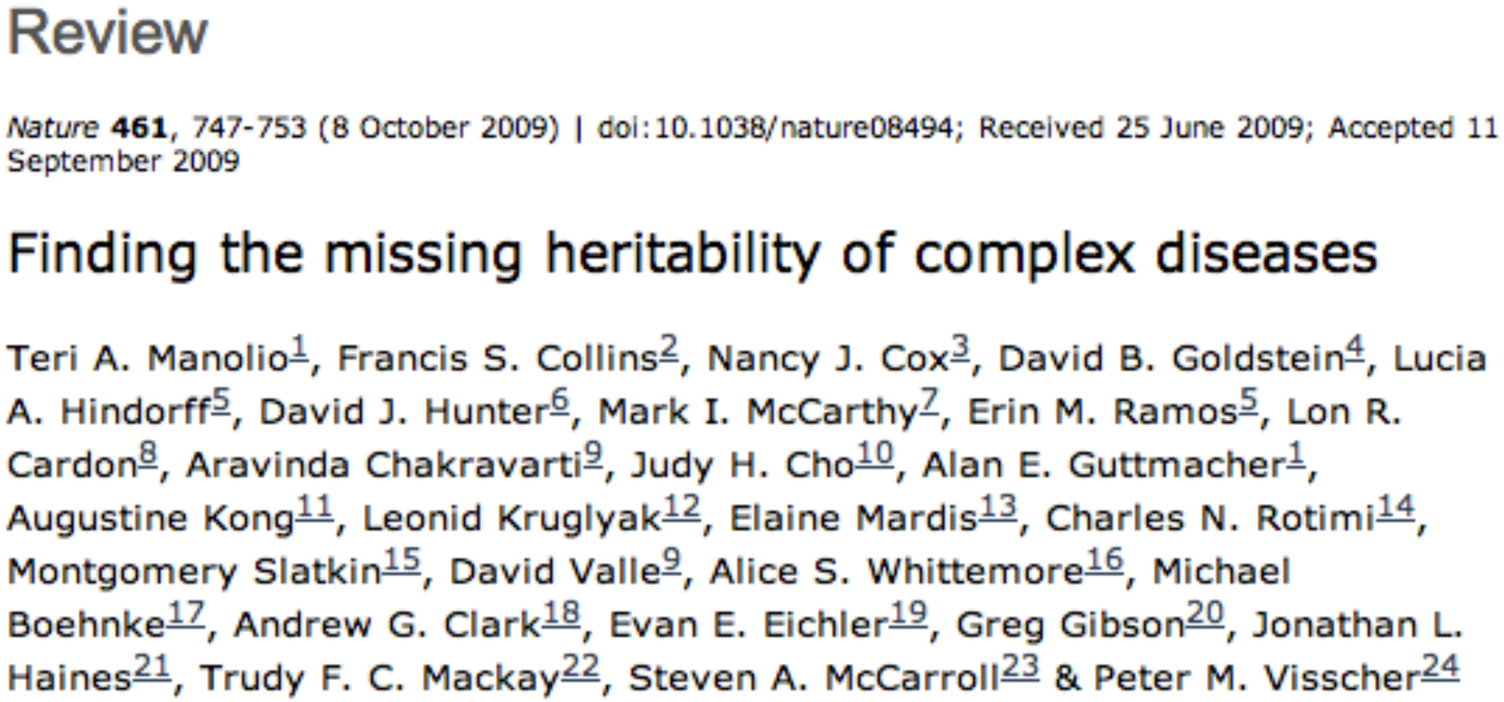
- The marker-effect in a GWAS is the amount of variation that a marker explains (effect weighted by frequency), $2a^2p(1-p)$
  - A marker tagging a causative site with a low frequency major effect allele gives a small marker variance
  - Poor power to detect all but major alleles when they are rare
  - Inverse relationship between frequency and effect size
- Is the bulk of genetic variation from
  - "common" alleles ($p > 5\%$) of small effect
  - Rare alleles ($p < 1\%$) of large effect

Take-home message from early GWAS (pre-2015):

- Many sites, each of small effect
  - Infinitesimal-like result (from a marker-variance standpoint)
  - Large-effect (variance) markers are very rare
    - DOES NOT imply alleles of large effect (a >> 1) are absent, rather just at very low frequency
  - Inverse relationship between allelic effect size and frequency
- > 80% of Hits are in noncoding regions
  - Importance of regulatory mutations

Problems with early GWAS

- Initially, human geneticists (and others) were thrilled with the ability of GWAS to localize chromosomal regions contributing to genetic variation to very small (kilobase) regions
- Found essentially all had very small effects (variance explained by a marker)
- Further problem: the sum of such detected markers fell far short of accounting for the known genetic variation



Review

Nature **461**, 747-753 (8 October 2009) | doi:10.1038/nature08494; Received 25 June 2009; Accepted 11 September 2009

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]

A number of GWAS workers noted that the sum of their significant SNP marker variances was much less (typically 10%) than the additive variance estimated from biometrical (relative-based) methods. Where is this “missing” heritability? Does this suggest a fundamental flaw in quantitative genetics?

Part III:
Measures of LD

$$\begin{aligned} r^2_{A_iB_j}(t) &= \frac{[D_{A_iB_j}(t)]^2}{p_{A_i}p_{B_j}(1-p_{A_i})(1-p_{B_j})} = \frac{(1-c)^{2t}\,[D_{A_iB_j}(0)]^2}{p_{A_i}p_{B_j}(1-p_{A_i})(1-p_{B_j})} \\ &= (1-c)^{2t}\,r^2_{A_iB_j}(0) \end{aligned} \tag{5.13d}$$

Example 5.5. Because the maximum value of D varies over allele frequencies, Lewontin (1964) proposed a standardized measure

$$D' = \frac{D}{\text{max}|D|}, \qquad \text{where} \quad -1 \leq D' \leq 1 \tag{5.15a}$$

with

$$\text{max}|D| = \begin{cases} \text{min}[\,p(1-q), (1-p)q\,] & \text{for } D \geq 0 \\ \text{max}[-pq, -(1-p)(1-q)] & \text{for } D < 0 \end{cases} \tag{5.15b}$$

To see how this metric differs from r^2, consider the two following situations:

	Case One			Case Two	
	M	m		M	m
Q	20	0	Q	20	0
q	4	6	q	0	10

In case one, $\widehat{p}_M$ = 24/30 = 8/10, $\widehat{p}_Q$ = 20/30 = 2/3, and $\widehat{p}_{MQ}$ = 20/30 = 2/3, yielding

$$\widehat{D}_{MQ} = \widehat{p}_{MQ} - \widehat{p}_M\, \widehat{p}_Q = 2/3 - (8/10)(2/3) = 2/15 = 0.133$$

$$\widehat{r^2}_{MQ} = \frac{(\widehat{D}_{MQ})^2}{\widehat{p}_M(1-\widehat{p}_M)\,\widehat{p}_Q(1-\widehat{p}_Q)} = \frac{(2/15)^2}{(8/10)(2/10)(2/3)(1/3)} = 1/2$$

Because $\widehat{p}_M(1-\widehat{p}_Q) = 4/15$ and $(1-\widehat{p}_M)\,\widehat{p}_Q = 2/15$, max$|D|$ = 2/15 (Equation 5.15b) and

$$D' = \frac{2/15}{2/15} = 1$$

Hence, while D' might suggest complete disequilibrium, its r^2 value is only 0.5. Note that while Q is always found on M-bearing chromosomes, some M-bearing chromosomes instead contain q. Hence, while Q is always associated with M, the converse is not true. The former gives $D' = 1$, while the latter implies that $r^2 < 1$.

Case One

	M	m
Q	20	0
q	4	6

Case Two

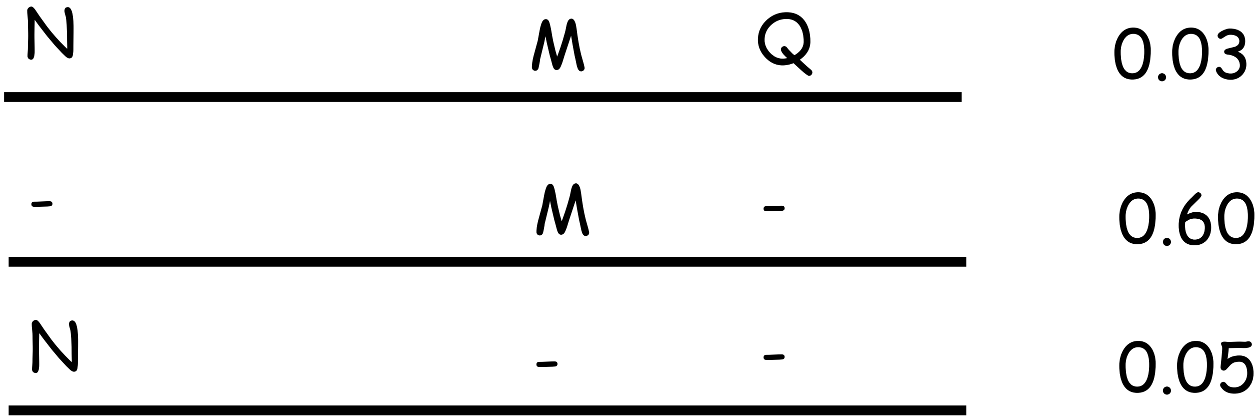
	M	m
Q	20	0
q	0	10

For case two, $\widehat{p}_M = \widehat{p}_Q = \widehat{p}_{MQ}$ = 2/3, yielding $D_{MQ} = 2/9 = 0.22$ and $\widehat{r^2}_{MQ} = 1$. Further, $\widehat{p}_M(1 - \widehat{p}_Q) = (1 - \widehat{p}_M)\,\widehat{p}_Q = 2/9$, giving $D' = 1$. In general, $r^2 = 1$ implies that $|D'| = 1$, but *the converse is not true*. In terms of the 2×2 contingency table, one zero off-diagonal element (one allele is *always* found on the background of the other) implies $|D'| = 1$, while $r^2 = 1$ only when *both* off-diagonal elements are zero.

