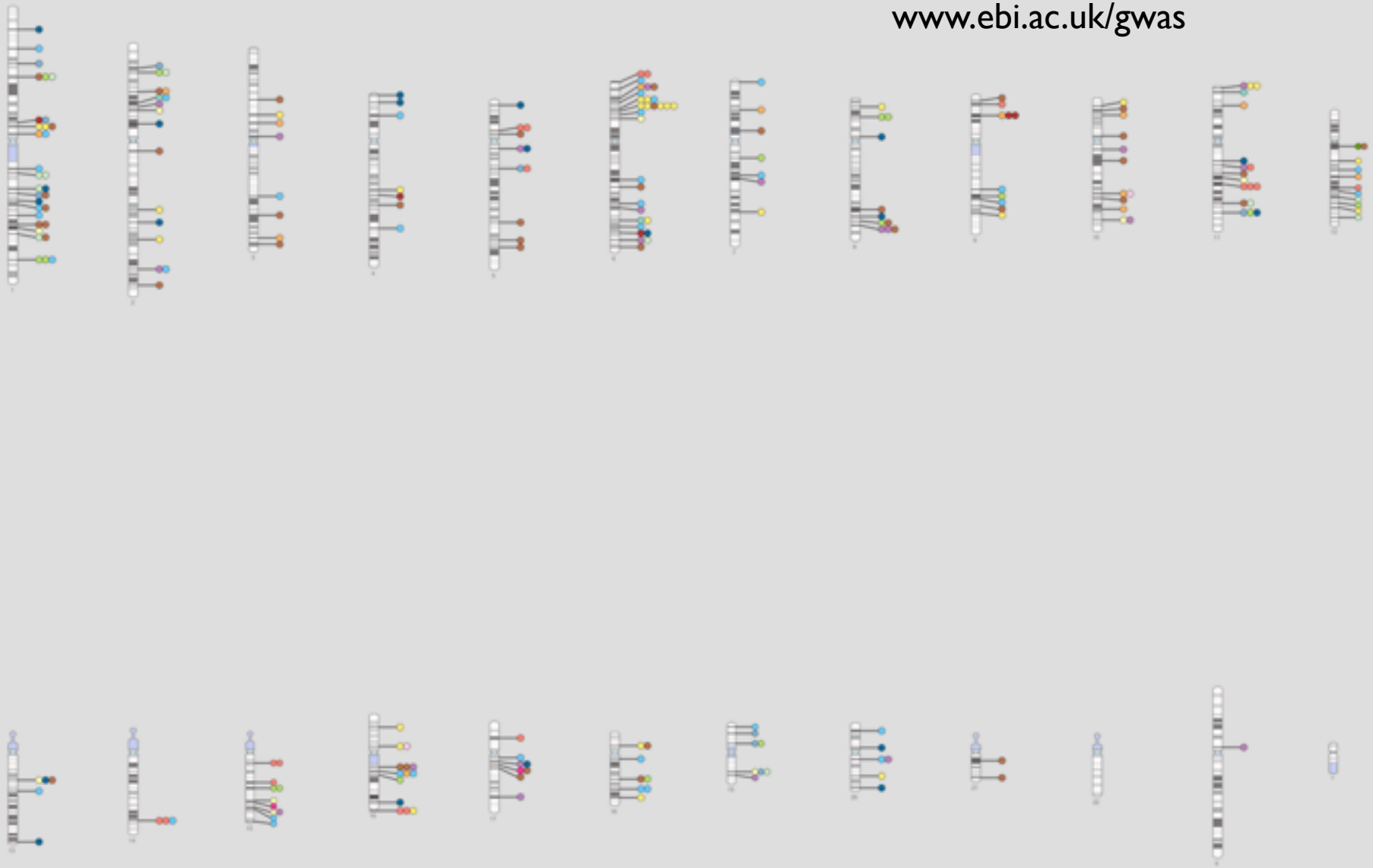
GWAS 12

Matti Pirinen
University of Helsinki
27.2.2019

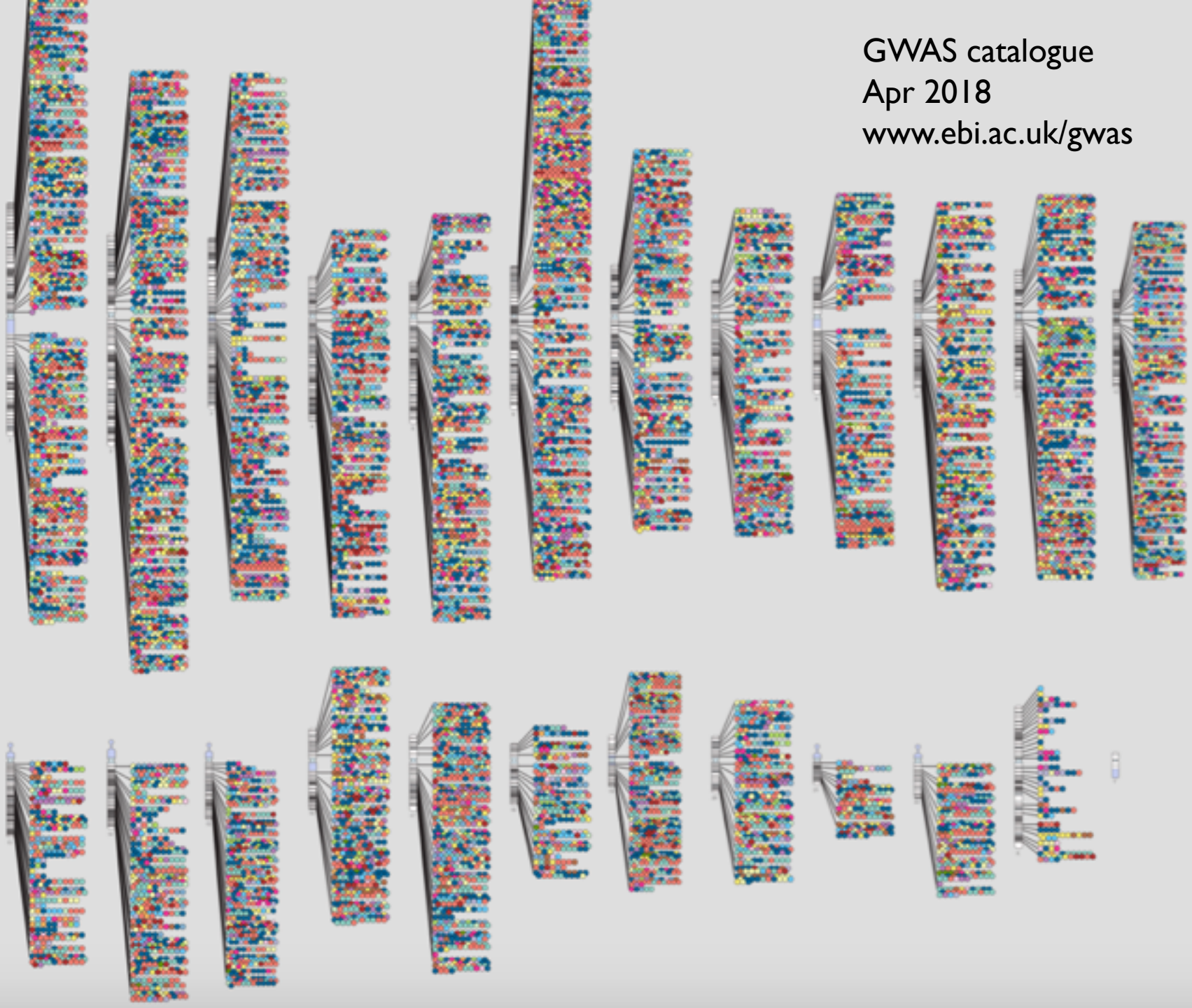
GWAS catalogue
Q2 / 2008
www.ebi.ac.uk/gwas



First GWAS findings were done in 2005.

In 2008 there were a few dozen SNPs with P<5e-8

- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait



GWAS catalogue
Apr 2018
www.ebi.ac.uk/gwas

Associations 69,885
Studies: 5,125
Papers: 3,378

- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

GWAS LOGIC AND GWAS CRITICISM

- GWAS catalogue (https://www.ebi.ac.uk/gwas/) reports 3,764 publications and 107,785 unique SNP-trait associations (31-Jan-2019)
  - Many common diseases (e.g. Schizophrenia, T2D, CAD have 100s of associations)
- "GWAS logic" is that these findings point us to biology and potential targets for therapies and make early prediction of future disease possible
- Next we'll look
  - Does the GWAS logic seem to work based on empirical data?
  - What kind of criticism is there about GWAS?

