Genome-wide association studies II: Identifying genetic associations with complex traits

Gavin Band gavin.band@well.ox.ac.uk

MSc Global Health Science and Epidemiology

Genetic Epidemiology Module

Feb 2023

Learning objectives

Understand a genome-wide association study (GWAS) and the concept of a hypothesis-free approach to studying genetic associations.

Have a working knowledge of the different steps involved in the conduct of GWAS, including study design, quality control and basic analyses.

Be able to interpret and critically appraise evidence from genome-wide association studies.

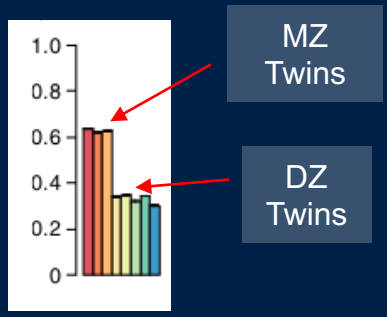
Understand the relevance of replication, meta-analysis and consortia, and multi-ancestry approaches, in genome-wide association studies.

Appreciate the use of post-GWAS analyses including fine mapping, gene and pathway analyses, and the concept of causal variants.

Lecture plan

- Recap & fallout from last lecture
- Gaining biological knowledge from GWAS
- Biological examples
- Heritability and prediction

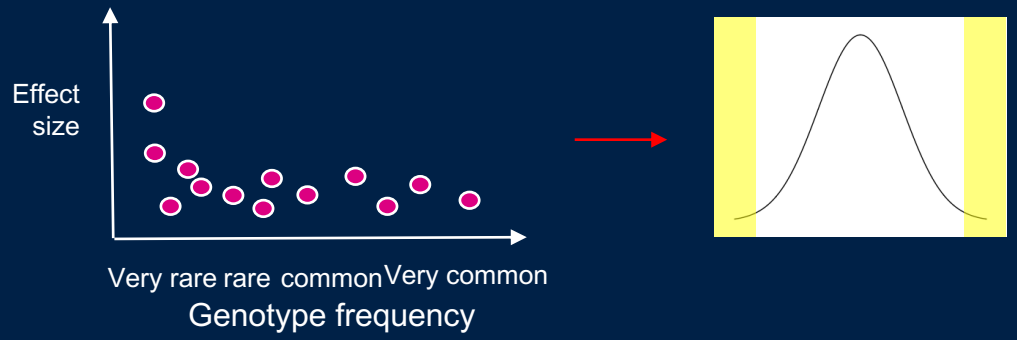
Recap



1. Most human traits are highly heritable

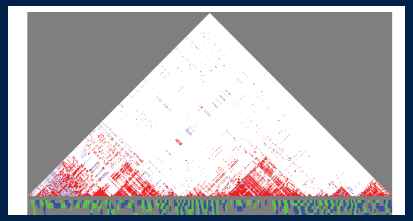
A large proportion of population variation is explained by genetics

2. For many 'complex' traits, this is caused by lots of variants with small effects



3. To find these genetic variants, we can use genome-wide association study methodology.

e.g. genotype cases and controls at a dense set of markers across the genome, and do a statistical test of association. Relies on block-like structure of LD to access untyped variants.

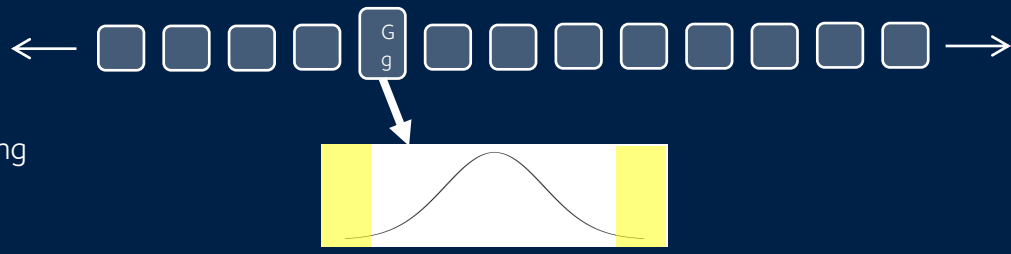


Aim to uncover the underlying biology of disease.

Last time – basic GWAS approach

Basic idea: try to find **causal** effects of genetic variants on phenotypes.

Many traits are heritable but *complex*: caused by many genetic variants with small effects across the genome (along with environmental factors, interactions, …)



Strategy: use genome-wide genotyping and imputation to access as much genetic variation as possible. For a disease phenotype, a case-control (or population control) design then allows us to directly estimate the relative risk of each variant.

$$\textit{Relative risk} = \frac{P(\text{disease}|\text{genotype } G)}{P(\text{disease}|\text{genotype } g)}$$

Measures the association between genotype and phenotype.
Estimated as an odds ratio in the study

The *accuracy of our estimates*, and the *power to detect nonzero effects*, depends mainly on the *sample size* and the *frequency of the variant:*

$$\text{Standard error}(\log OR) \approx \frac{1}{\sqrt{N \times f(1-f) \times \phi(1-\phi)}}$$

Sample size

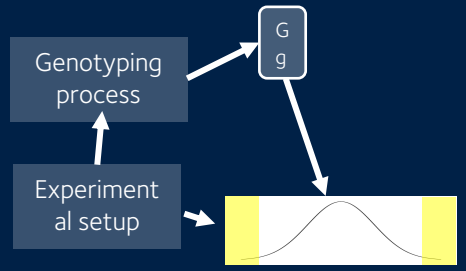
Genotype frequency

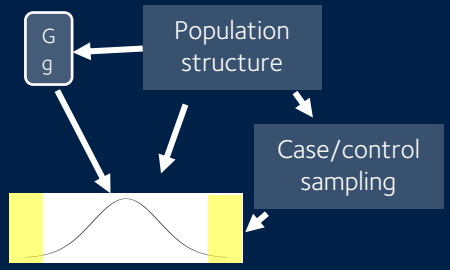
Ratio of cases to controls in study

Three potential problems

Case-control designs do not control for confounding – this has to be done in the analysis stage.
Association picks up all 'causal' paths from genotype to phenotype.
There are at least three important ways the study could be confounded:

Experimental confounding – for example, differential genotyping between cases and controls.

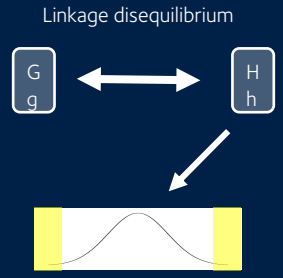




Confounding by population structure – for example, if the sampling structure, or the true distribution of the phenotype, happens to covary with genetic background

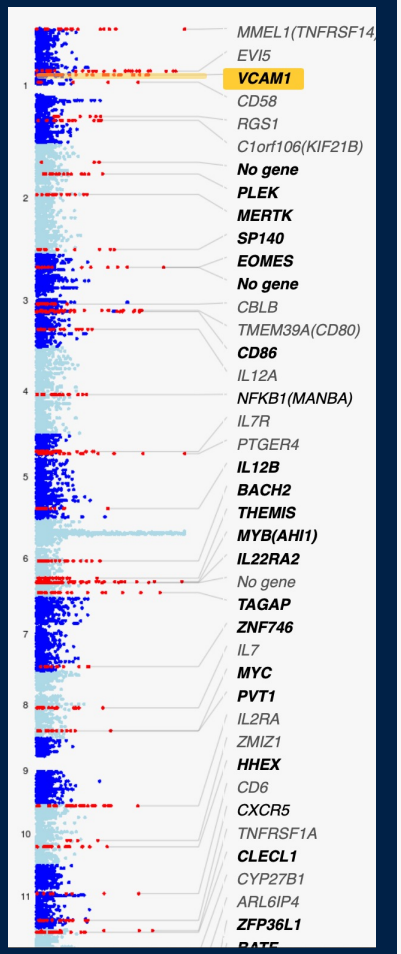
Confounding by LD

Nearby variants are correlated (in linkage disquilibrium) because of population genetic drift broken down by recombination. This makes it easier to detect association, but harder to narrow down to the actual causal variant.



Consolidation question from last lecture

WTCCC2 GWAS of multiple sclerosis (9,772 cases and 7,376 controls).



For further information about terms used below, hover over the red question marks.

Region

dbSNP id:?	rs11581062
status:?	novel association
physical position:?	01:101,180,107
association region:?	01:100,983,315-101,455,310
functional tag:?	*N/A*
nearest gene:?	*SLC30A7*
candidate gene:?	*VCAM1**

Signal

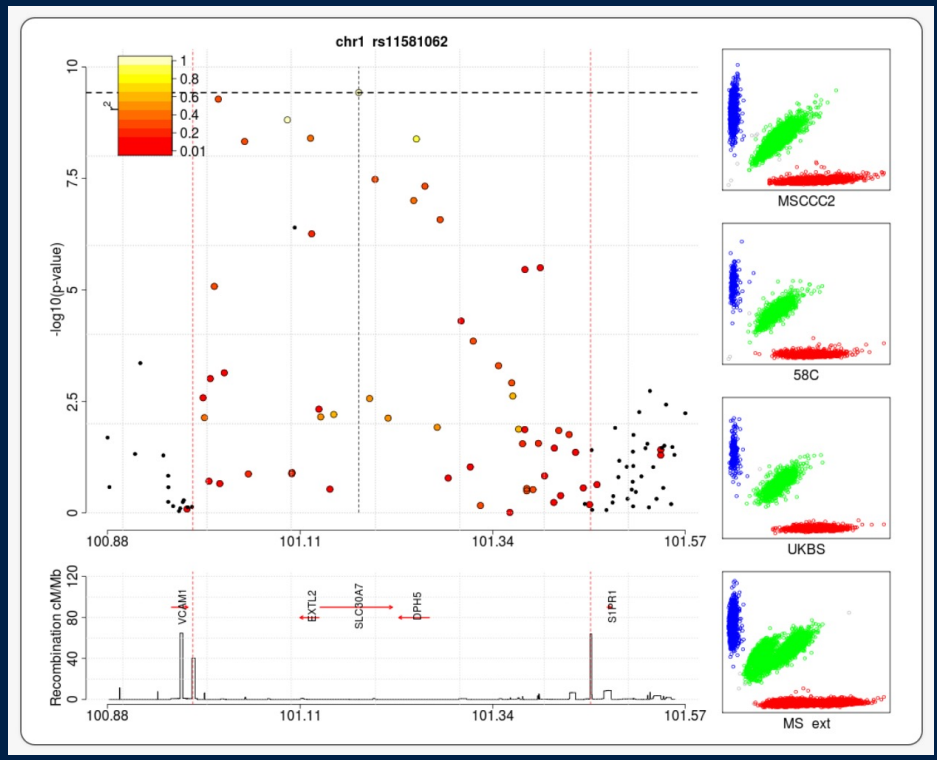
p-value discovery:?	3.7e-10
OR discovery (95% CI):?	1.13 (1.09-1.18)
p-value replication:?	4.20e-02 (one-sided)
OR replication (95% CI):?	1.07 (0.99-1.15)
p-value combined:?	2.50e-10
OR combined (95% CI):?	1.12 (1.1-1.13)
Risk (non-risk) allele:	G(A)

Allele frequencies?

Country	controls / cases	control / case frequency
Australia	- / 647	- / 0.32
Belgium	- / 544	- / 0.33
Denmark	- / 332	- / 0.32
Finland	2165 / 581	0.23 / 0.24
France	347 / 479	0.31 / 0.34
Germany	1699 / 1100	0.29 / 0.31
Ireland	- / 61	- / 0.34
Italy	571 / 745	0.30 / 0.33
Norway	121 / 953	0.26 / 0.28
Poland	- / 58	- / 0.27
Spain	- / 205	- / 0.36
Sweden	1928 / 685	0.27 / 0.28
UK	5175 / 1854	0.29 / 0.32
USA	5370 / 1382	0.29 / 0.32

Proximal genes?