As fully developed by Wray (2005), the key idea is that ***r^2* is large only when the marker and causative alleles have very similar allele frequencies.** To formally see this, first note that the maximal value of r^2 occurs when a causative *Q* is only found on one marker allele background (*M*). In this setting, let p_M be the frequency of *M*, and $p_Q = \alpha p_M$ be the frequency of *Q*, with $0 < \alpha \leq 1$. Given that the frequency of the *MQ* gamete is $p_Q = \alpha p_M$, then

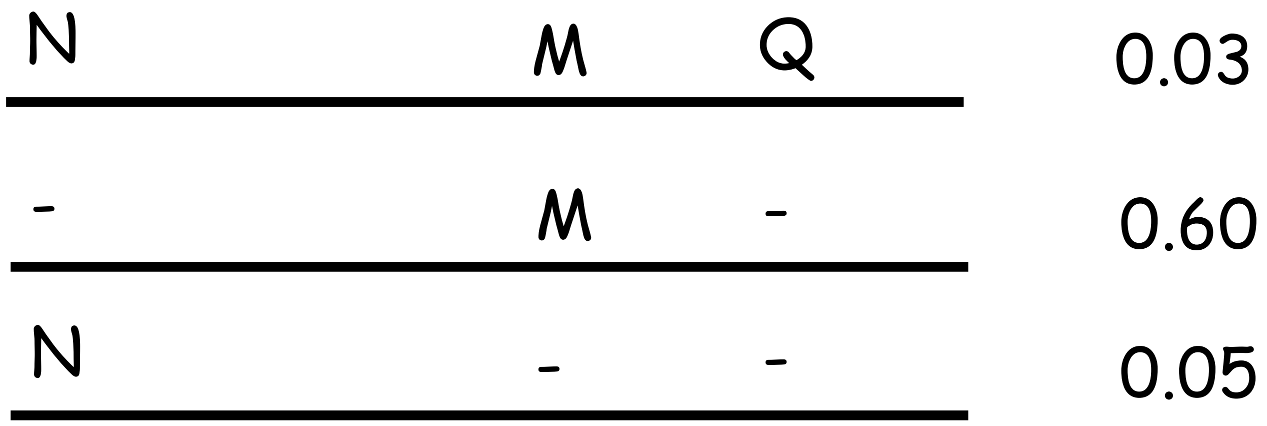
$$D_{MQ} = p_{MQ} - p_M\, p_Q = p_Q - p_M \cdot p_Q = \alpha p_M(1 - p_M)$$

$$r^2_{MQ} = \frac{[\,\alpha p_M(1 - p_M)]^2}{p_M(1 - p_M)\alpha p_M(1 - p_Q)} = \frac{\alpha(1 - p_M)}{1 - p_Q} \leq \alpha \qquad (5.15c)$$



N	M	Q	0.03
-	M	-	0.60
N	-	-	0.05

Example 20.1 A common misconception in GWAS studies is the assumption that *markers closer to a causal site will have larger LD values*, and hence larger marker effects (via larger r^2 values). ***Such need not be the case***. Consider a new causal allele, Q, that arose on an NM marker background (haplotype), where the QTL locus is much closer to M than N. Suppose that recombination is sufficiently rare such that no Q alleles are found on any other haplotypes in the GWAS sample. $D'_{MQ} = D'_{NQ} = 1$ in this case (as Q is *only* found on an N or an M background), but their r^2 values (which determine how much of the actual variance is accounted for by the marker variance) are a function of the marker allele frequencies.



N	M	Q	0.03
-	M	-	0.60
N	-	-	0.05

$$r^2_{MQ} = \frac{\alpha_M(1-p_M)}{1-p_Q} = \frac{\alpha_M - p_Q}{1-p_Q} \leq \alpha_M$$

where p_M is the frequency of the marker allele (M) and α_M the frequency of M alleles associated with Q (so that $\alpha_M p_M = p_Q$ is the frequency of Q). Suppose that the frequency of Q is 3%, while the frequencies of M and N are, respectively, 60% and 5%. Here $p_M = 0.6$ and $\alpha_M = 3/60 = 0.05$, giving $r^2_{MQ} = (0.05 - 0.03)/(1 - 0.03) = 0.0206$, so that the M marker variance accounts for only 2.06% of the causal (actual) variance. Conversely, for the more distant marker, N, $p_N = 0.05$ and $\alpha_N = 3/5 = 0.6$, giving $r^2_{NQ} = (0.6 - 0.03)/(1 - 0.03) = 0.588$. Thus, the marker variance for the *more distant site* captures almost 60% of the actual (causal) variance, a thirty-fold increase over the marker variance for the closer site, M. This is an illustration of the concept from Chapter 5 that r^2 LD values are ***largest when the causal and marker allele frequencies are similar***, and fall off as their absolute frequency difference increases (i.e., as α_M becomes smaller). As we will see, this impacts the power of a GWAS to detect common versus rare alleles.

Part IV:
Implications for GWAS

Key:

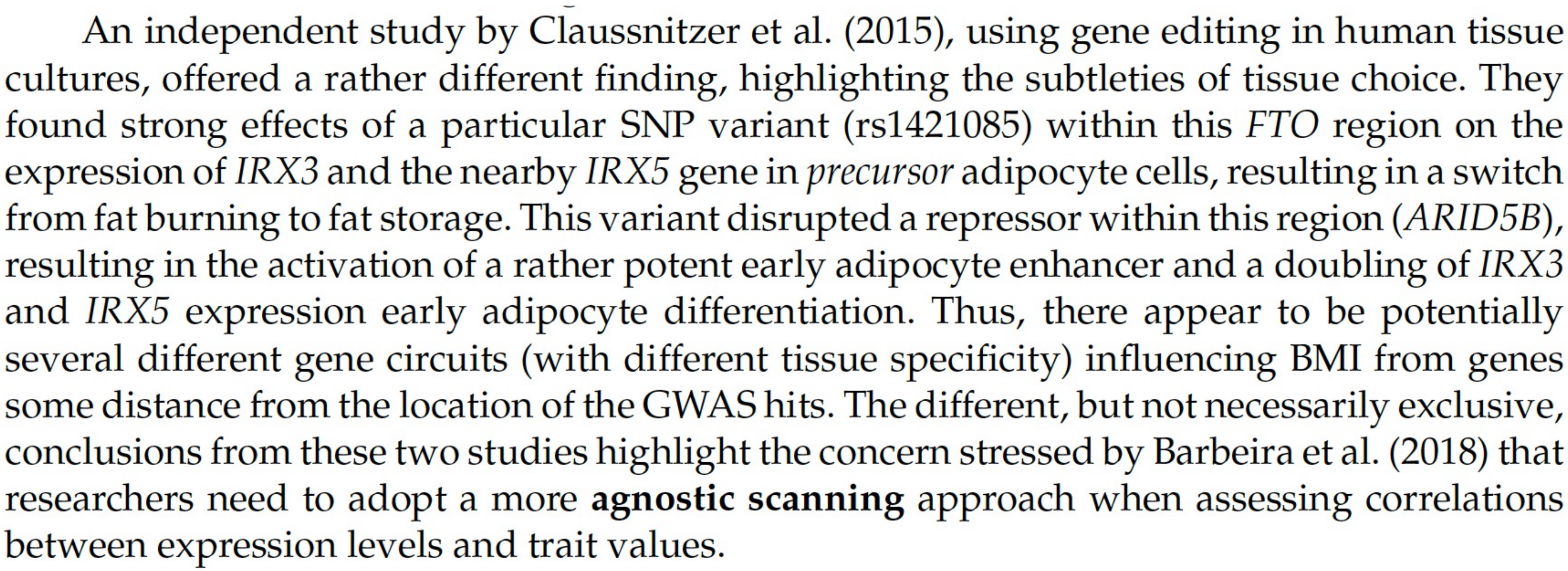
- If the marker and causative alleles have *miss-matched frequencies*, then the LD between them will be small
- Hence, common marker alleles very poorly tag rare alleles
  - Var(Marker) = r^2 Var(Causal site)
  - Most rare alleles not tagged
  - The marker SNP with the strongest signal is likely NOT the closest SNP

Even after sequencing the entire association block (so that *all* variants, including those that are causal, are scored), the **lead**, or **index**, **SNP** (that displaying the most significant p value) within that block is ***likely not the causal variant***, especially when power is low and LD is extreme (Ledur et al. 2010; Udler et al. 2010; van de Bunt et al. 2015; Wu et al. 2017; Huang et al. 2018; Schaid et al. 2018). Simulations by van de Bunt et al. (2015) assuming **whole-genome sequencing** (**WGS**) data still found that the lead SNP corresponded to the causal SNP only 80% of the time when the allele had high frequency and a strong effect ($p = 0.5$; odds ratio, OR, of 1.5), and less than 3% of the time when the allele was less common and of modest effect ($p = 0.05$, OR = 1.1). Hence, even with WGS data and a large population sample, determining the causal variants is far from trivial. The term **QTN** (**quantitative trait nucleotide**) has been used to declare a clear demonstration of a causal SNP (or some other variant, such as a CNV), but this has been very challenging to accomplish in most settings (see Example 21.8 for some exceptions).