GWAS CRITICISM

1. Effect sizes of GWAS findings are too small to be of any (practical) use
2. Mechanisms behind GWAS findings are missing
3. Missing heritability problem shows that GWAS do not explain much
4. GWAS have methodological flaws

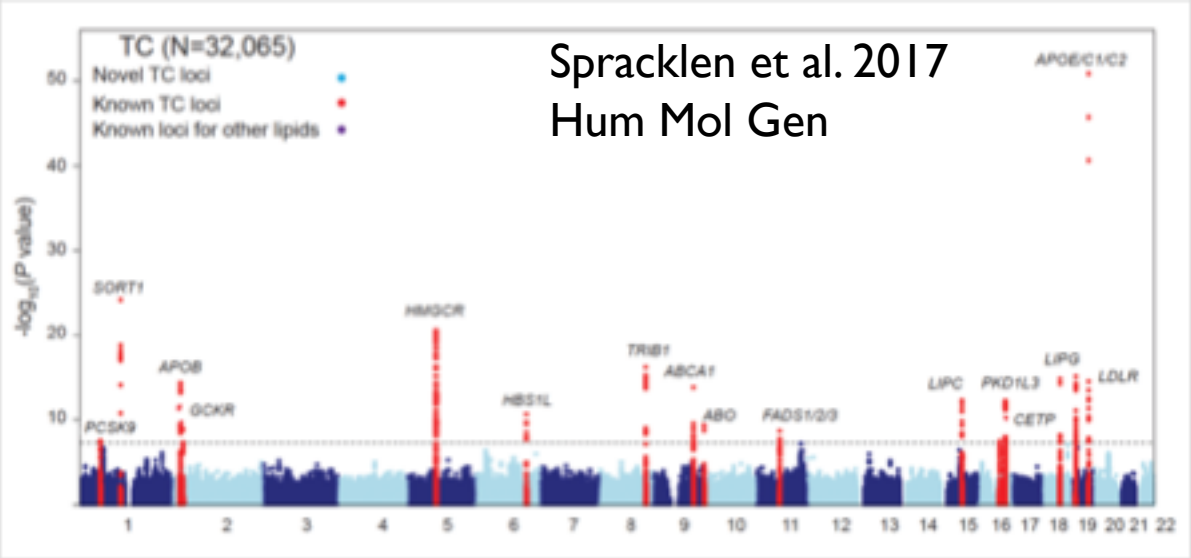
C1. SMALL EFFECTS

- If an effect allele contributes an average of 1 mm to height or 10% to risk of Multiple sclerosis is that of any use?
  - 11/10000 of risk variant carriers get MS while among the non-risk allele carriers the number is 10/10000. This is a tiny difference.

C1. SMALL EFFECTS

- Individual common variants have small effects likely because nature does not tolerate large effects at those loci
- By therapies, it can still be possible to perturb the system by a much larger effects than what exist naturally in a human population
- Effect size in a natural population is not the only, or even the most important, measure of therapeutic potential

"The HMGCR locus has a common variant at 40% frequency that changes LDL by a modest 2.8 mg/dl and no known rare mutations of large effect, presumably because they would be lethal. Yet, the encoded protein is the target of statins, drugs taken by tens of millions of patients that can significantly reduce both LDL levels and myocardial infarction risk." E Lander 2011, Nature