DPH5, *EXTL2*, *S1PR1*, *SLC30A7*, *VCAM1**



Can you explain?

Anatomy of an association analysis

All GWAS should report data in a way that can be re-used by future studies.
This study used several previous GWAS to conduct replication. All the details are given in a supplementary table:

		WAS + replicatio		GWAS			UK only GWAS		on-UK only GWAS		combined replication		eneMSA NL replicatio			ene MSA US replicatio			eneMSA CH replicatio			ANZ replication			BWH replication		
Gene	RiskAlle	pval	OR (95% CI	pval	OR (95% CI	log10 (BayesFacto	pval	OR (95% CI	pval	OR (95% CI	pval**	OR (95% C	pval**	OR (95% CI	inf	pval**	OR (95% CI	inf	pval**	OR (95% CI	inf	pval**	OR (95% CI	inf	pval	OR (95% CI	inf
MMEL1	C	1.00E-14	1.14 (1	3.10E-14	1.16 (1	11.39	0.0073	1.12 (1	7.10E-13	1.17 (1	0.0085	1.08 (1.01-1.15)	0.26	1.1 (0.	0.94	0.18	1.1 (0.	1.01	0.24	1.11 (0	1.03	0.006	1.15 (1	1	0.41	1.02 (0	
EVI5	A	5.80E-15	1.15 (1	6.50E-12	1.15 (1	9.15	2.90E-05	1.2 (1.	2.70E-08	1.14 (1	1.00E-04	1.14 (1.06-1.22)	0.088	1.23 (0	1.05	0.59	0.97 (0	0.91	0.71	0.92 (0	0.94	0.023	1.12 (1	0.97	0.0059	1.18 (1	1
SLC30A7	G	2.50E-10	1.12 (1	3.70E-10	1.13 (1	7.43	0.00047	1.16 (1	1.70E-07	1.13 (1	0.042	1.07 (0.99-1.15)	0.57	0.99 (0	1.01	0.095	1.09 (0	0.99	0.013	1.18 (0	0.91	0.57	0.99 (0	1.01	0.095	1.09 (0	0
EXTL2	*A*	*4.00E-08*	*1.09 (1*	*3.70E-07*	*1.1 (1.*	*4.52*	*0.00096*	*1.14 (1*	*6.00E-05*	*1.08 (1*	*0.017*	*1.08 (1.01-1.15)*	*0.025*	*1.11 (1*	*1*	*0.088*	*1.09 (0*	*1.01*	*0.45*	*1.01 (0*	*0.88*	*0.025*	*1.11 (1*	*1*	*0.088*	*1.09 (0*	*1.*

Discovery and overall data as on web page

Evidence for the same effect direction was seen separately in both arms of the discovery…

…and in the combined replication…

…and in most of the individual replication studies.

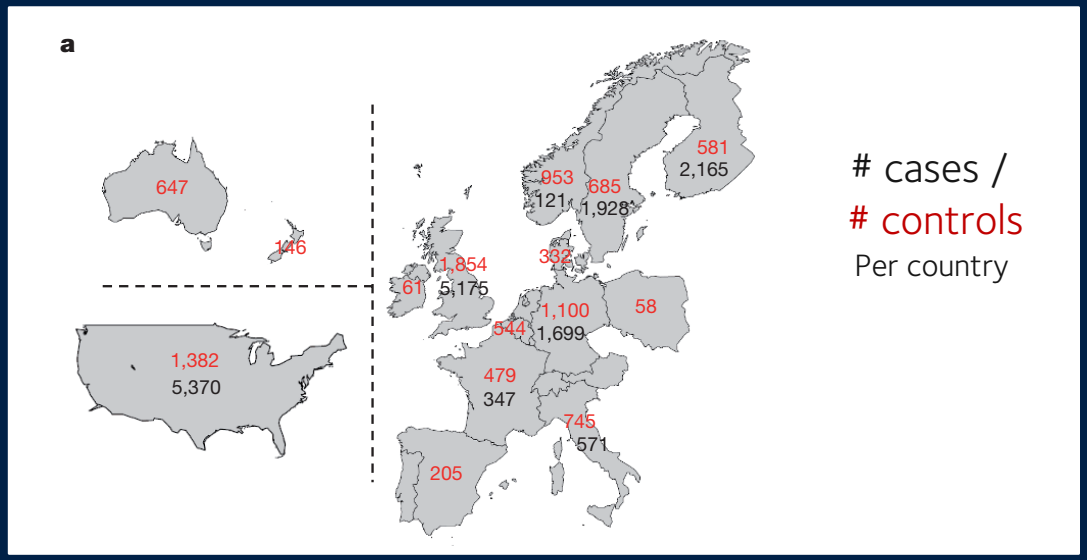
This is a common analysis approach: to gain sample size, use meta-analysis to combine results across several component studies. Then look for consistency between the studies.

$$v_{meta} = 1/\left(\sum_i \frac{1}{v_i}\right) \qquad \beta_{\text{meta}} = \left(\sum_i \frac{\beta_i}{v_i}\right) \times v_{meta}$$

(Where v denotes squared standard error)

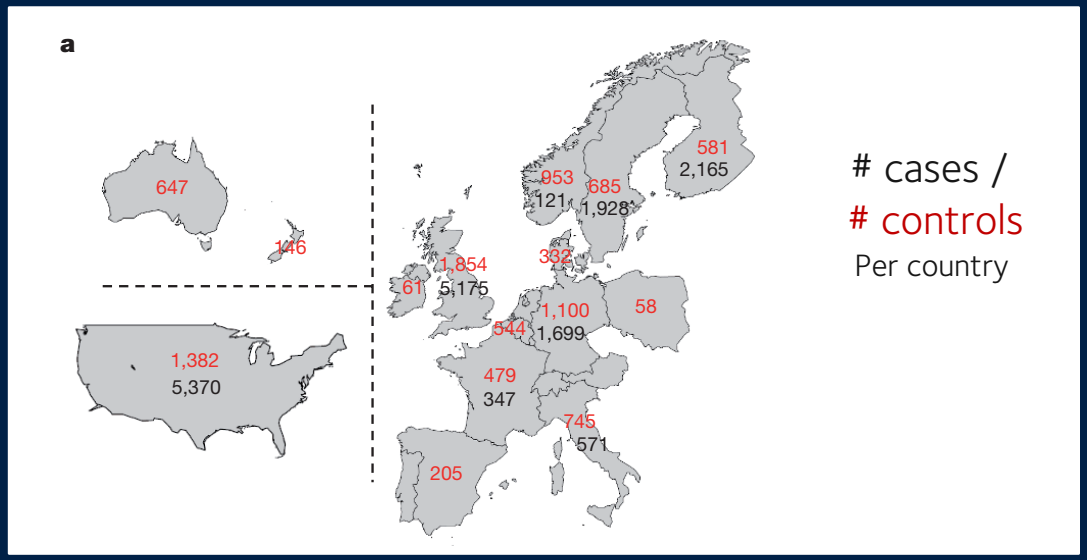
" Inverse variance weighted fixed-effect meta-analysis", gives results approximately equal to joint analysis of genotype data.

Dealing with population structure

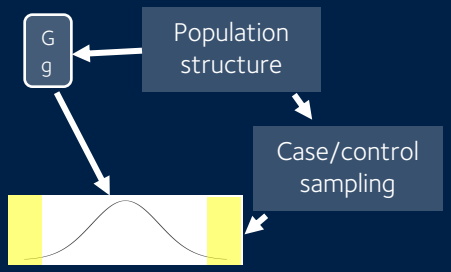


This study suffered from a key problem. Can you see what it is?

Dealing with population structure

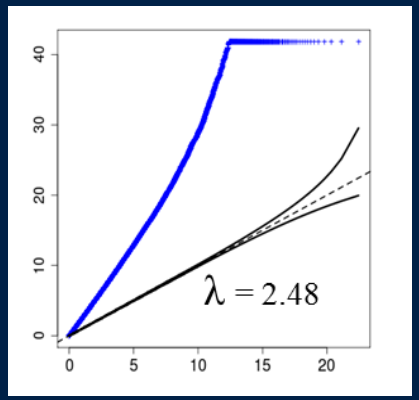


Answer: very strong confounding by population structure / sampling

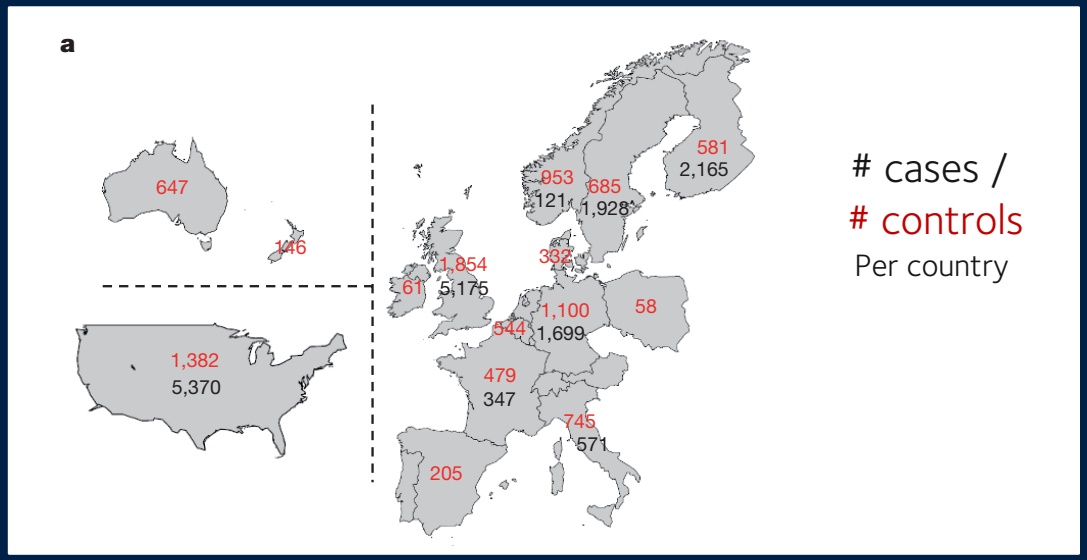


This is a quantile-quantile plot of all association tests genome-wide. It shows vastly inflated $-\log_{10}$ P-values.

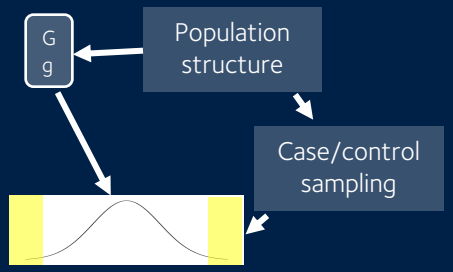
(A more advanced way to do this distinguishing structure from polygenicity is *LD score regression* – covered in a later lecture).