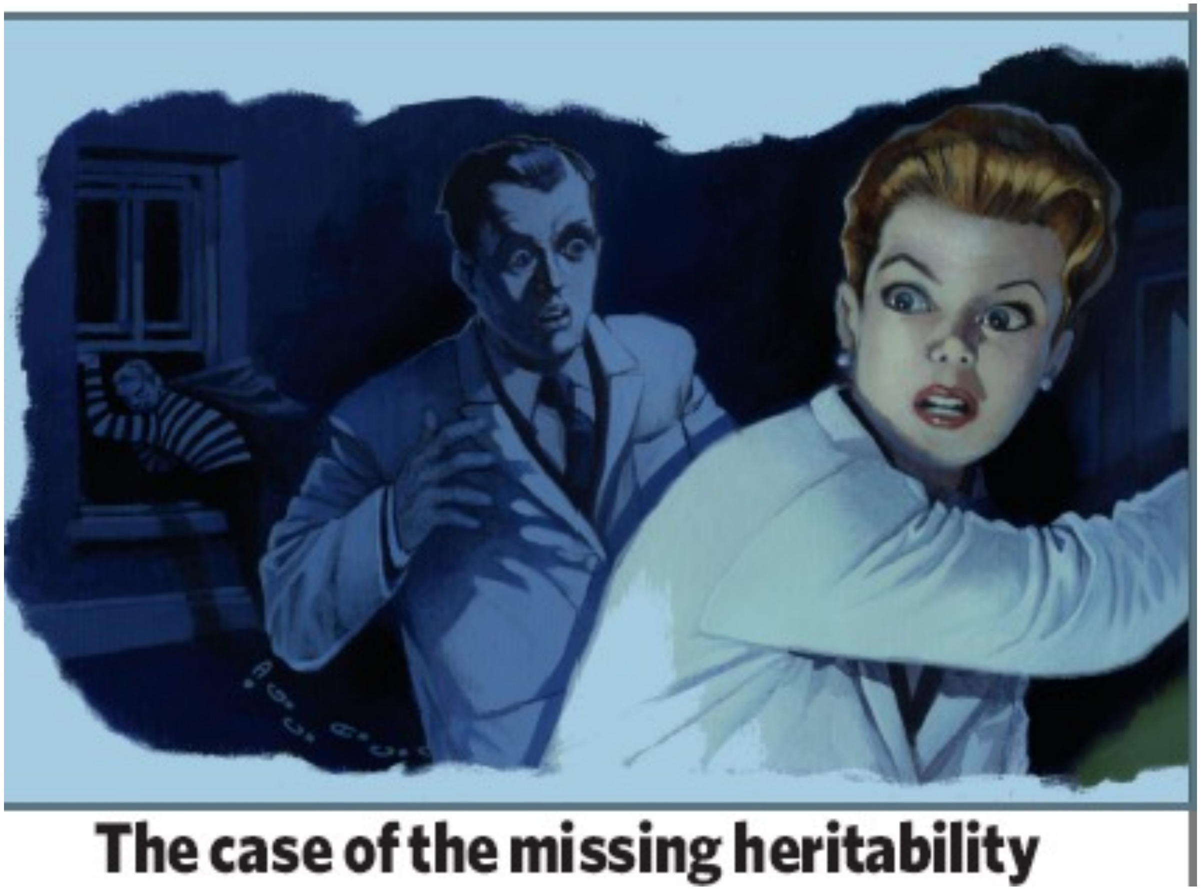
Lead SNP (LS) unlikely causal

- Simulations by Wu et al. (2017)
- If causal SNP (CS) is "common" $p > 0.01$
  - Under WGS, 80% of LS within 10kb of CS
  - Under imputation, goes to 25-35 kb
- If causal SNP is "rare" ($p < 0.01$)
  - Under WGS, 95% of LS are within 5kb
  - Under Imputation, only 37% within 5kb
- Key message: WGS not helpful for common causal SNPs, useful for rare causal SNPs

Example 21.5 An important cautionary tale on fine-mapping was offered by Smemo et al. (2014). A set of roughly 90 variants in very high LD that map within a 47 kb region spanning introns 1 and 2 of the *FTO* gene had very strong, and highly reproducible, GWAS hits for human obesity (measured by body mass index, BMI). Individuals homozygous for risk alleles averaged more than 3kg heavier than individuals homozygous for non-risk alleles. Deletion of *FTO* in mouse models results in leaner mice, while mice overexpressing *FTO* are heavier. Finally, this 47kb region is heavily enriched with *cis*-acting control factors (enhancers, repressors, DNAse I sensitivity sites, TF binding sites). However, *none* of the variants within this region map as eQTLs for *FTO* expression. Smemo et al. found that this regions is involved in chromatin looping to a region over a megabase away containing the gene *IRX3*. In a human EWAS using brain tissue, 11 of the *FTO* SNPs associated with BMI were also eSNPs for *IRX3*, but not *FTO*, expression. Further, of the eSNPs associated with *IRX3* expression in either brain or mature adipose tissue, only those expressed in the brain showed highly significant associations with BMI. Hence, the *FTO* GWAS hits appear to be distal eSNPs that impact expression levels of *IRX3* in the brain. The apparent colocalization of *FTO* GWAS hits and mouse knockout effects gave a misleading picture of how these specific causal sites influence human body mass. Further, focusing expression studies solely on one obvious target, mature adipose tissue, would have missed this signal.



An independent study by Claussnitzer et al. (2015), using gene editing in human tissue cultures, offered a rather different finding, highlighting the subtleties of tissue choice. They found strong effects of a particular SNP variant (rs1421085) within this *FTO* region on the expression of *IRX3* and the nearby *IRX5* gene in *precursor* adipocyte cells, resulting in a switch from fat burning to fat storage. This variant disrupted a repressor within this region (*ARID5B*), resulting in the activation of a rather potent early adipocyte enhancer and a doubling of *IRX3* and *IRX5* expression early adipocyte differentiation. Thus, there appear to be potentially several different gene circuits (with different tissue specificity) influencing BMI from genes some distance from the location of the GWAS hits. The different, but not necessarily exclusive, conclusions from these two studies highlight the concern stressed by Barbeira et al. (2018) that researchers need to adopt a more **agnostic scanning** approach when assessing correlations between expression levels and trait values.



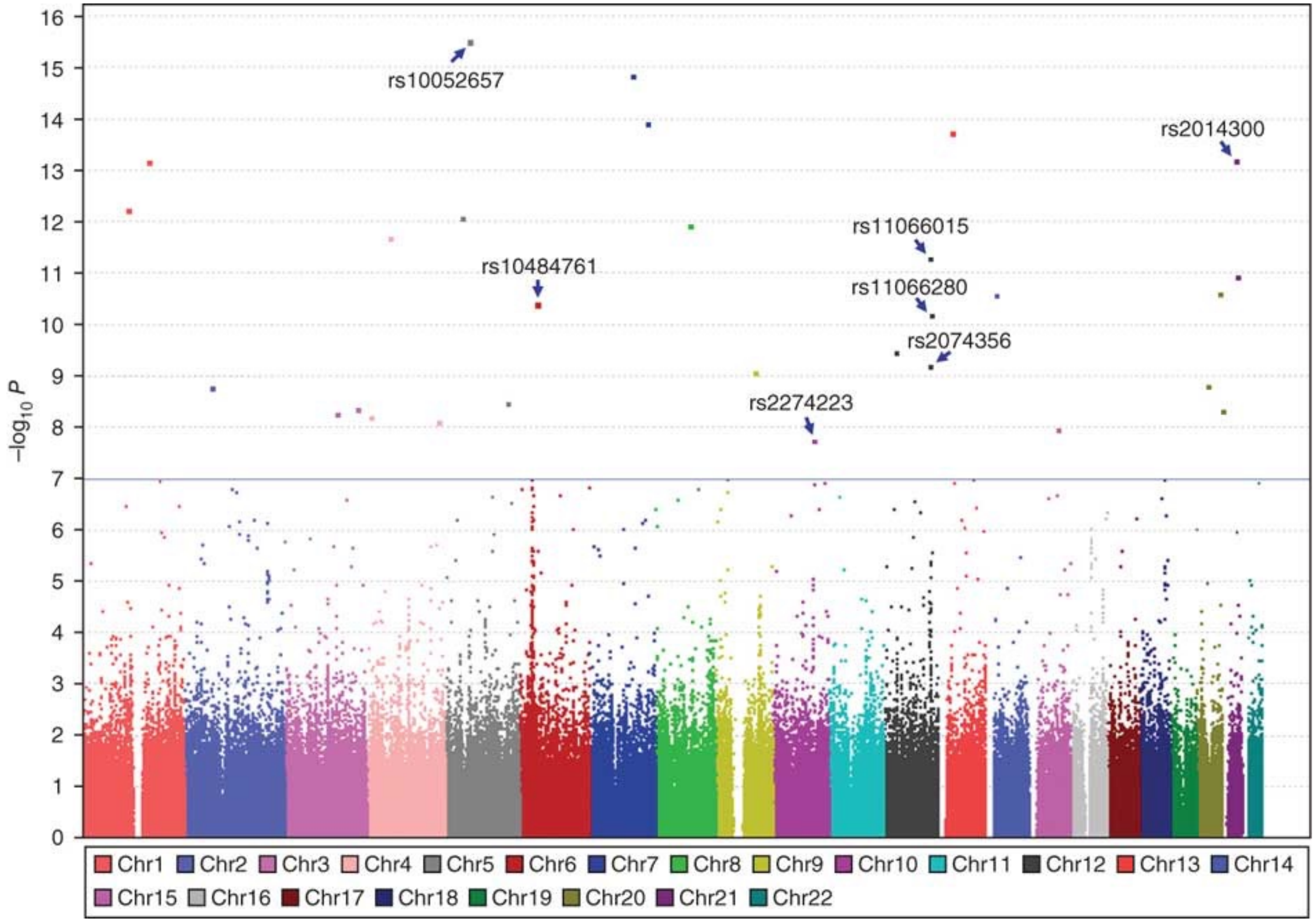
Infamous figure from *Nature* on the angst of human geneticists over the finding that all of their discovered SNPs only accounted for a fraction (~ 10%) of relative-based heritability estimates of human disease.