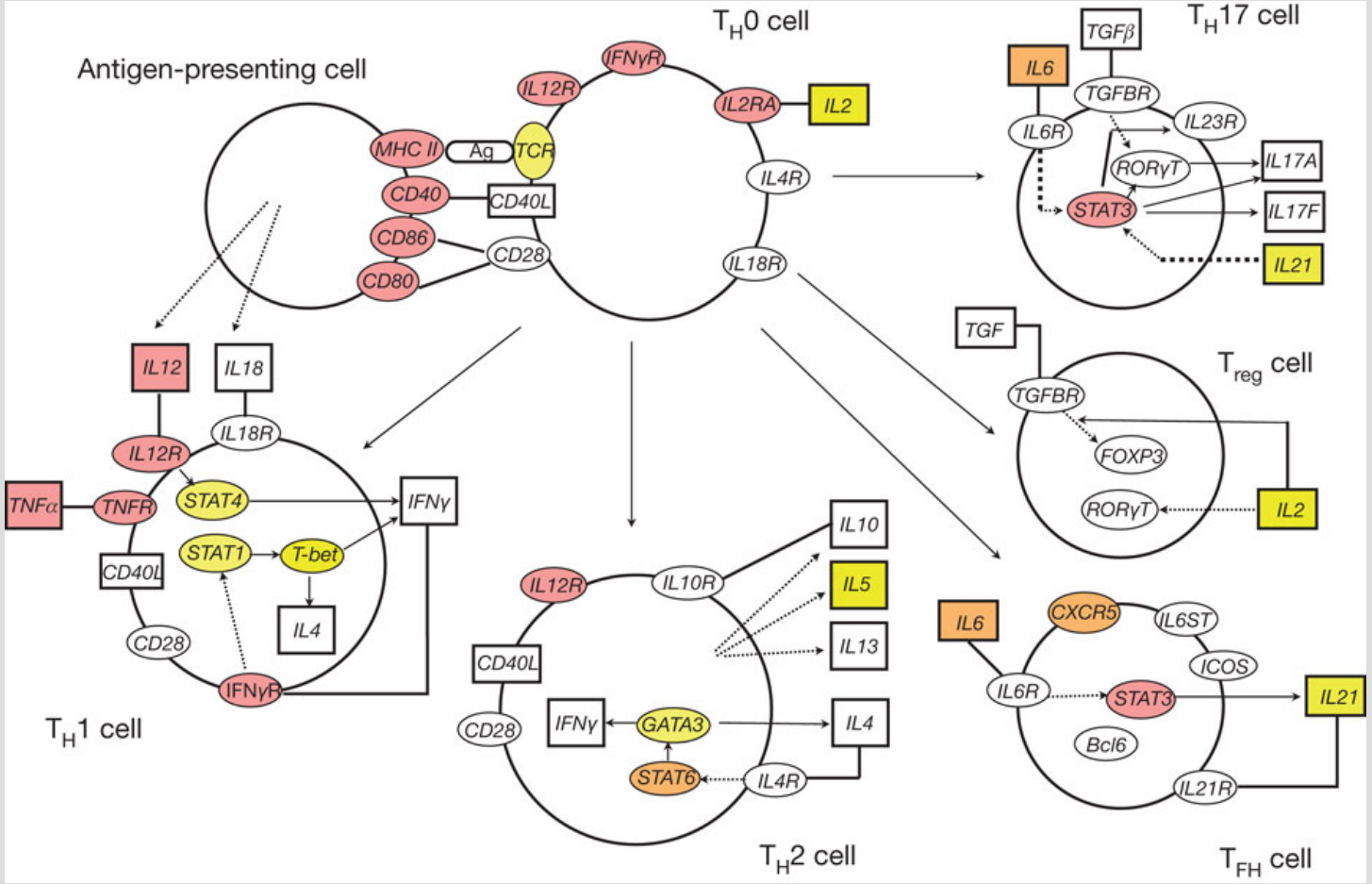


C1. SMALL EFFECTS

Examples of GWAS loci pointing to targets of existing drugs

Multiple sclerosis:
VCAM1 – natalizumab;
IL2RA – daclizumab.

Schizophrenia:
DRD2 - antipsychotic drugs.



Tiny effects can point to pathways.
GWAS of Sawcer et al. Nature 2011
points to immune mechanism in MS (rather than neuronal mechanism). Fig. shows GWAS hits in T-helper cell pathway.