Dealing with population structure

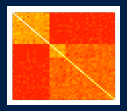


Answer: very strong confounding by population structure / sampling



Solution:

1. Use genome-wide genotypes to estimate genetic relatedness between samples

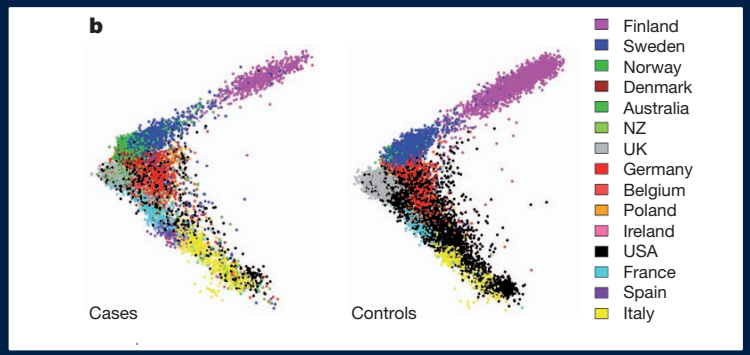


2. Include the relatedness as a covariate in the association test

Using regression to test for association (instead of the 2x2 table method)

1. Logistic regression including principal components

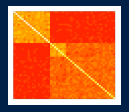
$$\text{outcome} \sim \text{genotype} + PCs$$



Plot of first two principal components obtained from the genetic relatedness matrix

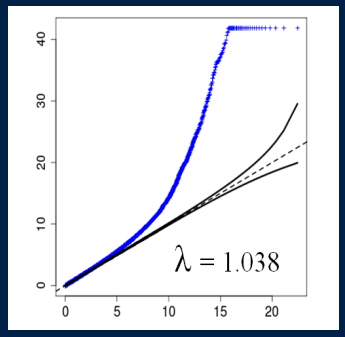
Uses just the strongest directions of variation in relatedness (population structure)

2. Linear mixed model

$$\text{outcome} \sim \text{genotype} +$$ 

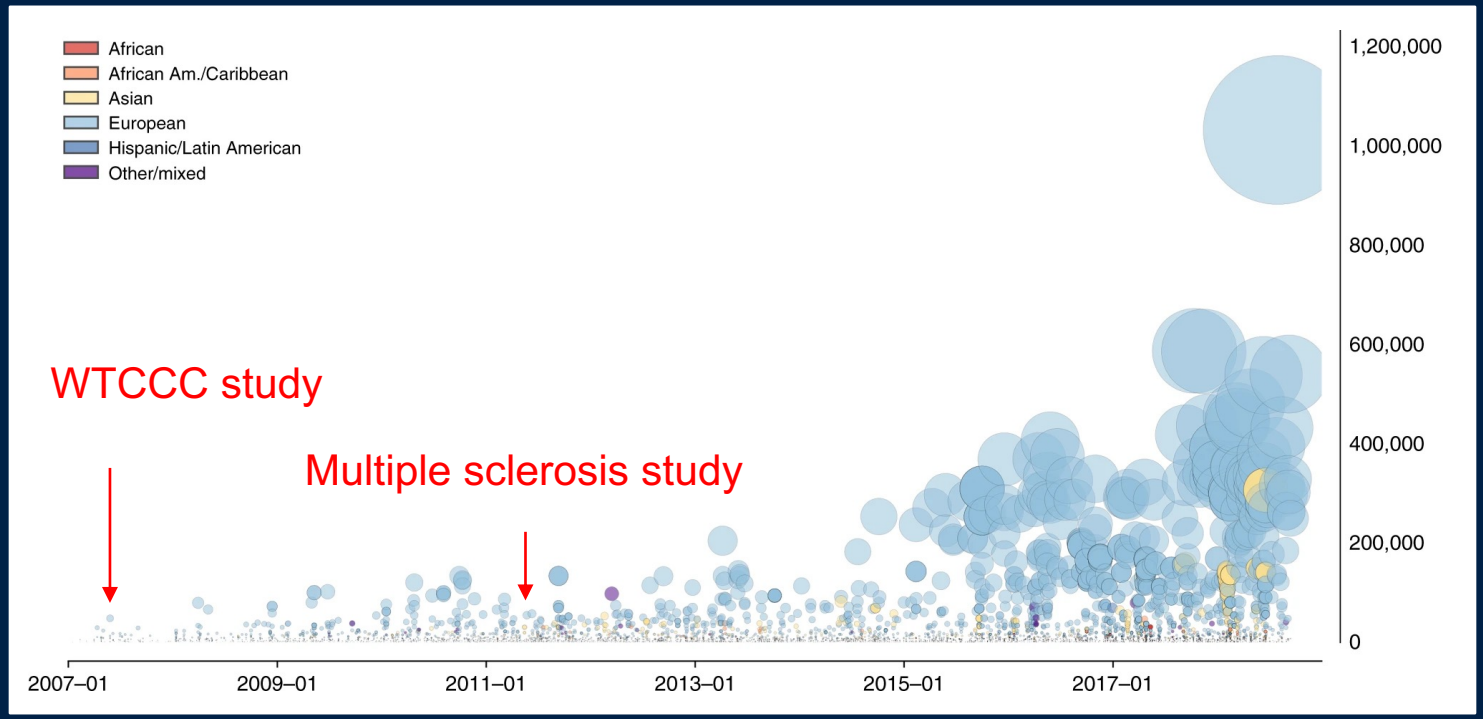
Include a **genetic relatedness matrix computed from genome-wide genotypes** in the association test

Uses the **entire matrix** of relationships



Most p-values are now not inflated

GWAS revolution

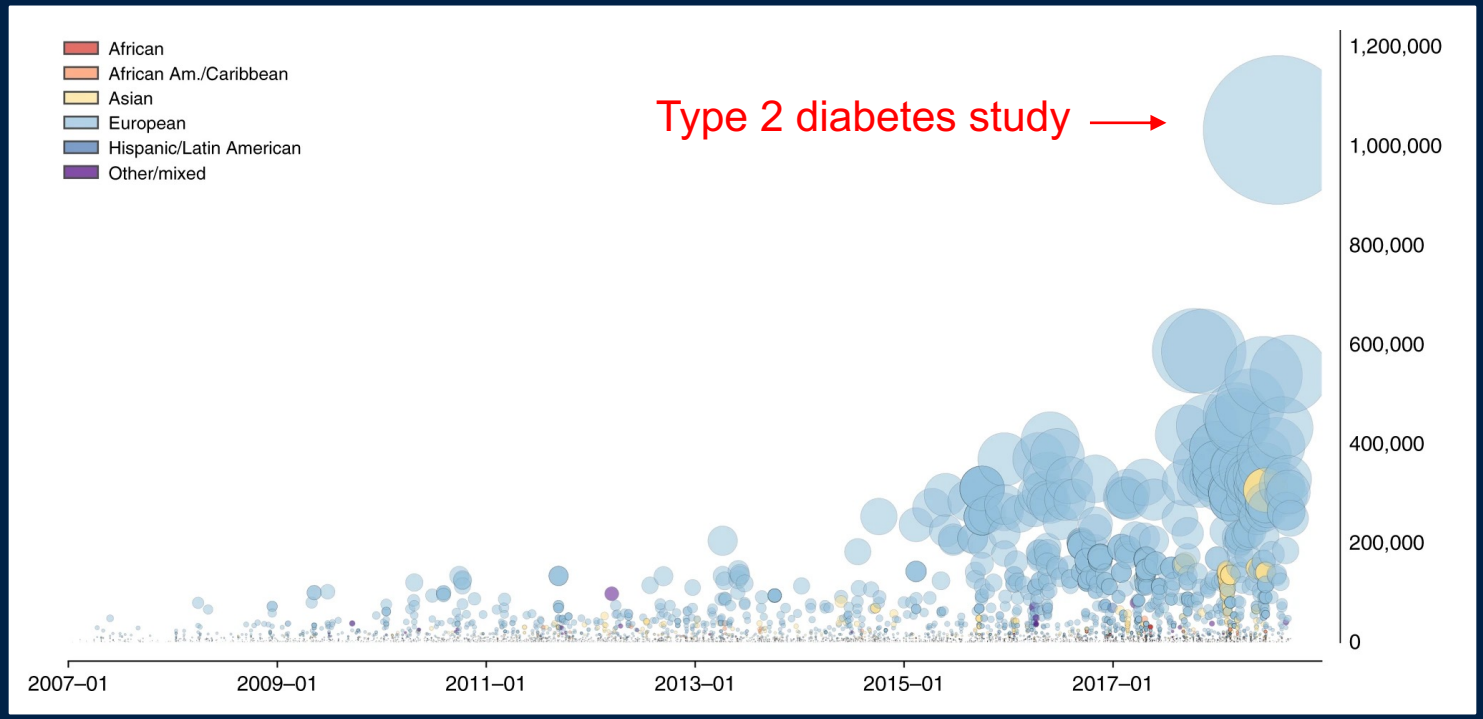


Mills & Rahal, "**A scientometric review of genome-wide association studies"**, Communications Biology 2019

Lecture plan

- Recap & fallout from last lecture
- Gaining biological knowledge from GWAS
- Uncovering biology: examples
- Pleiotropy, heritability and prediction

GWAS revolution



Mills & Rahal, "**A scientometric review of genome-wide association studies"**, Communications Biology 2019

Type 2 diabetes study

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0241-6

Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps

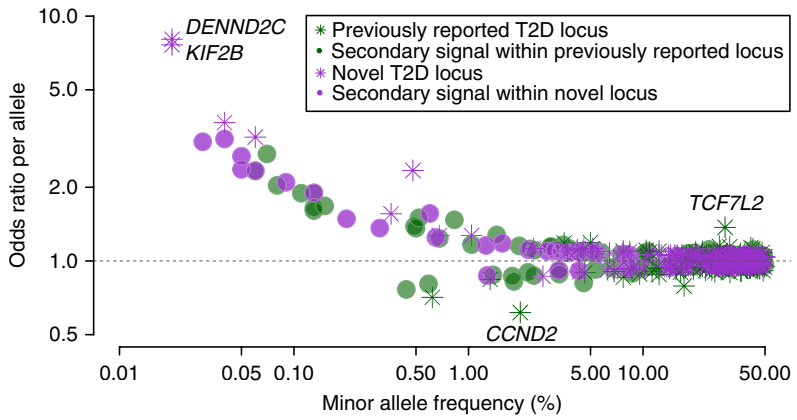


Fig. 5 | The relationship between effect size and MAF. Conditional- and joint-analysis effect size (*y* axis) and MAF (*x* axis) for 403 conditionally independent SNPs. Previously reported T2D-associated variants are shown in green, and novel variants are shown in purple. Stars and circles represent the 'strongest regional lead at a locus' and 'lead variants for secondary signals', respectively.

N = 74,000 T2D cases
And 824,000 controls

Have gone from a handful of T2D signals in 2007 to **403** in 2018

These loci give a detailed view of the 'genetic architecture' of this trait.