“There is something simultaneously remarkable and encouraging about the fact that a centuries-old method requiring no more than a ruler, a pencil and (I suppose) a slide rule out performed, by an order of magnitude, the fruits of the genomic revolution”

--Ben Sheldon (2013)

No “missing heritability”

- Low power because sites of small effect are unlikely to be called as significant, esp. given the high stringency associated with control of false positives over tens of thousands of tests.



Only these markers included (as they are declared significant)

Huge number of important, but small effect, markers not declared significant

Haloptype	Frequency	effect
QM	rp	a
qM	$(1-r)p$	0
qm	$1-p$	0

Resulting SNP effects:

Effect of m = 0

Effect of M = ar

Genetic variation associated with causative site Q
= $2(rp)(1-rp)\, a^2 \sim 2rpa^2$ when Q rare – gives a small signal.

Genetic variation associated with marker SNP M is
$2p(1-p)(ar)^2 \sim 2pa^2r^2$

Ratio of marker/true effect variance is ~ r

Thus, if Q rare within the M class ($r \ll 1$), even a completely linked SNP marker captures only a fraction of variance

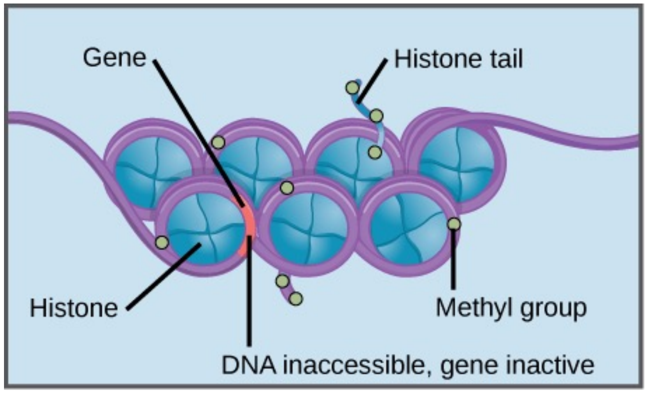
The tide starts to shift as N grows

- Human height, h^2 ~ 0.7 to 0.8
- Using only markers reaching genome-wide significance, and scoring hits using common SNPs
- 2008, GWAS N = 40,000
  - 27 “hits”, 6%
- 2010, N = 180,000
  - 200 “hits”, 14%
- 2014, N = 253,000
  - 700 hits, 20%
  - Using 2000, 3700 & 9500 SNPs, 26%, 30%, 36%
- 2018, N = 700,000
  - 3,000 hits, 35%

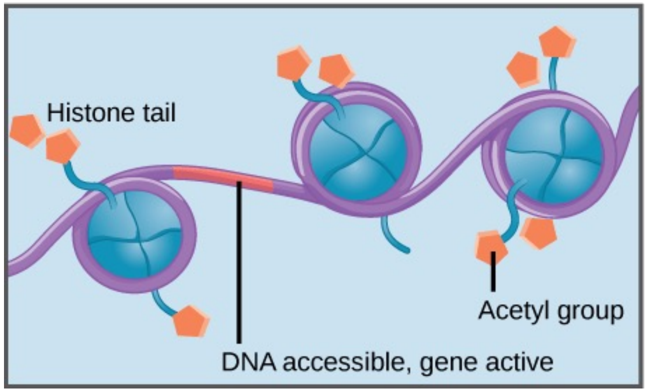
- If one includes all markers, and asks that variance they explain,
- 2010, N = 4,000 -> 62% of heritability
- 2015, N = 44,000 -> 80%
- Variance of variance explained by a chromosome is very highly correlated with its length
  - Schizophrenia, 70% of all 1 MB regions contain a risk site
- 2022, N = 25,500 WGS
  - 68% using common markers (p > 1%)
  - 50-56% using imputed markers (predicting rare alleles)
  - > 90% using WGS. Most gain from very rare alleles in low LD with common sites, hence poorly imputed
- Current estimates (based on estimated distribution of effect sizes) of 95,000 to > 100,000

Part V:
GWAS hits &
Gene regulation

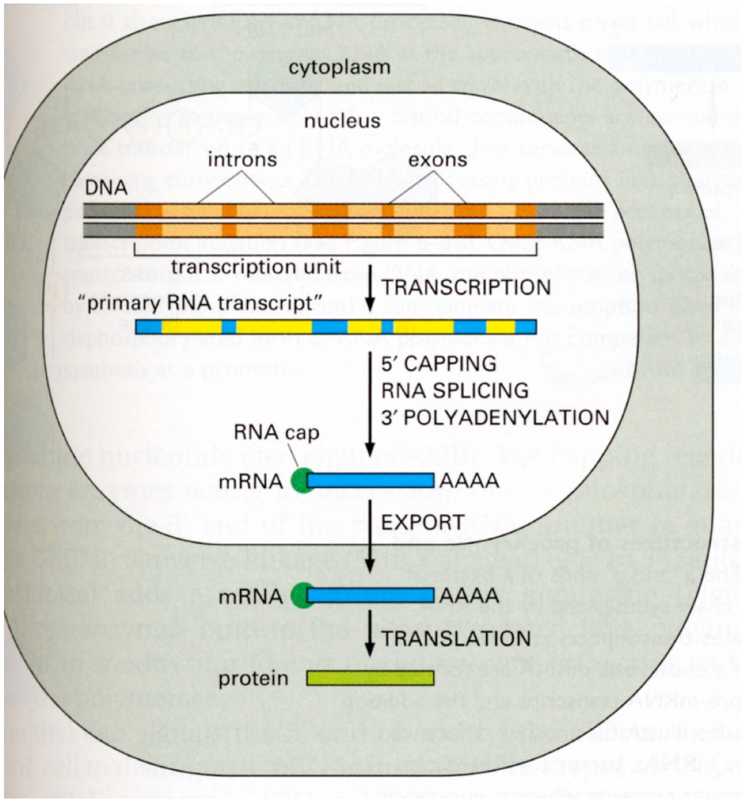
Steps from gene regulation to trait variation

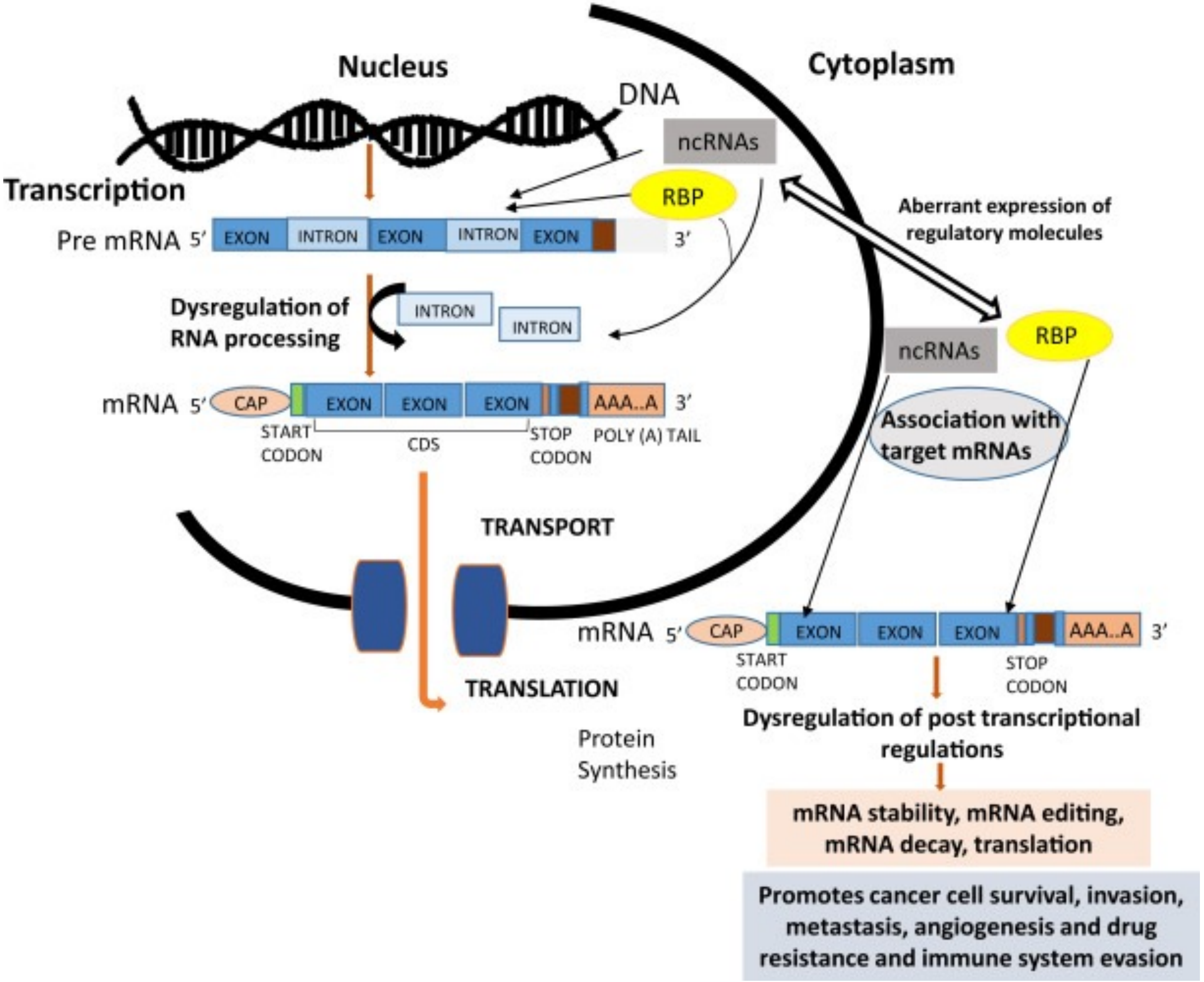


Methylation of DNA and histones causes nucleosomes to pack tightly together. Transcription factors cannot bind the DNA, and genes are not expressed.