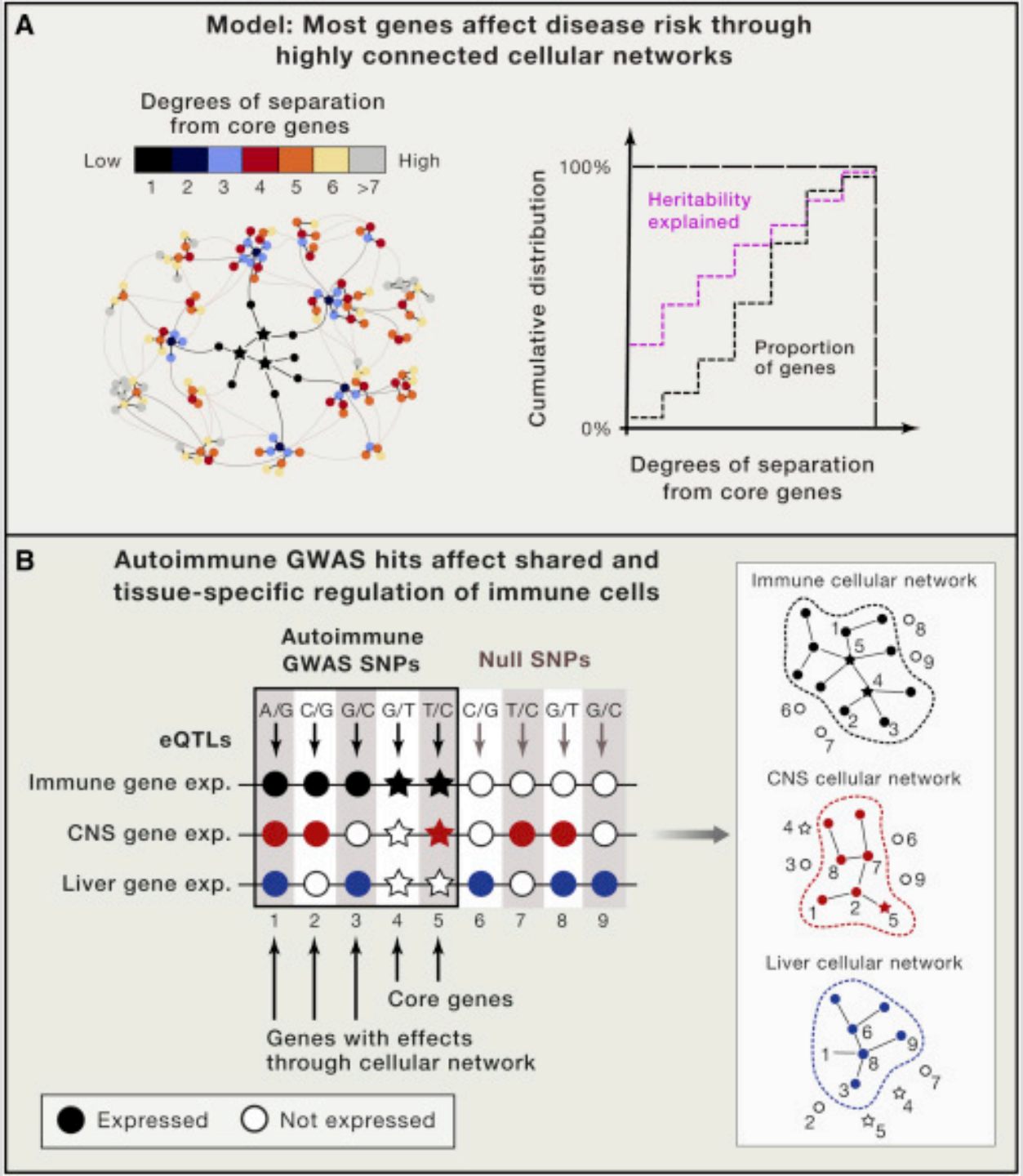
C1. SMALL EFFECTS

- OK, so some top GWAS hits clearly point to biology, but is there a good reason to keep on going or are we eventually associating everything with everything?

Weiss & Terwilliger 2001

OMNIGENIC MODEL (BOYLE ET AL. 2017 CELL)

Intuitively, one might expect disease-causing variants to cluster into key pathways that drive disease etiology. But for complex traits, association signals tend to be spread across most of the genome—including near many genes without an obvious connection to disease. We propose that gene regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes and that most heritability can be explained by effects on genes outside core pathways. We refer to this hypothesis as an "omnigenic" model.



FLATTENING
O'CONNOR ET AL. 2018 BIORXIV

Our results suggest that the omnigenic model is incomplete, and that negative selection can
help explain why core genes explain a limited proportion of trait heritability.

Our results support a potential distinction between core and peripheral genes:
among a large set of genes with similarly strong common-variant associations, a small subset may have much larger phenotypic effects, and
It may be useful to label these as core genes.