Type 2 diabetes study

But finding biology is hard

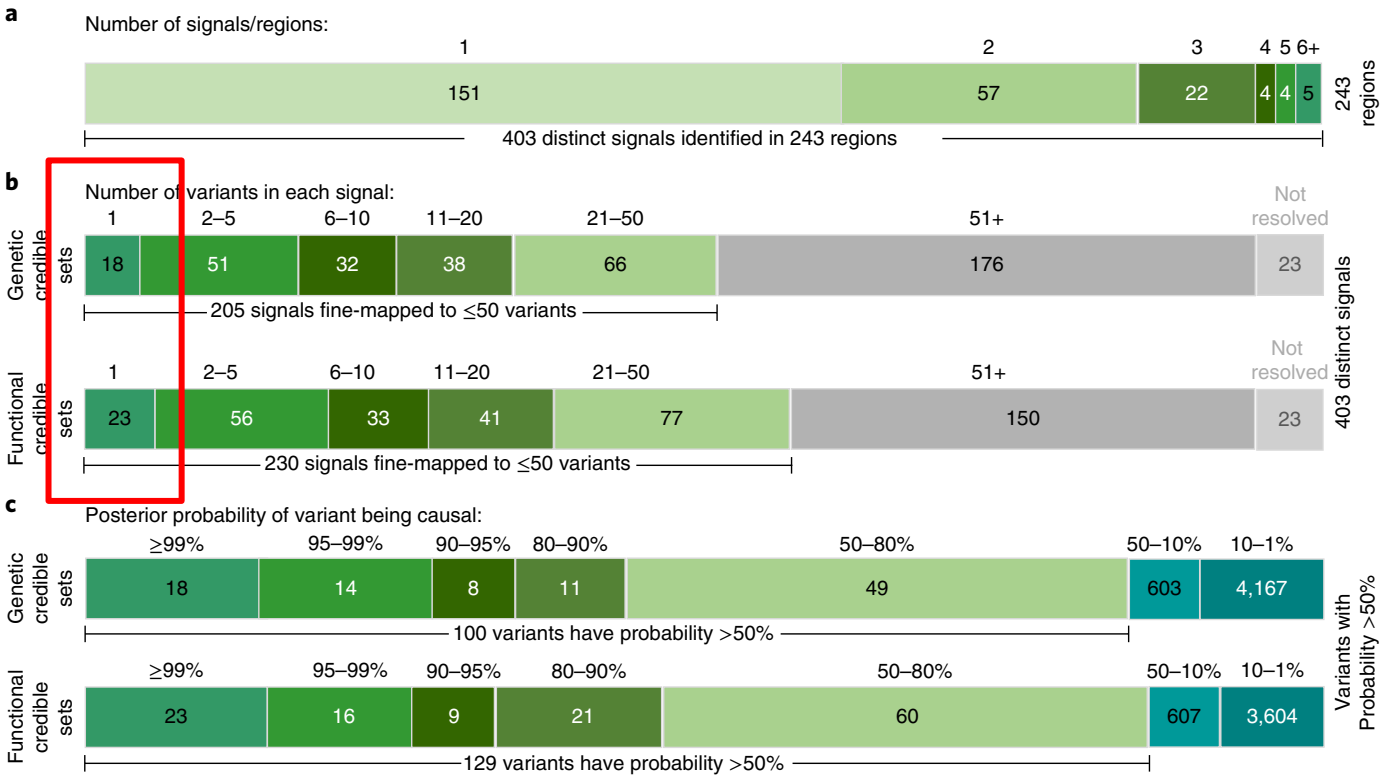


Fig. 3 | Summary of fine-mapped associations. a, Distinct association signals. A single signal at 151 loci, and two to ten signals at 92. **b**, Number of variants in genetic and functional 99%-credible sets. Eighteen and 23 signals were mapped to a single variant in genetic and functional credible sets, respectively. **c**, Distribution of the PPA of the variants in credible sets. Four of the 51 variants with PPA >80% in the genetic credible sets have lower PPAs in the functional credible set, thus giving a total of 73 variants with PPA >80% in either.

Even using this large sample, and exploiting functional data in relevant cell types, only a handful of these signals could be unambiguously mapped to individual variants.

Another example - IBD


ARTICLE

doi:10.1038/nature22969

Fine-mapping inflammatory bowel disease loci to single-variant resolution

Hailiang Huang[1,2]§, Ming Fang[3,4]*, Luke Jostins[5,6]*, Maša Umićević Mirkov[7], Gabrielle Boucher[8], Carl A. Anderson[7],

Huang et al Nature 2017

Attempted fine-mapping of 139 signals of association with inflammatory bowel disease (IBD), using genotype data on 67,852 individuals, and data on the functional state in relevant cell types.

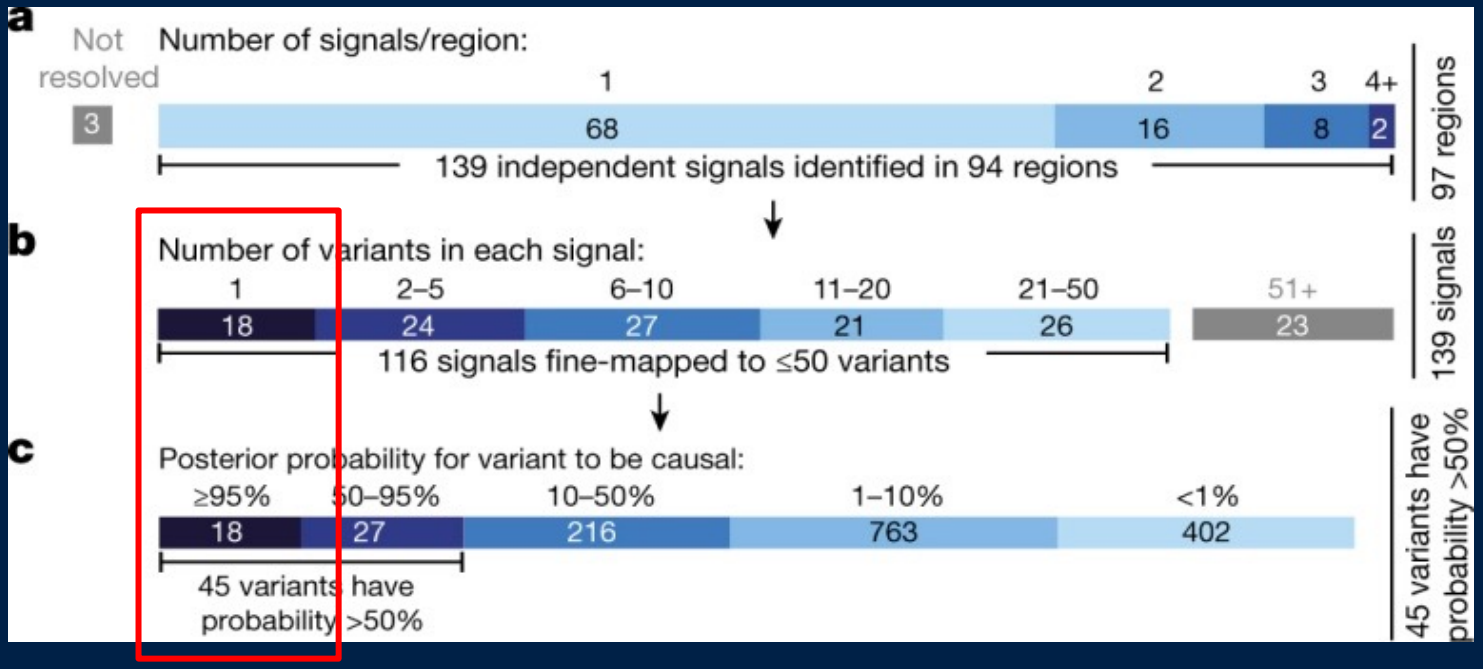
...with mixed success:

Among 45 likely causal variants:

13 protein-coding changes

3 = disruption of transcription factor binding

10 = tissue specific epigenetic marks



At least 21 loci could not be assigned a plausible function despite the extensive data.

Another example - IBD

ARTICLE doi:10.1038/nature22969

Fine-mapping inflammatory bowel disease loci to single-variant resolution

Hailiang Huang[1,2]§, Ming Fang[3,4]*, Luke Jostins[5,6]*, Maša Umićević Mirkov[7], Gabrielle Boucher[8], Carl A. Anderson[7],

Huang et al Nature 2017

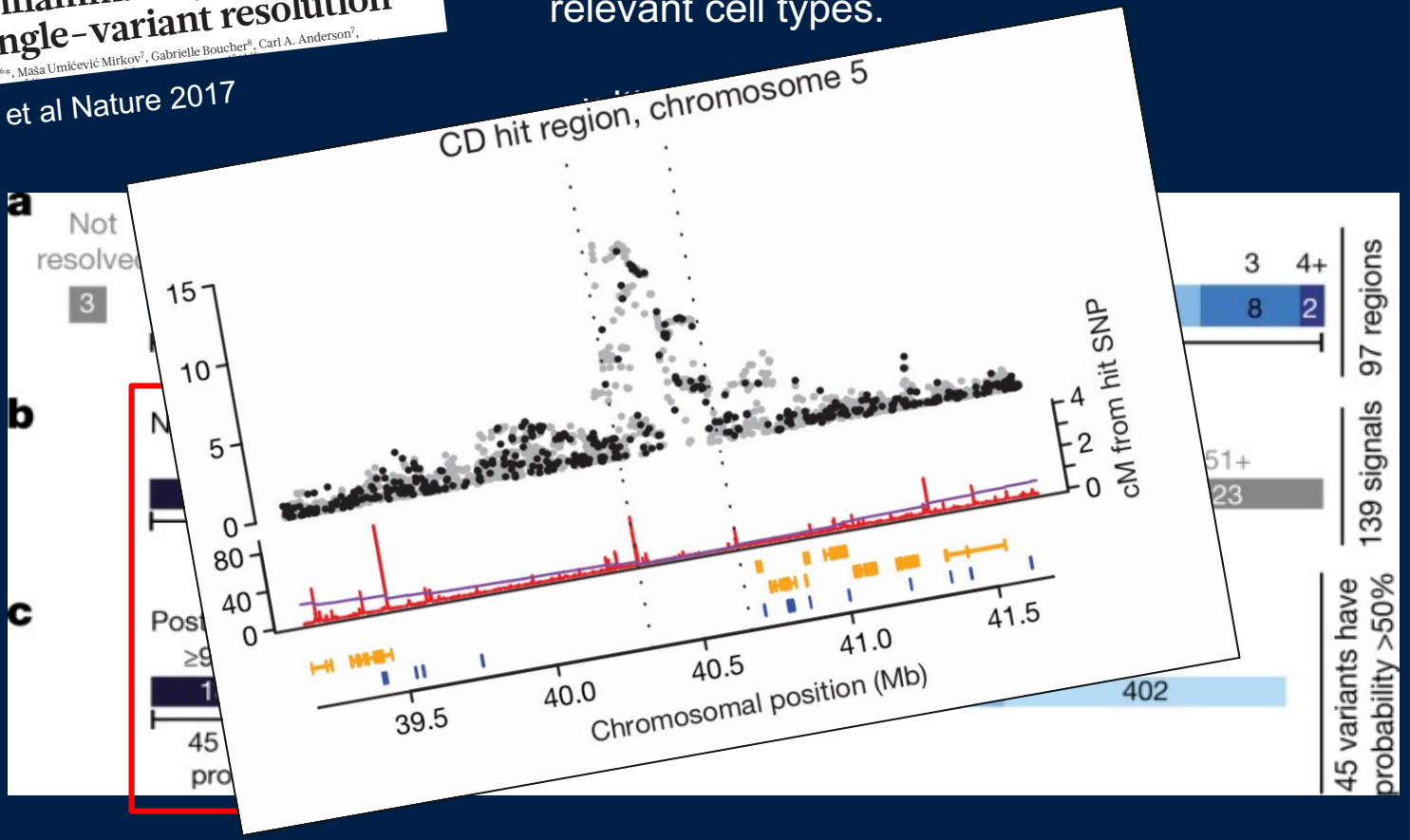
Attempted fine-mapping of 139 signals of association with inflammatory bowel disease (IBD), using genotype data on 67,852 individuals, and data on the functional state in relevant cell types.

Among 45 likely causal variants:

13 protein-coding changes

3 = disruption of transcription factor binding

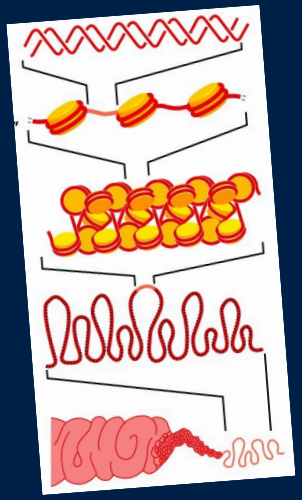
10 = tissue specific epigenetic marks



At least 21 loci could not be assigned a plausible function despite the extensive data.

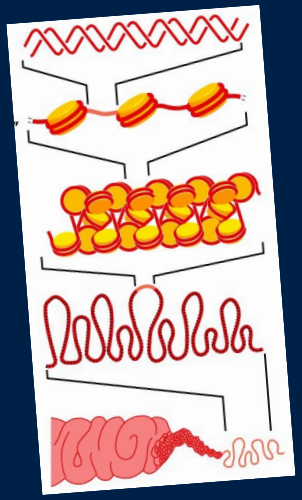
The circle of genetic causation

DNA gets physically packaged up into chromosomes...