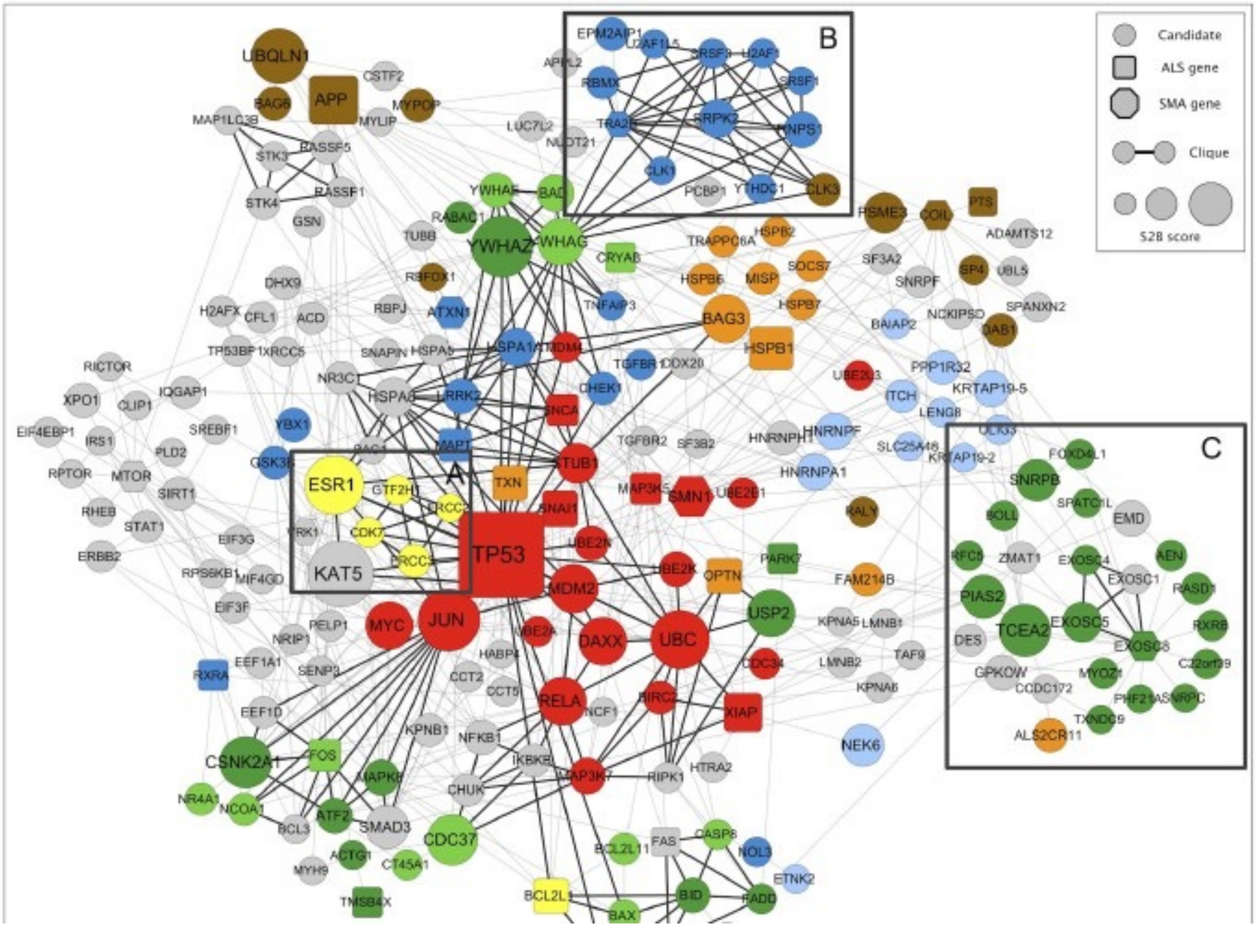


Histone acetylation results in loose packing of nucleosomes. Transcription factors can bind the DNA and genes are expressed.




Nucleus
Cytoplasm
DNA
ncRNAs
RBP
Transcription
Pre mRNA 5′
EXON
INTRON
EXON
INTRON
EXON
3′
Aberrant expression of regulatory molecules
Dysregulation of RNA processing
INTRON
INTRON
ncRNAs
RBP
mRNA 5′
CAP
EXON
EXON
EXON
AAA..A
3′
START CODON
CDS
STOP CODON
POLY (A) TAIL
Association with target mRNAs
TRANSPORT
mRNA 5′
CAP
EXON
EXON
EXON
AAA..A
3′
START CODON
STOP CODON
TRANSLATION
Protein Synthesis
Dysregulation of post transcriptional regulations
mRNA stability, mRNA editing, mRNA decay, translation
Promotes cancer cell survival, invasion, metastasis, angiogenesis and drug resistance and immune system evasion

Gene products



Trait value

QTL/GWAS & regulation

- A number of QTL and GWAS studies looked at QTLs showing variation for a number of regulatory features, such as
  - mRNA levels
  - Isoform variation in splicing
  - Histone modification
  - Large number of such regQTLs/regSNPs found (despite the low power, N < 5000) of most designs.
  - Trait GWAS hits enriched for regSNPs

Table 21.1 A few of the different classes of QTLs (SNPs). The general terminology is to use QTLs generically, especially in a linkage-based analysis to indicate a region, and SNP in a GWAS setting to refer to a SNP showing an association. The QTL/SNP terminology is a bit idiosyncratic, with different versions for some of these abbreviations appearing in the literature.

acQTL/acSNP	Chromatin acetylation QTL/SNP
aseQTL/aseSNP	Allele-specific expression QTL/SNP
caQTL/caSNP	Chromatin accessibility QTL/SNP
cis-xQTL/*cis*-xSNP	*Cis* (local) QTL/SNP for feature x
dsQTL/dsSNP	DNAse I sensitivity QTL/SNP
eQTL/eSNP	RNA expression QTL/SNP
epiQTL/epiSNP	Chromosome epiallele QTL/SNP
hQTL/hSNP	Histone QTL/SNP
haQTL/haSNP	Histone acetylation QTL/SNP
hmQTL/hmSNP	Histone methylation QTL/SNP
meQTL/meSNP	DNA methylation QTL/SNP
methQTL/methSNP	DNA methylation QTL/SNP
miR-QTL/miR-SNP	MicroRNA QTL/SNP
pQTL/pSNP	Protein expression QTL/SNP
pb-xQTL/pb-xSNP	Population-based QTL/SNP for feature x
QTN	Quantitative trait nucleotide
QTT	Quantitative trait transcript
rQTL/rSNP	Ribsome occupancy QTL/SNP
regQTL/regSNP	Regulatory QTL/SNP
sQTL/sSNP	Splicing QTL/SNP
sb-xQTL/sb-xSNP	Sex-based QTL/SNP for feature x
tQTL/tSNP	Trait QTL/SNP
trans-xQTL/*trans*-xSNP	*Trans* (distal) QTL/SNP for feature x
vQTL/vSNP	Variance QTL/SNP