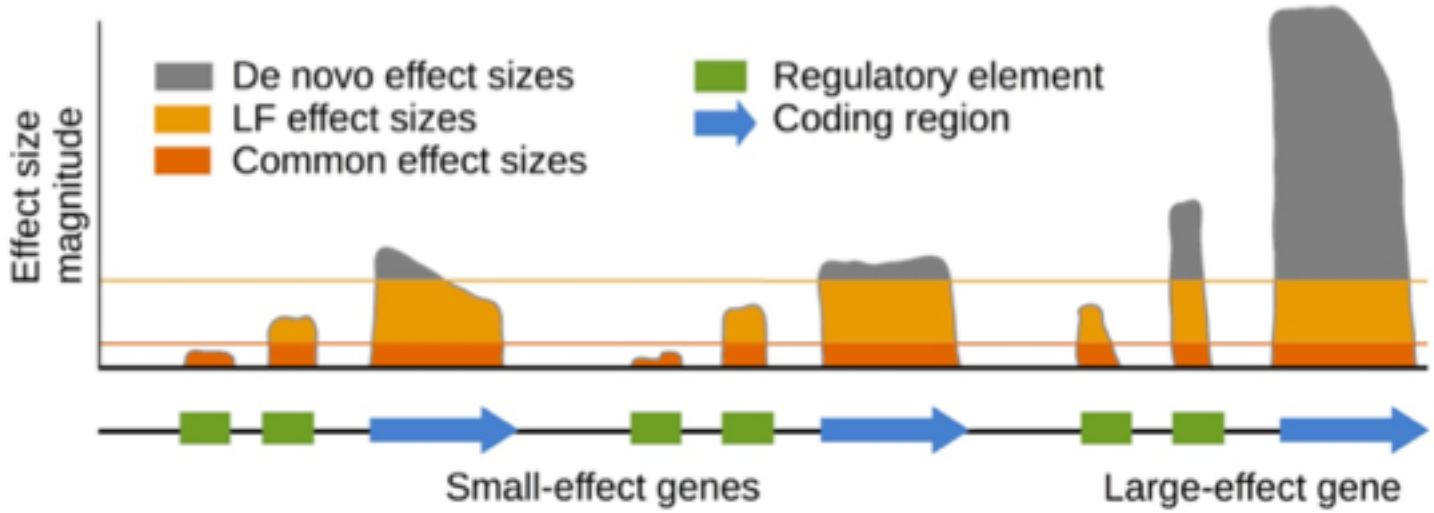


Figure 1. Illustration of flattening due to natural selection. We display the maximum per-allele effect size of a SNP at each site for a toy example of three genes and nearby regulatory regions. Here, the distribution of *de novo* effects is not highly polygenic; it is dominated by coding variants in a single large-effect gene (although other genes also harbor small effects). Negative selection imposes an upper effect size bound (possibly soft), resulting in increased polygenicity. The effect is stronger for common SNPs than for low-frequency SNPs. Within functionally important regions (e.g. coding), a larger proportion of variants have effect sizes near the bound, leading to especially large polygenicity. In practice, this bound may vary across the genome; we expect that the resulting distribution will be more flat than the distribution of *de novo* variants, but not perfectly flat.

C1: ”1000S OF SMALL EFFECTS FROM GWAS ARE NOT INTERESTING BECAUSE WE WOULD HAD LIKED TO FIND ONLY A FEW LARGE EFFECTS”

People say to me, "Are you looking for the ultimate laws of physics?" No, I'm not. I'm just looking to find out more about the world and if it turns out there is a simple ultimate law which explains everything, so be it; that would be very nice to discover. If it turns out it's like an onion with millions of layers and we're just sick and tired of looking at the layers, then that's the way it is!… **And therefore when we go to investigate we shouldn’t pre-decide what it is we are trying to do except to find out more about it**… My interest in science is to simply find out more about the world.

(Richard Feynmann in book *The Pleasure of Finding Things Out* (1999) p. 23.)

C2. MISSING MECHANISMS

- GWAS give merely statistical associations
  - We do not know what are the causal variants (if there are such) and what are their causal mechanisms?
- Answer: With time and (serious) efforts some loci have been explained (next slides). But it won't be quick and easy in near future given how many loci are being identified by GWAS and how little idea we have about many of them
- We need to understand non-coding genome much better to understand GWAS hits

"FROM NONCODING VARIANT TO PHENOTYPE VIA SORT1 AT THE 1P13 CHOLESTEROL LOCUS" BY MUSUNURU ET AL. NATURE 2010

Recent genome-wide association studies (GWASs) have identified a locus on chromosome 1p13 strongly associated with both plasma low-density lipoprotein cholesterol (LDL-C) and myocardial infarction (MI) in humans. Here we show through a series of studies in human cohorts and human-derived hepatocytes that a common noncoding polymorphism at the 1p13 locus, rs12740374, creates a C/EBP (CCAAT/enhancer binding protein) transcription factor binding site and alters the hepatic expression of the SORT1 gene. With small interfering RNA (siRNA) knockdown and viral overexpression in mouse liver, we demonstrate that Sort1 alters plasma LDL-C and very low-density lipoprotein (VLDL) particle levels by modulating hepatic VLDL secretion. Thus, we provide functional evidence for a novel regulatory pathway for lipoprotein metabolism and suggest that modulation of this pathway may alter risk for MI in humans. We also demonstrate that common noncoding DNA variants identified by GWASs can directly contribute to clinical phenotypes.