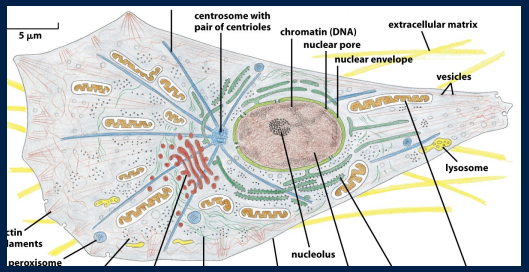


The circle of genetic causation

DNA gets physically packaged up into chromosomes...

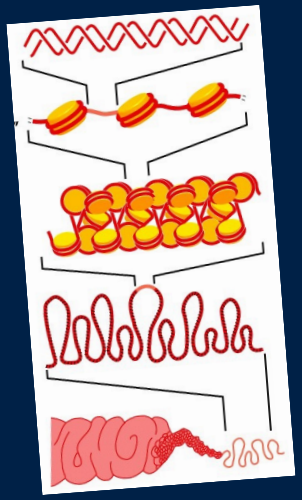


...inside cells, where it is **transcribed** to form proteins and other molecules...

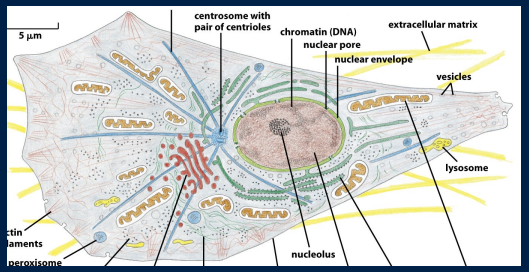


The circle of genetic causation

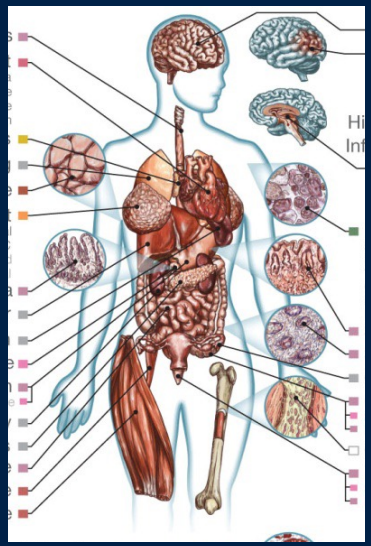
DNA gets physically packaged up into chromosomes...



...inside cells, where it is **transcribed** to form proteins and other molecules...



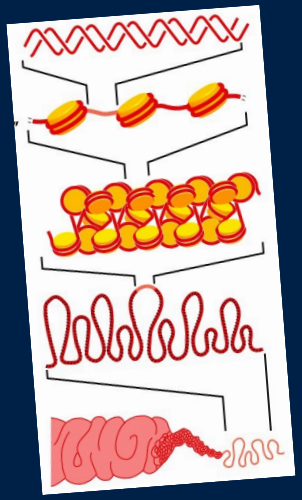
...that affect how the cells behave, forming different organs...



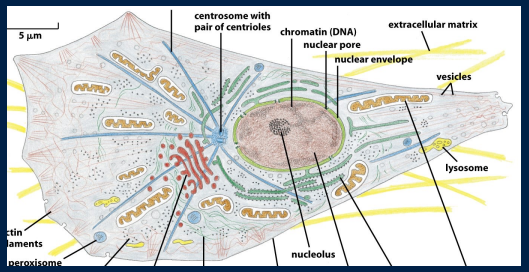
...that combine to make individuals...

The circle of genetic causation

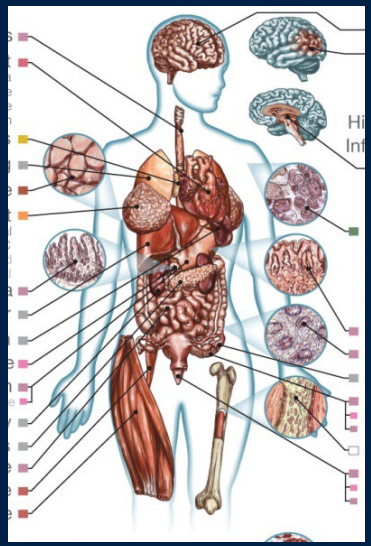
DNA gets physically packaged up into chromosomes...



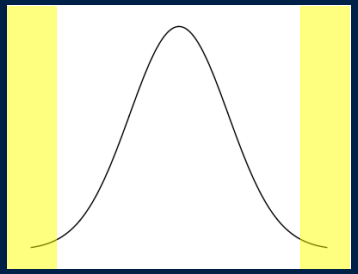
...inside cells, where it is **transcribed** to form proteins and other molecules...



...that affect how the cells behave, forming different organs...



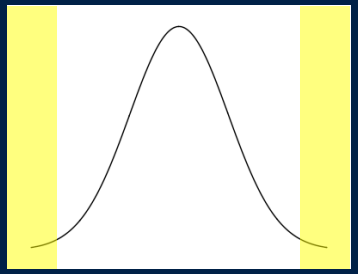
...that combine to make individuals...



...whose success is affected by the traits they have...

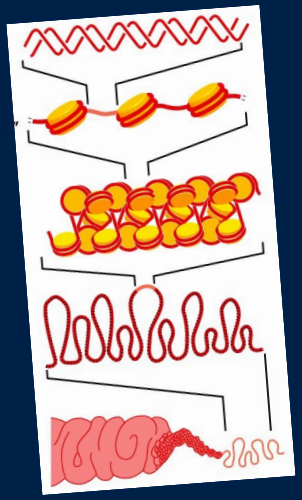
The circle of genetic causation

...passing on DNA, with **mutations** and **recombination**, to new generations...

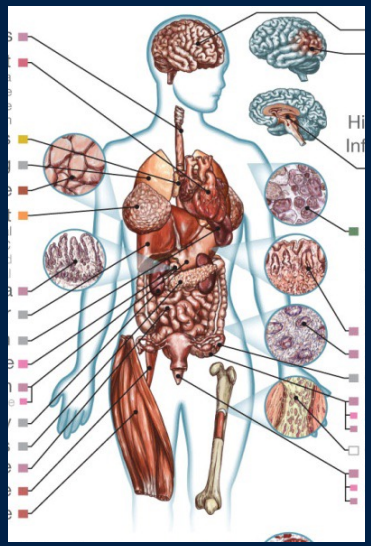


...whose success is affected by the traits they have...

...that gets physically packaged up into chromosomes...

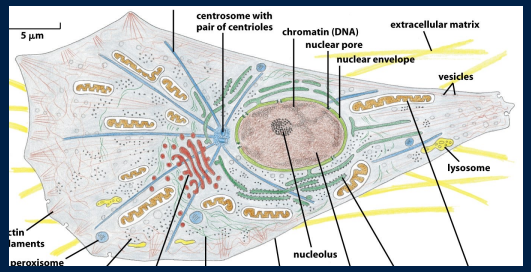


There is complex biology at all stages



...that combine to make individuals...

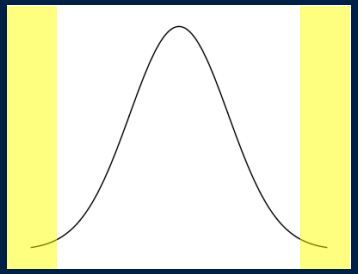
...inside cells, where it is **transcribed** to form proteins and other molecules...



...that affect how the cells behave, forming different organs...

The circle of genetic causation

...passing on DNA, with **mutations** and **recombination**, to new generations...



...whose success is affected by the traits they have...

microarrays, genome sequencing

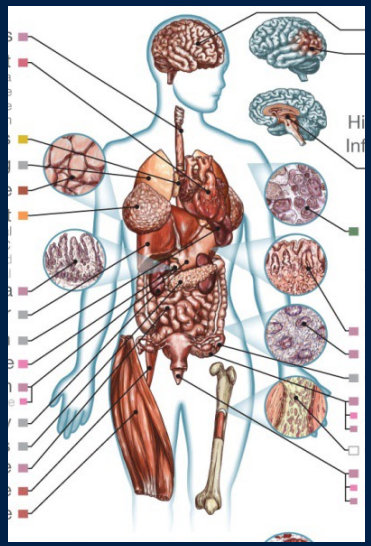
Clinical phenotype measurements

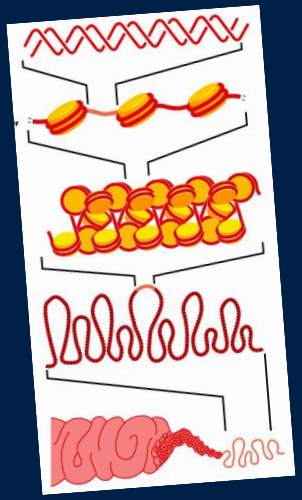
...that gets physically packaged up into chromosomes...

There is complex biology at all stages

And we can measure it.

Biomarker measurements



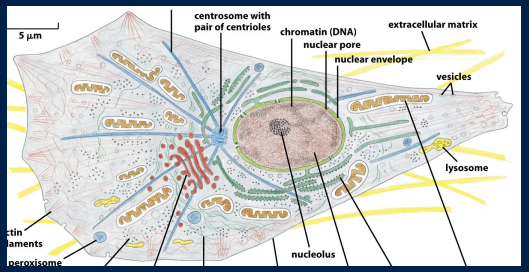


Chromatin state marker assays, ChIP-seq, ...

RNA-seq, spectroscopy, antibody binding

...that combine to make individuals...

...inside cells, where it is **transcribed** to form proteins and other molecules...



...that affect how the cells behave, forming different organs...

Gaining biological knowledge from GWAS

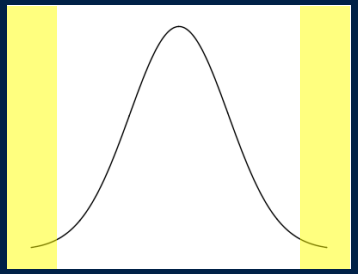
There are several ways we can try to translate knowledge of associations into new biological insights. I will try to describe a few of these.

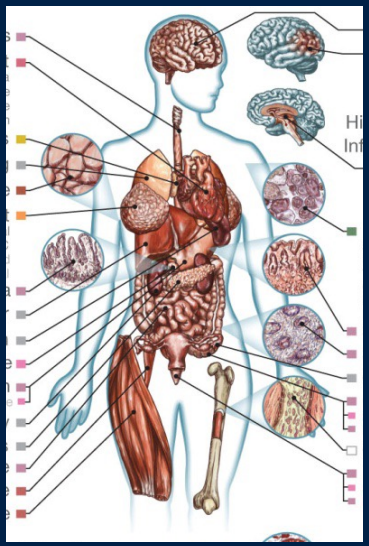
- Fine-mapping – can we identify the actual causal variants underlying these associations, and hence discover specific proteins and disease pathways?
- Pathway analysis – even if we can't fine-map, we can still try to assess whether associations group into particular biological pathways that might shed light on biology
- Pleiotropy analysis – are associations shared between traits, improving our understanding of disease etiology?
- Heritability analysis – how much of the heritability do the signals explain?

Lecture plan

- Recap & fallout from last lecture
- Gaining biological knowledge from GWAS
- Uncovering biology: examples
- Pleiotropy, heritability and prediction

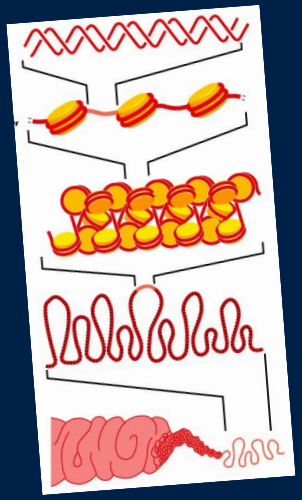
The circle of genetic causation

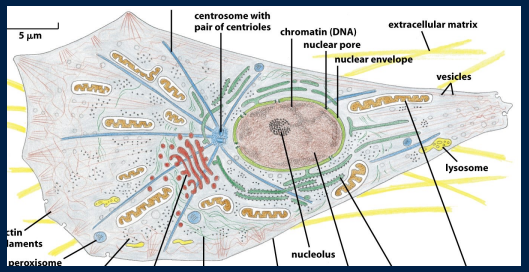




...that combine to make individuals...