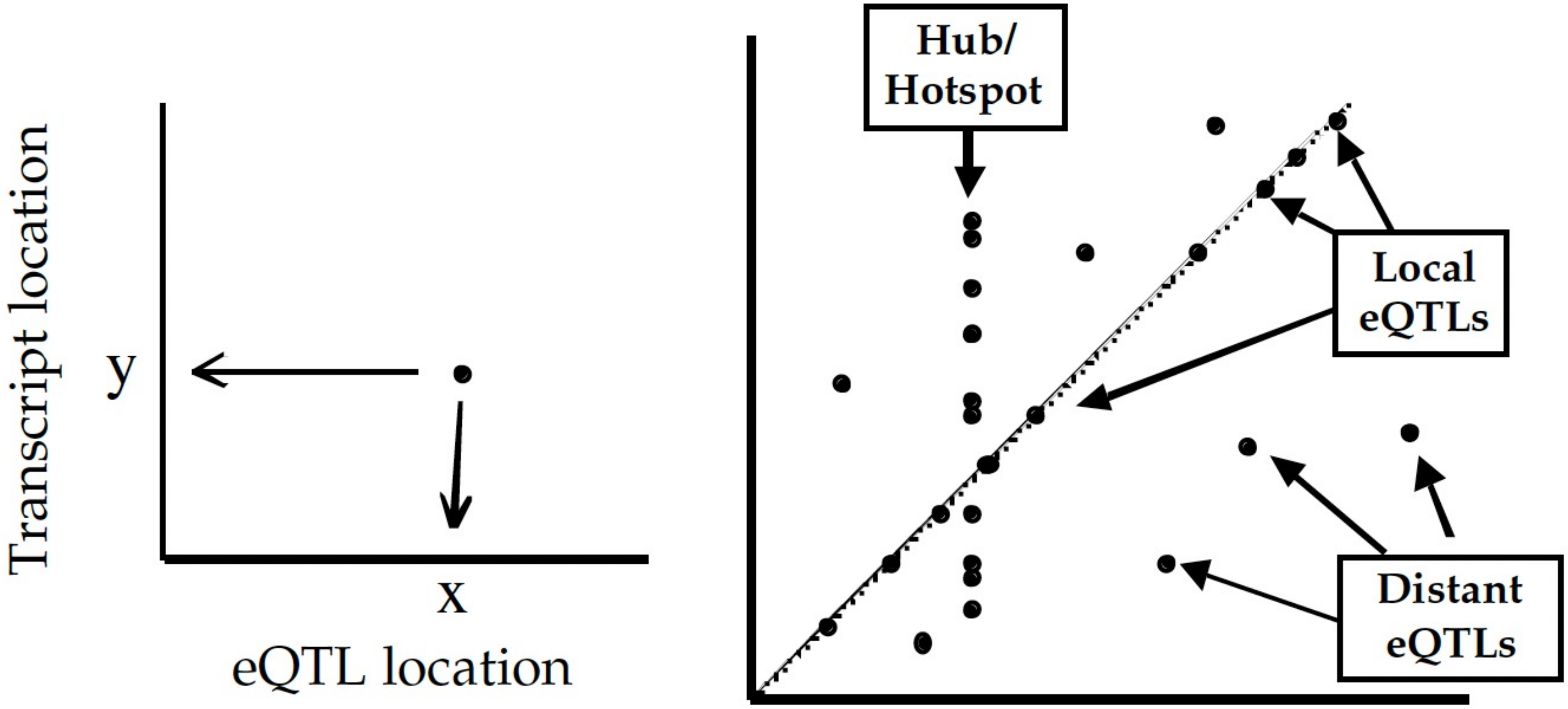


Figure 21.1 A stylized **transcriptome map**, plotting eQTL locations versus the location of the coding region for a transcript. Both axes correspond to genome position, with the horizontal (x) axis denoting a region/marker being tested as an eQTL and the vertical (y) axis the location of the coding region for a transcript (occasionally in the literature these two axes are reversed). A point or pixel at position (x, y) on this map indicates a significant association between a transcript whose coding region is at genomic position y and a marker/region at genomic position x. Points falling on the diagonal correspond to eQTLs that map very close to, or at, the same location as the coding region for the transcript they influence. These have been called ***cis* eQTLs**, but as discussed in the text are better referred to as **local** (**proximal**) **eQTLs**. Points falling off the diagonal correspond to eQTL locations that influence transcripts whose coding regions are at a different location from the eQTL. These have been called ***trans* eQTLs**, but are better referred to as **distant (distal) eQTLs**. A vertical stack of points corresponds to a (small) genomic region that is enriched for eQTLs, and is called a **hotspot** or **hub**, with the eQTLs in that region impacting numerous transcripts.

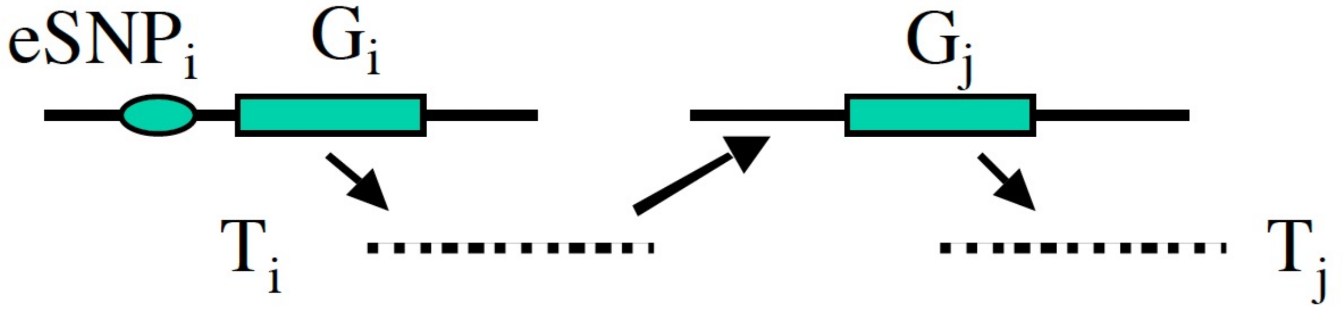


Figure 21.2 The concept of ***cis*-mediation**. The observation is that eSNP i acts at some distance away from a coding region (G_j) to regulate the level of its transcript T_j (eSNP i is a *trans*-eSNP for transcript j). The **mediation hypothesis** is that the impact of eSNP i is through a *cis* effect on the transcript from (local) gene G_i, whose transcript T_i then influences the regulation of transcript T_j of a distant gene G_j. Path analysis methods (Figure 21.3; Appendix 2) allow this idea to be extended over much more complex regulatory networks, as well as providing a framework for estimating direct and indirect effects of any component player (Example 21.3).

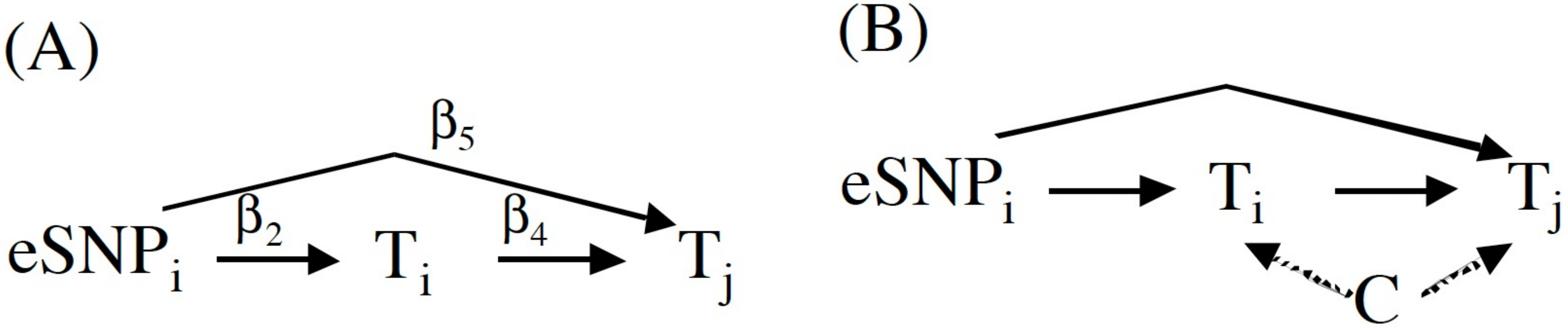


Figure 21.3 Path-analysis of a trio (eSNP$_i$, T_i, and T_j) to separate direct and indirect effects. **(A):** The path diagram when only these trio elements are involved (Example 21.3). The direct effect eSNP$_i$ $\rightarrow$ T_j (avoiding T_i) is given by β_5, and the indirect (**mediated**) effect via T_i, eSNP$_i$ $\rightarrow T_i \rightarrow T_j$, is $\beta_2 \cdot \beta_4$. The total effect is given by $\beta_1 = \beta_5 + \beta_2 \cdot \beta_4$. This same logic can be applied to a trio of an eSNP, a transcript it impacts, and a trait value (i.e., replacing T_j with z_k, the value for trait k), or other more complex regulatory pathways (e.g., Example A2.2). **(B):** Mediator confounding occurs, in the simplest case, when an unmeasured factor (a confounding variable C; U, for unmeasured variable, is also used in the literature) impacts both T_i and T_j. In this setting, estimates of the direct and indirect effects can be biased.

Example 21.3 As developed by a number of investigators, one can use conditional regressions (path analysis methods; Appendix 2) to both detect, and quantify, the amount of mediation that gene i has on transcript j (Chen et al 2007; Jiang et al. 2013; Pierce et al. 2014; Yang et al. 2017; Yao et al. 2017; Shan et al. 2019). This is done using a nested series of regressions to establish causality. Using the notation in Figures 21.2 and 21.3, first consider the association between the dosage of SNP i (the minor allele number copy number N_i = 0, 1, or 2) and the transcript associated with coding region j (T_j),

$$T_j = \alpha_1 + \beta_1 N_i + e_1 \tag{21.1a}$$

One declares SNP i to be a *trans*-eSNP for coding region j when the slope β_1 is significant. This slope measures the **total effect** of SNP i on T_j, the contributions from both direct effects and indirect effects (such as through T_i). Next, we declare SNP i to be a *cis*-eSNP for coding region G_i when the regression

$$T_i = \alpha_2 + \beta_2 N_i + e_2 \tag{21.1b}$$

has a significant slope. Similarly, we declare the T_i has an effect on T_j when β_3 is significant for the regression

$$T_j = \alpha_3 + \beta_3 T_i + e_3 \tag{21.1c}$$

Significant slopes in the above three regressions establish that: i) SNP i is associated with T_j ($\beta_1 \neq 0$); ii) SNP i is associated with T_i ($\beta_2 \neq 0$); and iii) T_i is associated with T_j ($\beta_3 \neq 0$). These univariate regressions, by themselves, do not separate direct from indirect effects. To do so, a multiple regression of T_j is constructed based on both N_i and T_i,