”SCHIZOPHRENIA RISK FROM COMPLEX VARIATION OF COMPLEMENT COMPONENT 4” BY SEKAR ET AL. 2016 NATURE

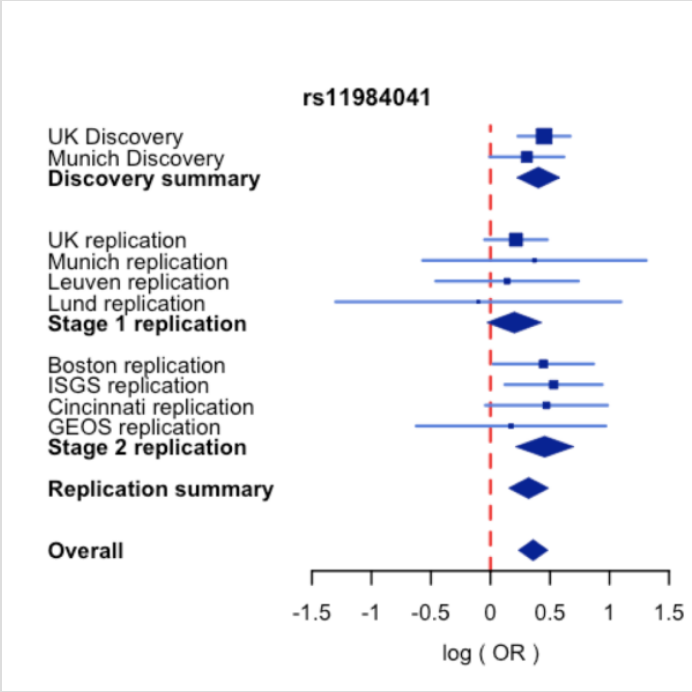
Schizophrenia’s strongest genetic association at a population level involves variation in the major histocompatibility complex (MHC) locus, but the genes and molecular mechanisms accounting for this have been challenging to identify. Here we show that this association arises in part from many structurally diverse alleles of the complement component 4 (*C4*) genes. We found that these alleles generated widely varying levels of *C4A* and *C4B* expression in the brain, with each common *C4* allele associating with schizophrenia in proportion to its tendency to generate greater expression of *C4A*. Human C4 protein localized to neuronal synapses, dendrites, axons, and cell bodies. In mice, C4 mediated synapse elimination during postnatal development. These results implicate excessive complement activity in the development of schizophrenia and may help explain the reduced numbers of synapses in the brains of individuals with schizophrenia.

C3. MISSING HERITABILITY

- As we saw in GWAS 8, most of the missing heritability (i.e. the gap between heritability estimates from twin studies and heritability explained by GWAS loci) could be explained by small effect sizes that have not been identified in GWAS with the stringent threshold.
  - Linear mixed model can identify heritability close to twin estimates in quantitative traits when whole genome data are used (Yang et al. 2015)
  - So we might not miss that much heritability but we miss the exact information of causal variants
- We do not yet cover well rare variants in our heritability estimates. Individually rare variants contribute little, but there are a lot of them
- The estimates of heritability from twin studies may be biased upwards due to non-additive effects that jump in to the additive estimates from closely related individuals (Zuk et al. PNAS 2012)

C4. "FLAWED" DESIGN OF GWAS

- **Population structure** can be handled with PCs and mixed models (although never perfectly)
- Differences in **genotyping procedures** between cases and controls must be accounted for (and this is not always easy)
- In large **meta-analyses**, unlikely that bias would be consistent in most studies
- **Replication** in (several) other cohorts provides convincing evidence
- **Stringent statistics** (P<5e-8) is applied
  - An order of magnitude more stringent threshold than physicists used to declare Higgs' boson found in 2012



Nat Gen 2012 44(3):328-33