Example 1: a pathway analysis





Pathway analysis

Pathway analyses and gene enrichment analysis seek to determine whether there is a statistical tendency for association signals to fall into known groups of related genes. These can be

- Known biological pathways (functional networks of proteins and molecules, performing known specific biological functions) – such as those available from the KEGG and Reactome databases

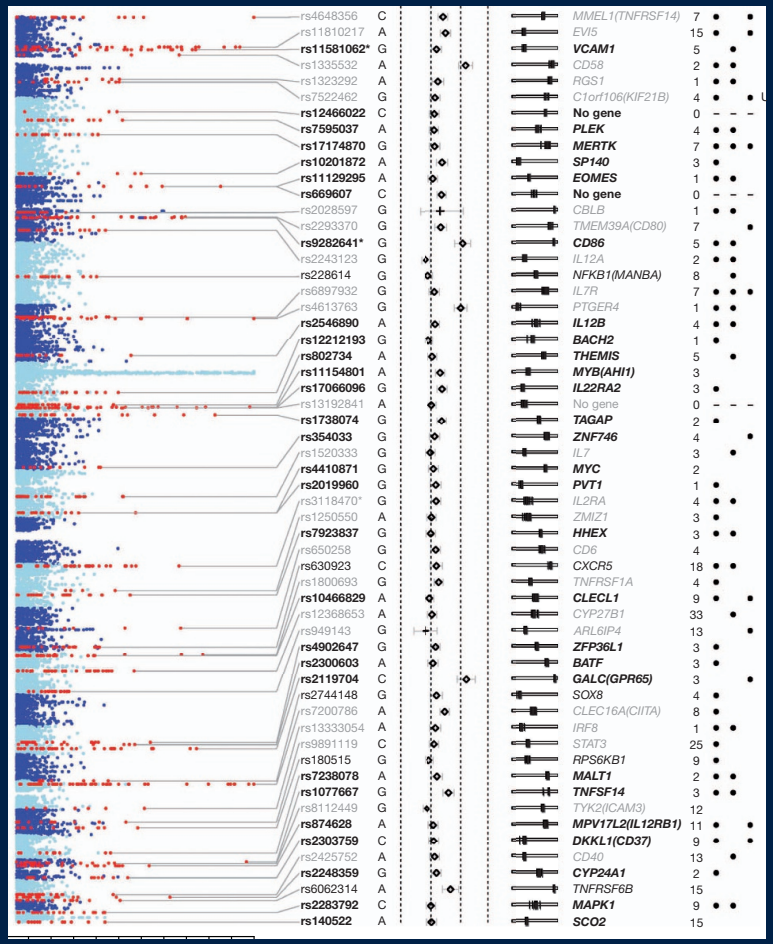
- More general classifications of genes by function, such as those from the Gene Ontology Project

A slightly different direction is to try to group signals by genome function – for example, do they lie in exons? Or gene promoters? Or in regulatory regions active in particular cells?

https://www.genome.jp/kegg/ https://reactome.org http://geneontology.org

Pathway analysis example

The primary cause of MS has typically been thought to be inflammation causing downstream neurodegeneration – with some debate about this. Can the GWAS of MS we discussed shed light on this?



Clinical and Experimental Neuroimmunology **1** (2010) 2–11

REVIEW ARTICLE

What drives disease in multiple sclerosis: Inflammation or neurodegeneration?

Hans Lassmann

Center for Brain Research, Medical University of Vienna, Vienna, Austria

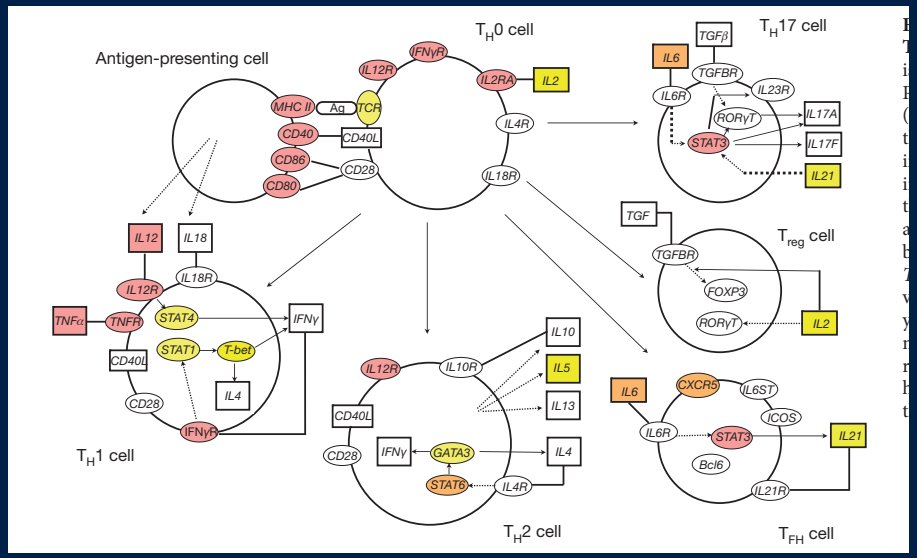
As the main figure shows, many of the association signals looked like they were near immune-system related genes.

www.well.ox.ac.uk/wtccc2/ms/

Pathway analysis example

We:

- Assigned SNPs to their nearest gene using the available annotation
- Used the Gene Ontology Project to classify genes into functionally related groups
- Conducted a statistical test (Fisher's exact test) to identify whether the nearest genes were enriched in each group.



T-helper-cell differentiation pathway
(from Ingenuity Pathway Analysis software)

Particularly strong enrichment was observed for immune system pathways – notably in "T cell activation and proliferation" ($P=1.9\times10^{-9}$)

"Although GO immune system genes only account for 7% of human genes, in 30% of our association regions the nearest gene to the lead SNP is an immune system gene"

Published: 10 August 2011

Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis

The International Multiple Sclerosis Genetics Consortium & The Wellcome Trust Case Control Consortium 2

www.well.ox.ac.uk/wtccc2/ms/

Fine-mapping

"Fine-mapping" is the general term used for attempts to narrow down association signals to the underlying causal variants. A typical process involves:

- Gathering complete information on genetic variation in the region of interest – for example by deep-sequencing a large number of individuals. (Large databases such as gnomAD / TopMed now make this easier.)

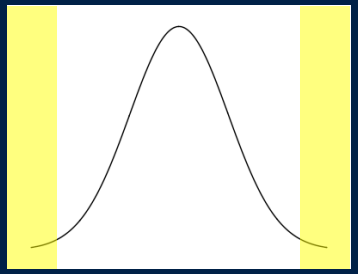
- Gathering information on genome function – including gene structure and regulatory regions.

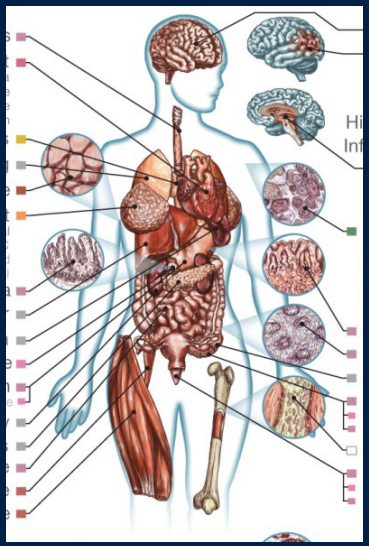
- Potentially leveraging data from different ancestral backgrounds, hoping that differences in LD patterns will help narrow down signals.

- Fitting models that attempt to parse apart multiple associations in the same region

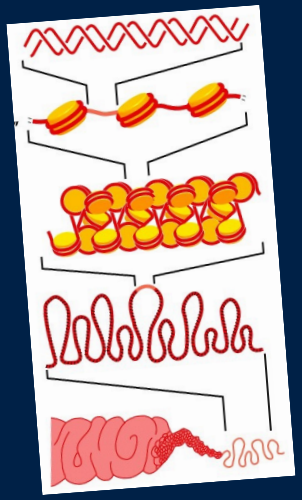
Possible underlying mechanisms are pretty diverse and a healthy dose of genomic detective work is often needed.

The circle of genetic causation

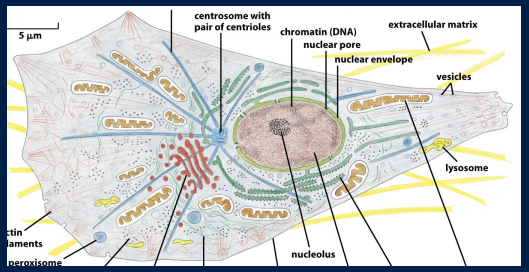


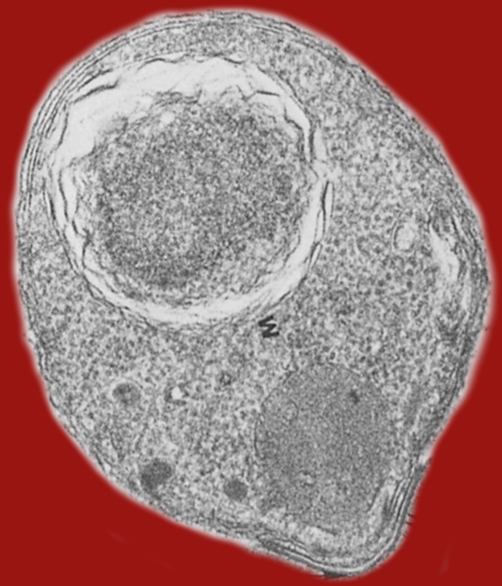


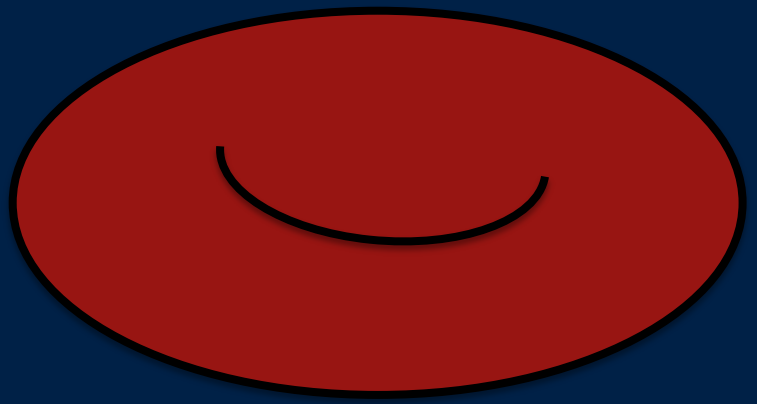
...that combine to make individuals...