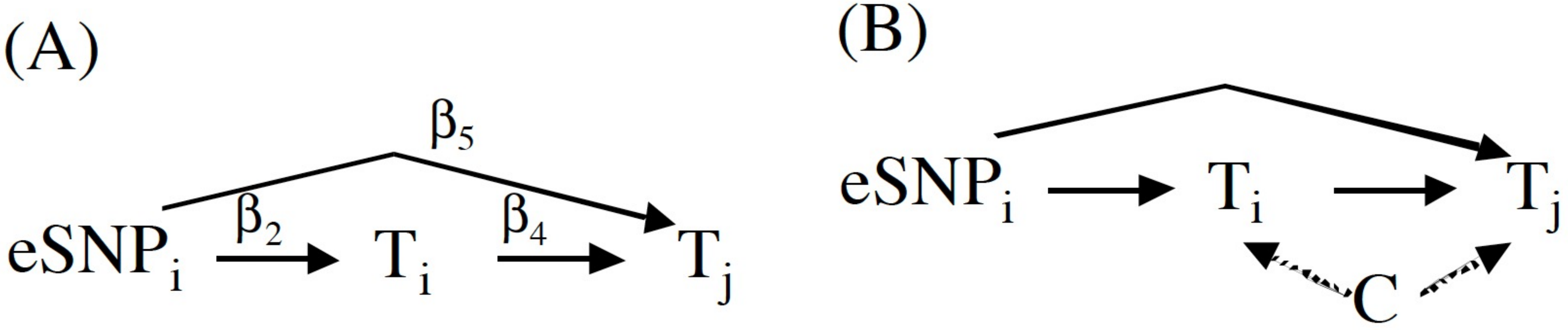
$$T_j = \alpha_4 + \beta_4 T_i + \beta_5 N_i + e_4 \tag{21.1c}$$

$$T_j = \alpha_4 + \beta_4 T_i + \beta_5 N_i + e_4 \tag{21.1c}$$

If $\beta_5 = 0$, then any effect from SNP i on T_j is simply through its effect on T_i, namely, **full mediation** (the effect of SNP i on T_j is entirely through its *cis*-effect on T_i). When both β_4 and β_5 are significant, then **partial mediation** occurs, where *both* T_i and SNP i (the latter through a path independent of T_i; Figure 21.3A) impact T_j. Note that this logic need not be restricted to just transcripts, one could measure (say) P_i, the level of protein from gene i, or some other regulatory measure such as methylation, splicing, etc. Modifications of permutation tests to accommodate the correlation structure of mediation analysis are discussed by Jiang et al. (2013) and T. Wang et al. (2020). An excellent overview of mediation analysis is given by Otter et al. (2018).

From the theory of path analysis (Appendix 2), the indirect effect of SNP i on T_j through the path given by T_i, is just the product of the path coefficients, which turns out to be $\beta_2 \cdot \beta_4$ from the above regressions. As shown in Figure 21.3A, the total path effect β_1 assumes the potential of a direct effect β_5 from eSNP$_i$ to T_j ($N_i \rightarrow T_j$) and an indirect of eSNP$_i$ via paths through T_j ($N_i \rightarrow T_i \rightarrow T_j$) with effect $\beta_2 \cdot \beta_4$. Hence, the proportion of the total effect on T_j from eSNP$_i$ mediated via T_i is

$$(\beta_1 - \beta_5)/\beta_1 = \beta_2\beta_4/\beta_1 \tag{21.1d}$$



From the theory of path analysis (Appendix 2), the indirect effect of SNP i on T_j through the path given by T_i, is just the product of the path coefficients, which turns out to be $\beta_2 \cdot \beta_4$ from the above regressions. As shown in Figure 21.3A, the total path effect β_1 assumes the potential of a direct effect β_5 from eSNP$_i$ to T_j ($N_i \rightarrow T_j$) and an indirect of eSNP$_i$ via paths through T_j ($N_i \rightarrow T_i \rightarrow T_j$) with effect $\beta_2 \cdot \beta_4$. Hence, the proportion of the total effect on T_j from eSNP$_i$ mediated via T_i is

$$(\beta_1 - \beta_5)/\beta_1 = \beta_2\beta_4/\beta_1 \tag{21.1d}$$

If there are no unscored correlated factors that impact members of this trio, then the relation $\beta_1 = \beta_5 + \beta_2 \cdot \beta_4$, namely total effect = direct effect plus indirect effect, should hold. If it does not, one is likely missing correlated elements (confounders). Figure 21.3B shows one example. Such confounding could be caused by the focal eSNP$_i$ being in LD with different causal SNPs for the *cis* effect on T_i and the *trans* effect on T_j (Pierce et al. 2014).

Colocalization analysis

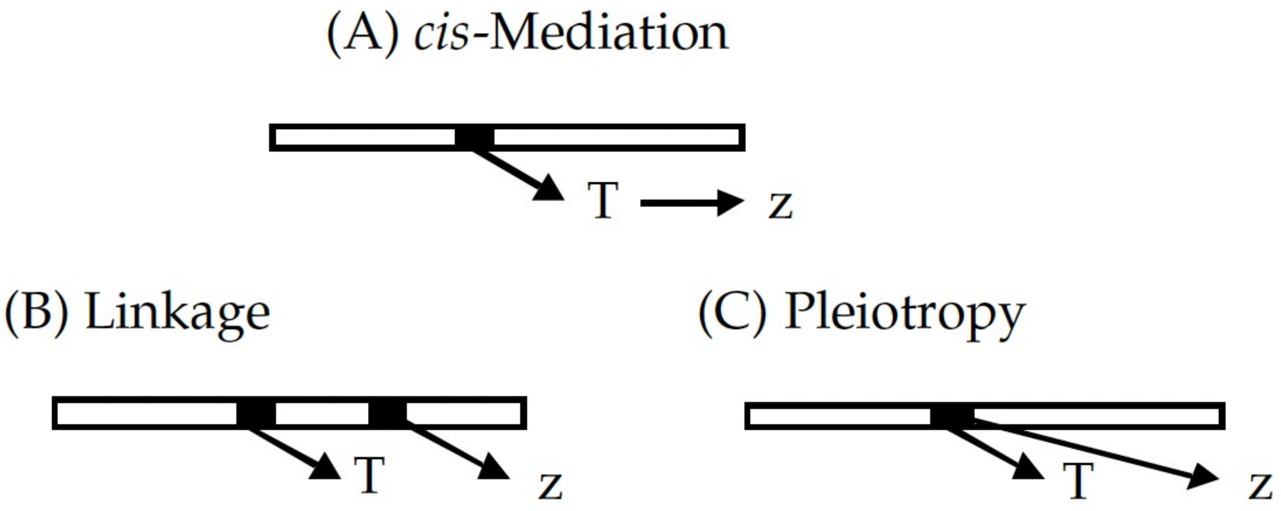


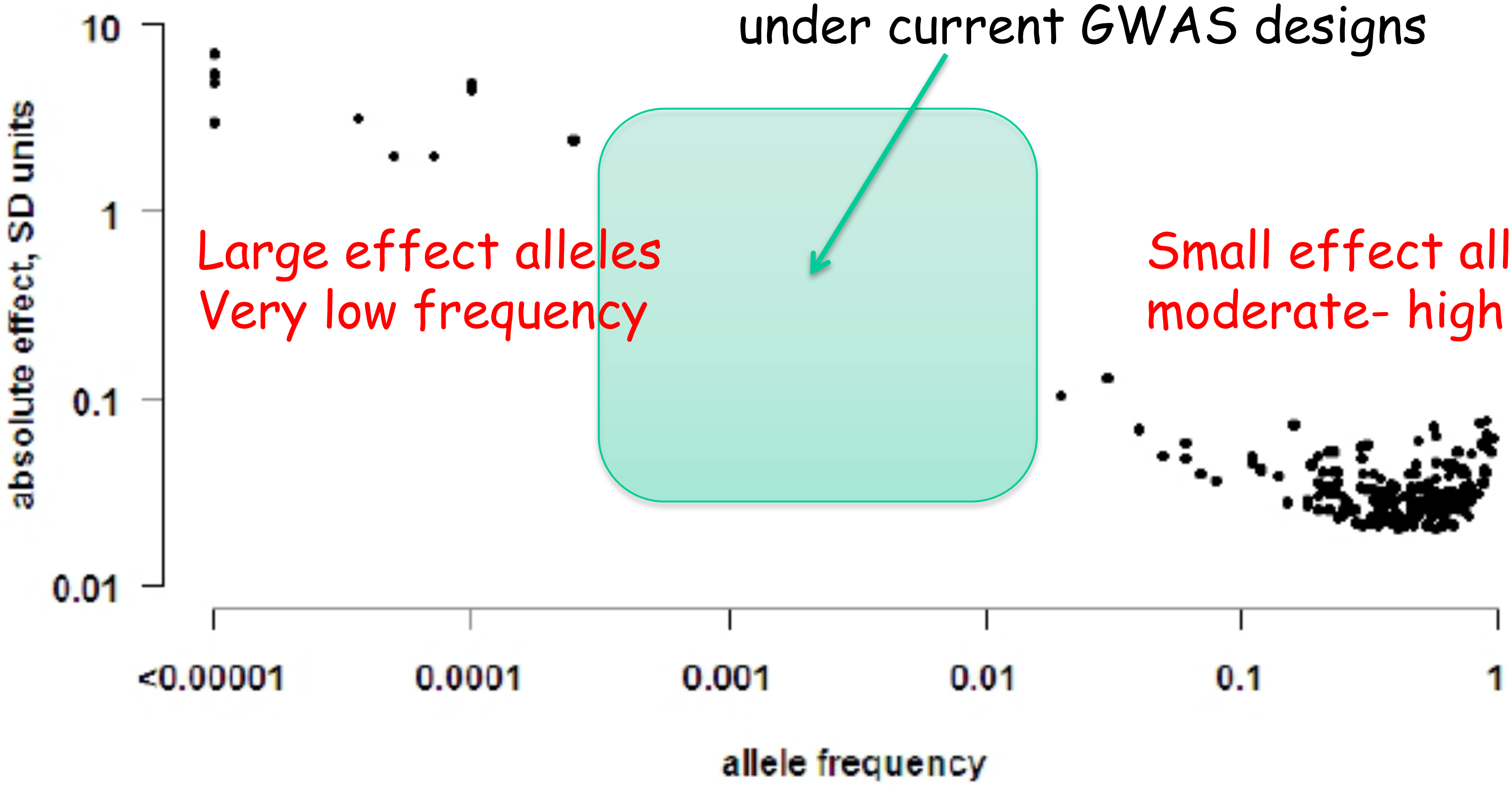
Figure 21.4 An apparent colocalization between a GWAS SNP for trait z (the small black box), transcript (T), and trait (z) could occur through three different pathways. **(A):** Direct *cis*-mediation. The GWAS SNP is an eSNP which directly influences the transcript, which in turn directly influences the trait. **(B):** Linkage. Two tightly linked SNPs are involved. One directly impacts the transcript, the second directly impacts the trait. **(C):** Pleiotropy. The same SNP directly impacts transcript levels and trait values separately, but the transcript level does not impact the trait value. Any combination of these different pathways could be involved, such as a direct *cis*-mediated SNP tightly linked to a separate SNP that only impacts the trait.

Part VI:
GWAS, Architectures, and selection

Common vs rare alleles

- Common alleles are old (under drift)
  - Mostly regulatory?
- Rare alleles are recent
  - Mostly structural>
- Most Hits have small variances, but could be common alleles of small effect or rare alleles of large effect
- Inverse correlations between effect size and allele frequency commonly seen

Data on human height loci



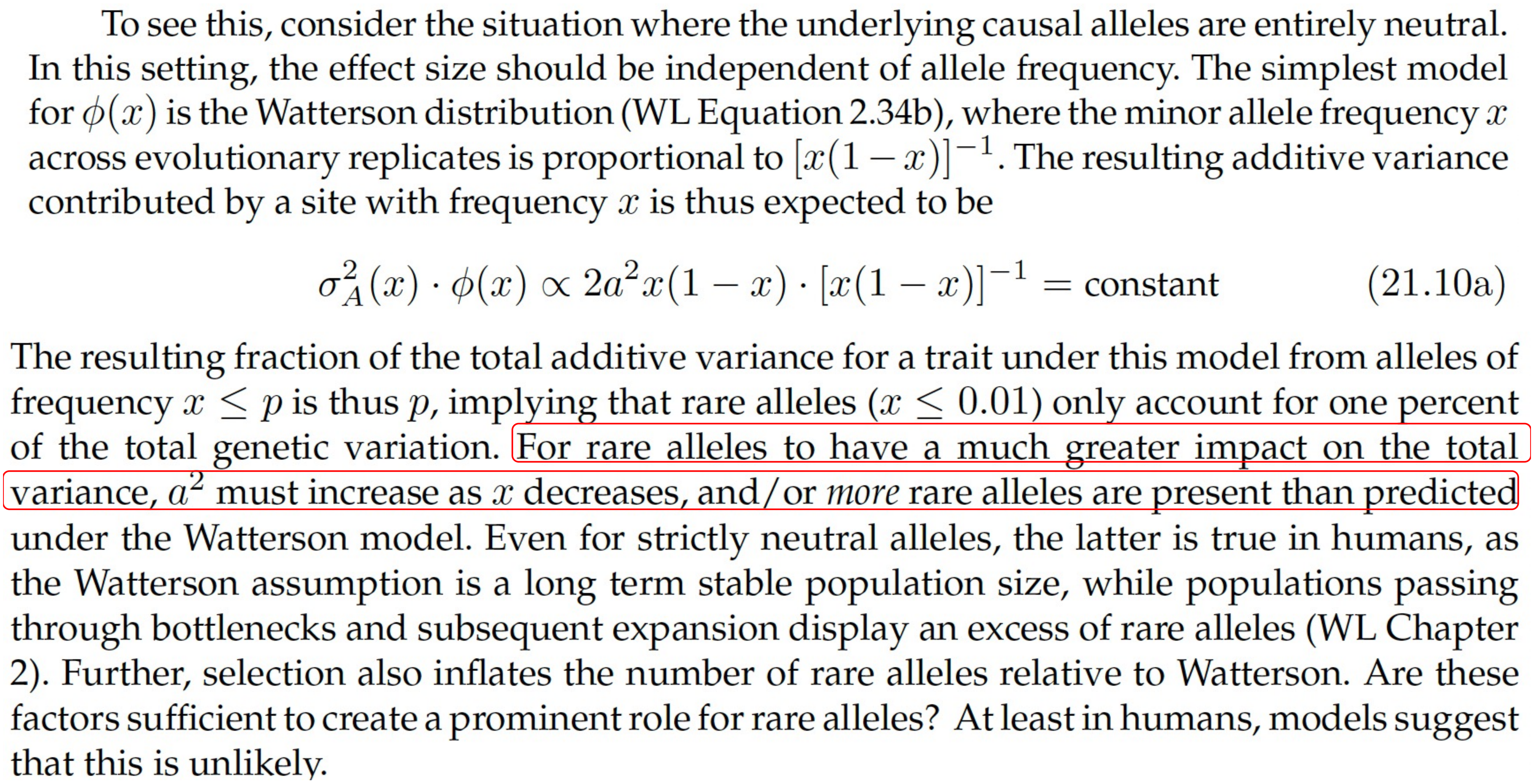
Zone of no power
under current GWAS designs

Large effect alleles
Very low frequency

Small effect alleles
moderate- high frequen

Genes of large effect most likely with deleterious effects of other traits, resulting in their low frequency

Peter Visscher's group at UQ



To see this, consider the situation where the underlying causal alleles are entirely neutral. In this setting, the effect size should be independent of allele frequency. The simplest model for $\phi(x)$ is the Watterson distribution (WL Equation 2.34b), where the minor allele frequency x across evolutionary replicates is proportional to $[x(1-x)]^{-1}$. The resulting additive variance contributed by a site with frequency x is thus expected to be

$$\sigma_A^2(x) \cdot \phi(x) \propto 2a^2 x(1-x) \cdot [x(1-x)]^{-1} = \text{constant} \tag{21.10a}$$

The resulting fraction of the total additive variance for a trait under this model from alleles of frequency $x \leq p$ is thus p, implying that rare alleles ($x \leq 0.01$) only account for one percent of the total genetic variation. For rare alleles to have a much greater impact on the total variance, a^2 must increase as x decreases, and/or *more* rare alleles are present than predicted under the Watterson model. Even for strictly neutral alleles, the latter is true in humans, as the Watterson assumption is a long term stable population size, while populations passing through bottlenecks and subsequent expansion display an excess of rare alleles (WL Chapter 2). Further, selection also inflates the number of rare alleles relative to Watterson. Are these factors sufficient to create a prominent role for rare alleles? At least in humans, models suggest that this is unlikely.

This was nicely illustrated by Zeng et al. (2018) and Schoech et al. (2019), who used a two-component mixture model for the additive effect a of an allele by assuming (for SNP j),

$$p(a_j \,|\, x) = \delta_0 \cdot \pi_0 + N(0, [2x_j(1-x_j)]^S \sigma_a^2) \cdot (1-\pi_0) \tag{21.10b}$$

where $(1-\pi_0)$ is the **polygenicity**, the fraction of all SNPs that impact a trait, the delta function δ_0 denotes a point mass at zero ($a = 0$), and S is a selection parameter. A value of $S < 0$ implies that average a^2 values increase as x decreases (corresponding to negative selection against alleles), while $S = 0$ corresponds to a neutral assumption of no correlation between a^2 and x. MCMC (Appendix 8) can be used to estimate the model parameters S, σ_a^2, and π_0, an approach Zeng et al. called **BayesS**. Note that by rearranging Equation 2.3b,

$$E[a^2 \,|\, x, a^2 > 0] = \sigma^2(a \,|\, a^2 > 0) + (E[a \,|\, a^2 > 0])^2 = [\,2x_j(1-x_j)]^S \sigma_a^2 + 0^2 \tag{21.10c}$$

showing that the variation associated with SNPs with an MAF of x is

$$\left([\,2x_j(1-x_j)]^S \sigma_a^2\right) \cdot 2x_j(1-x_j) = [\,2x_j(1-x_j)]^{1+S} \sigma_a^2 \tag{21.10d}$$

showing that the variation associated with SNPs with an MAF of x is

$$\left([\,2x_j(1-x_j)]^S \sigma_a^2 \right) \cdot 2x_j(1-x_j) = [\,2x_j(1-x_j)]^{1+S} \sigma_a^2 \tag{21.10d}$$

Both Zeng et al. (2018) and Schoech et al. (2019) considered over two dozen, largely non-overlapping, traits/diseases from the UK Biobank. Zeng et al. found that all but one of their traits had a negative estimate of S (ranging from -0.609 to 0.012), 24 of which were significantly negative, with a median S of -0.37. The polygenicity $(1-\pi_0)$ had a median value of 5.4% and ranged from 0.6% to 14.0%. Schoech et al. obtained very similar results for S. Substituting these S values into various population genetic models for $\phi(x|s)$ showed that no more than 10% of the variance could be due to rare alleles $(x \leq 0.01)$.

Key GWAS observations

- Traits are massively polygenic (> 10K sites)
- The per-site variation is typically very small
- Effect sizes range over at least two orders of magnitude
- Inverse correlation between effect size and allele frequency
- Most significant sites in noncoding regions, but highly enriched for regSNPs

The Omnigenic Model

- Pritchard suggested a model for the molecular basis of QG variation: the Omnigenic (all-genes) model
- A few core genes active in the tissues directly impacting the trait/disease of interest
- A massive number of peripheral genes that show regulatory variants in these tissues
  - Implication: any gene showing regulatory variation in a tissue can have a nontrivial impact on trait variation
  - Large fractions of the genome can host regulations with regulatory impact

An old friend

- The omnigenic model is, in part, a recasting of the large (& rare) vs. small (& common) debate
- Variants in core genes are expected to often have large effects and thus be rare
- Variants in peripheral genes are expected to have small effects and thus be common

Current view of QG Variation:

- Regulatory regions are at least as important as structural regions
  - Focusing on coding regions may be misguided
- Much of the low-frequency variation is likely due to novel mutations, restricted to a lineage or extended pedigree, but often in the same gene
  - Hence, unlikely to show up in association studies unless related lines are used
  - However, there are likely candidate genes at which novel variation arises