Example 2: fine-mapping



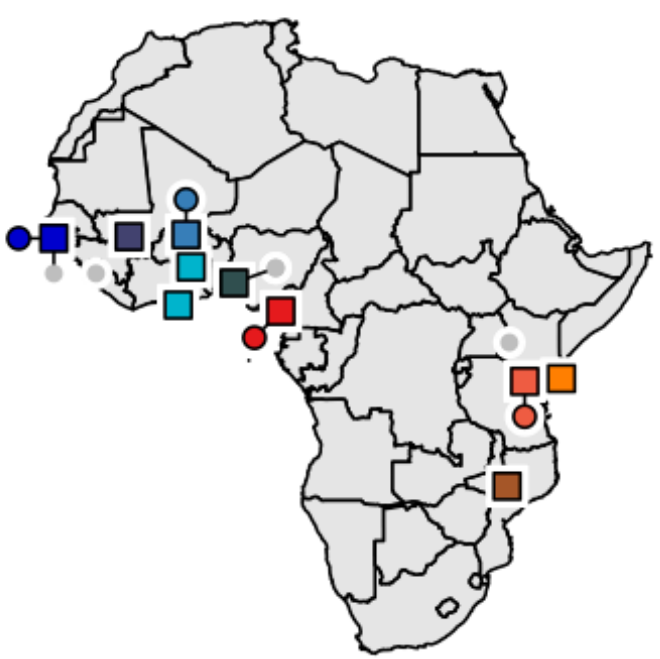




Plasmodium falciparum **VS** humans

Nature Communications: doi.org/10.1038/s41467-019-13480-z
or on bioArxiv: doi.org/10.1101/535898

GWAS of susceptibility to severe malaria



Study samples

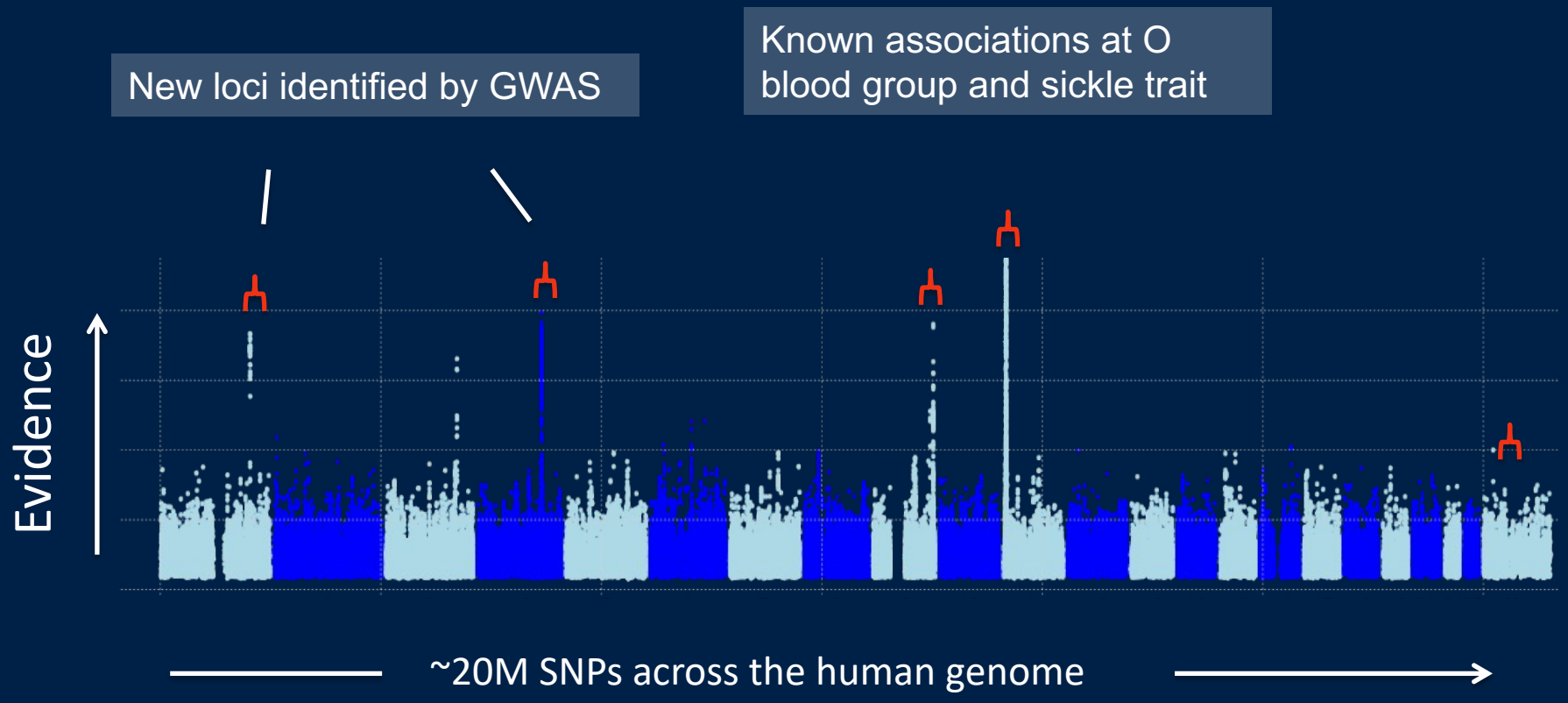
group	cases	controls	TOTAL
Africa			
Gambia	2567	2605	**5172**
Mali	274	183	**457**
Burkina Faso	733	596	**1329**
Ghana	399	320	**719**
Nigeria	113	22	**135**
Cameroon	592	685	**1277**
Malawi	1182	1317	**2499**
Tanzania	416	403	**819**
Kenya	1681	1615	**3296**
Asia			
Vietnam	718	546	**1264**
Oceania			
PNG	402	374	**776**

~17,000 clinical samples from West and East Africa, Oceania and South East Asia.
Genotyped on the Illumina Omni 2.5M array

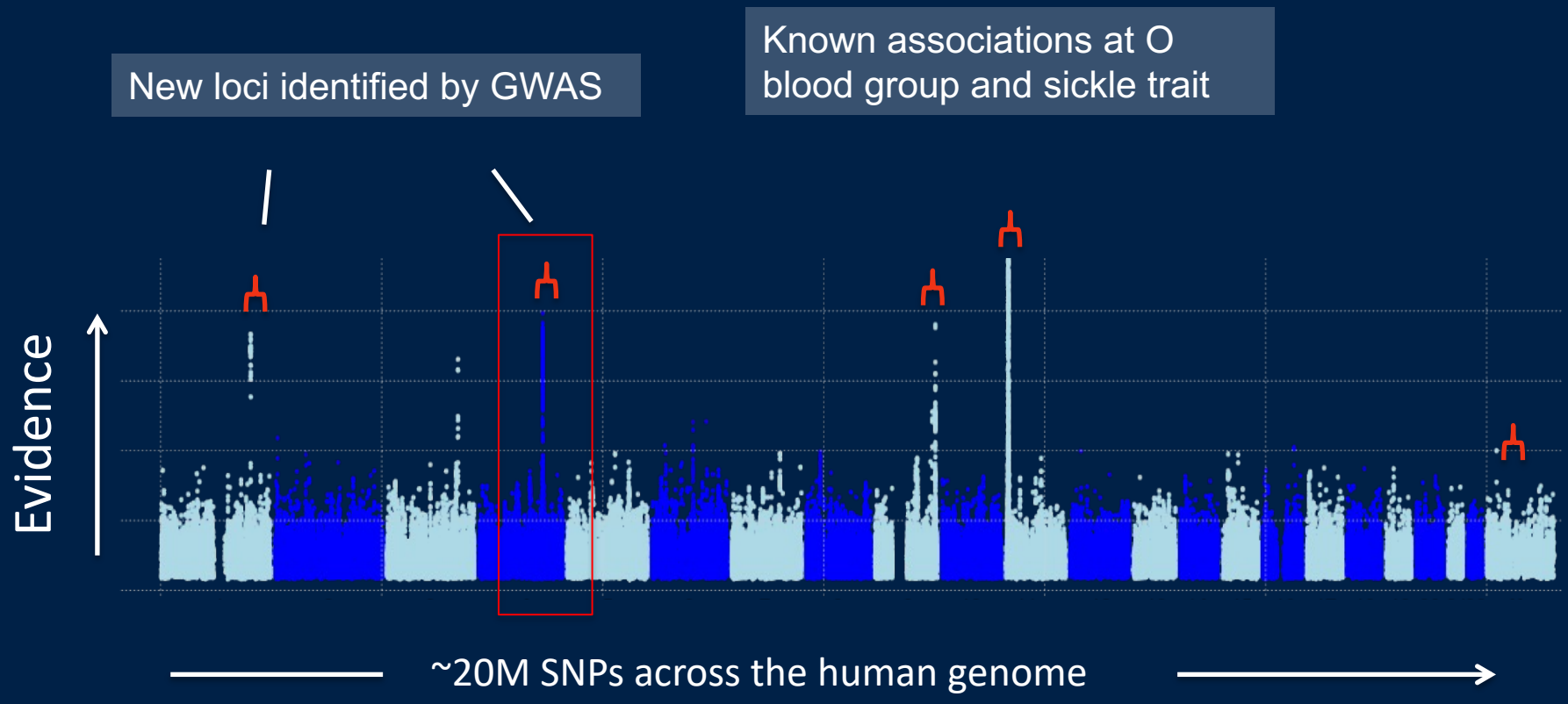
MalariaGEN
GENOMIC EPIDEMIOLOGY NETWORK

www.malariagen.net

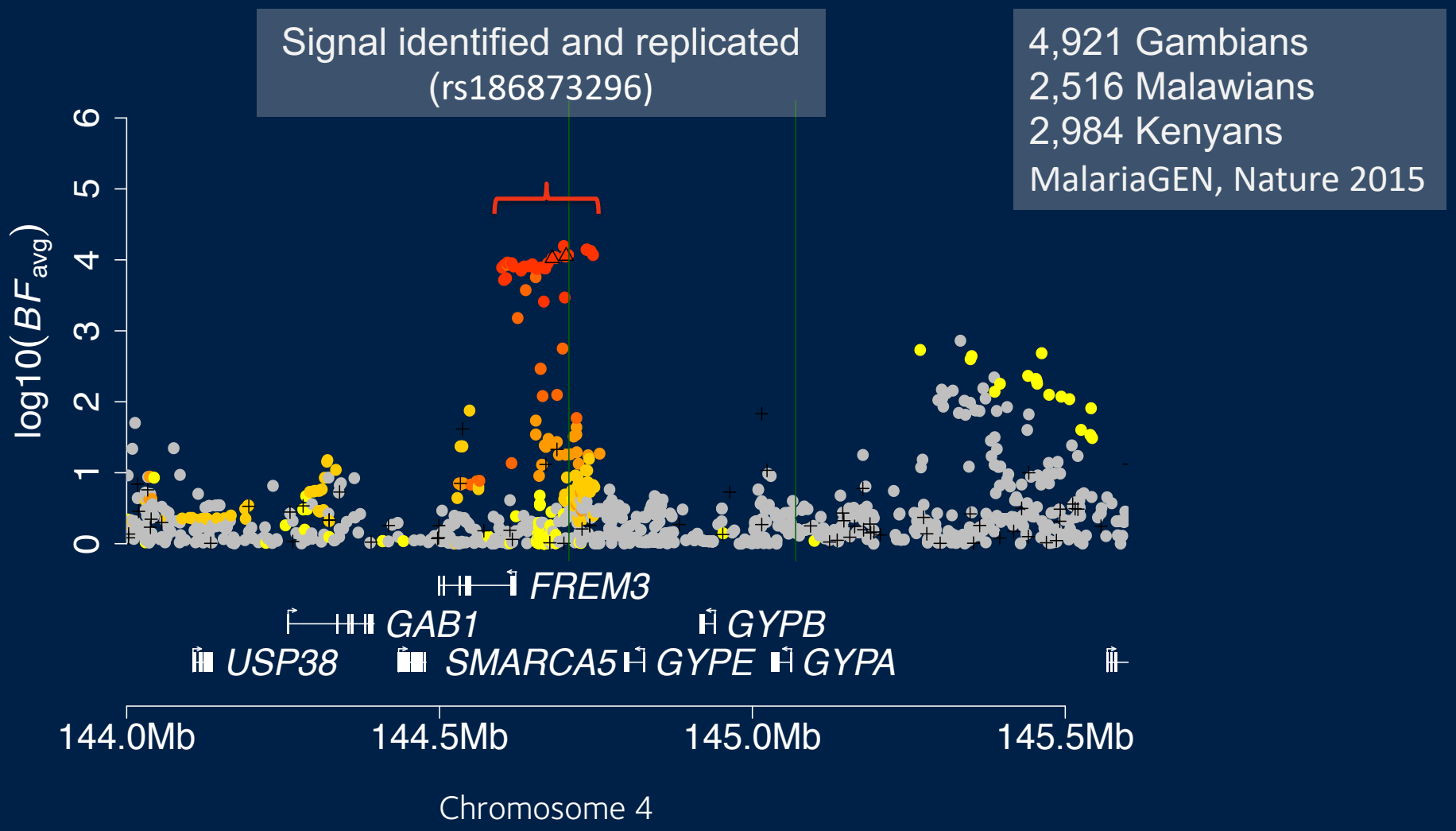
Natural resistance is driven by red blood cell variation



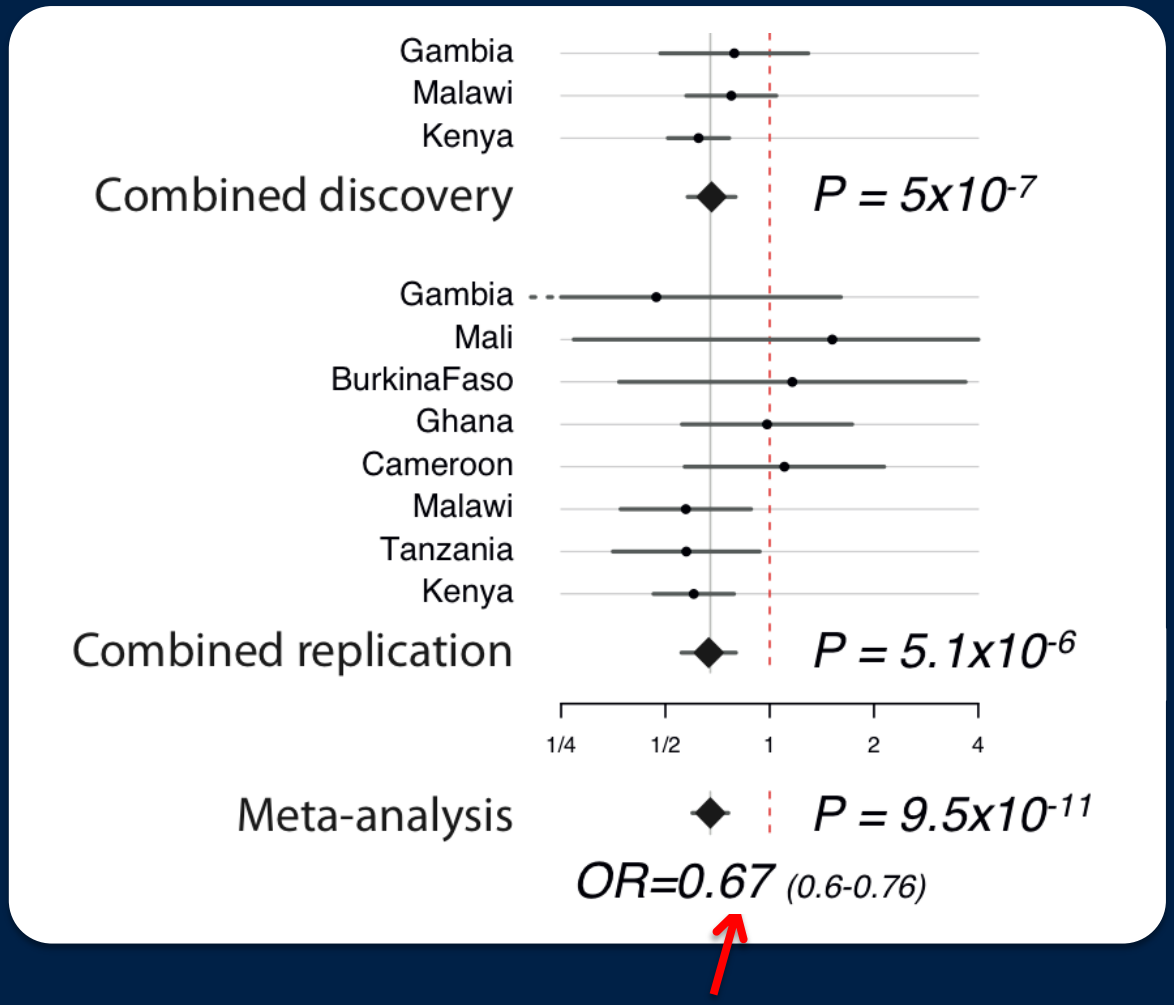
Natural resistance is driven by red blood cell variation



SNPs on chromosome 4 are associated with proection against severe malaria



The association has quite large effect



> 30% protective effect per copy of the derived allele

$$\text{Standard error}(\log OR) \approx \frac{1}{\sqrt{N \times f(1-f) \times \phi(1-\phi)}}$$

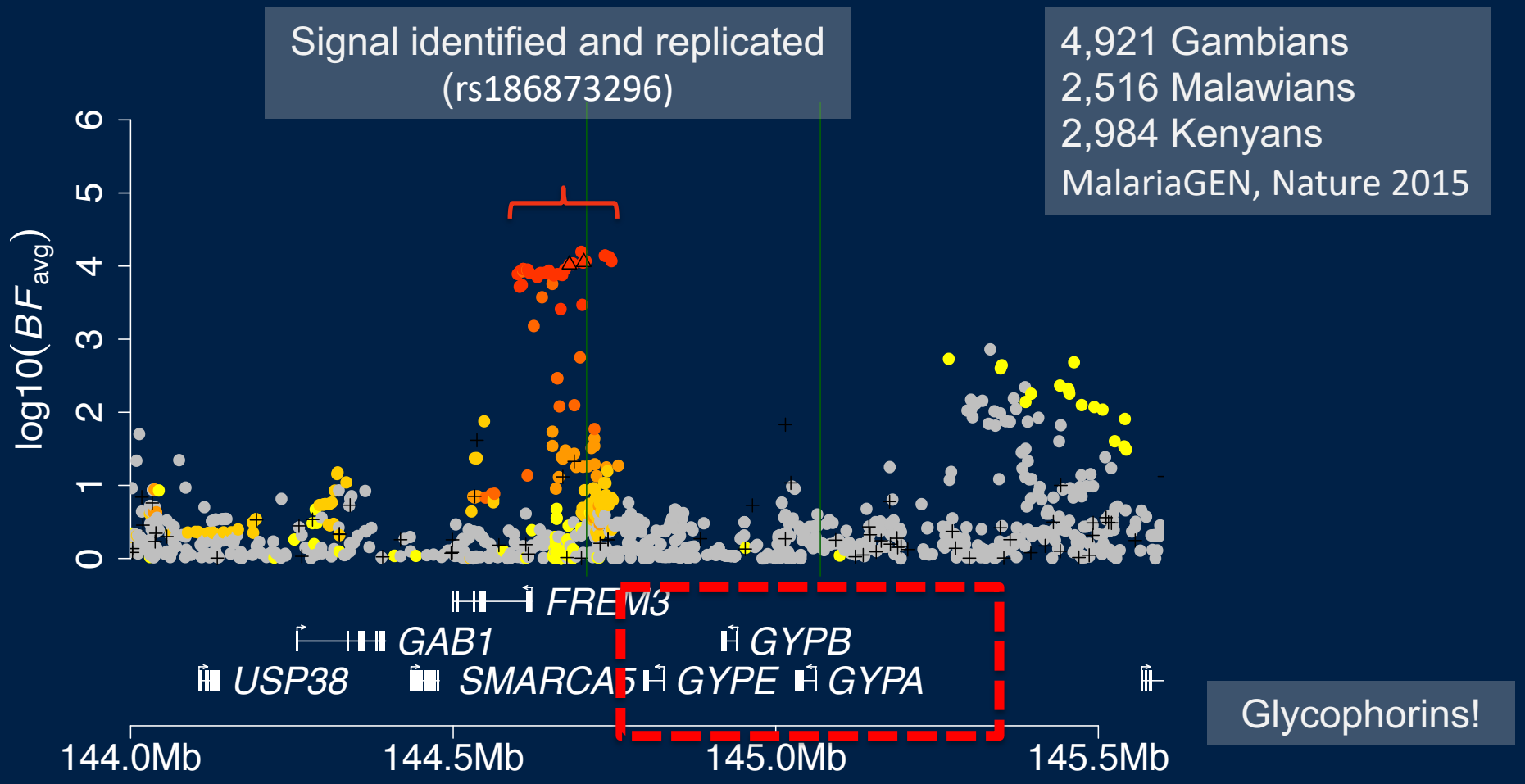
Can we finemap?

We had an exciting association. But fine-mapping has proven to be difficult for many GWAS loci.

To hope for success we might need:

- Good candidates for the functional gene?
- Good candidates for the causal mutation(s)?

SNPs on chromosome 4 are associated with proection against severe malaria



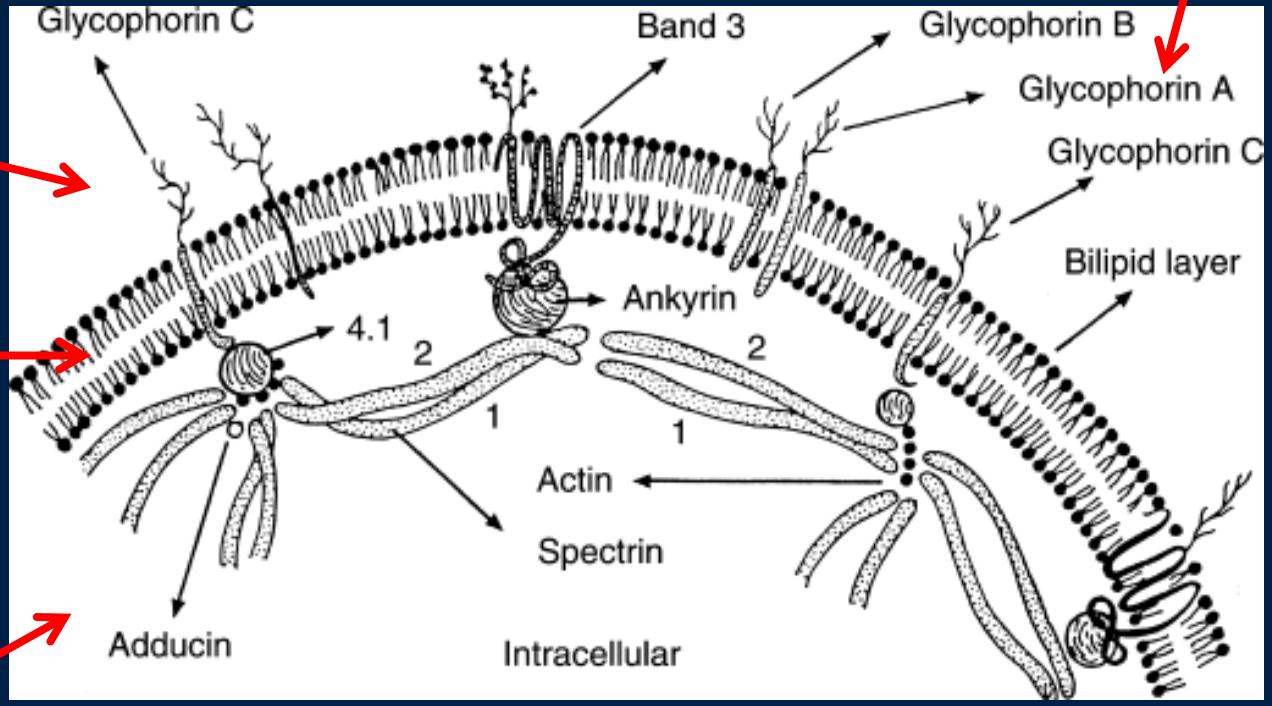
Glycophorins encode the ‘MNS’ blood group
(antigenic molecules on RBC surface)

Glycophorins

Outside red blood cell

Red blood cell membrane

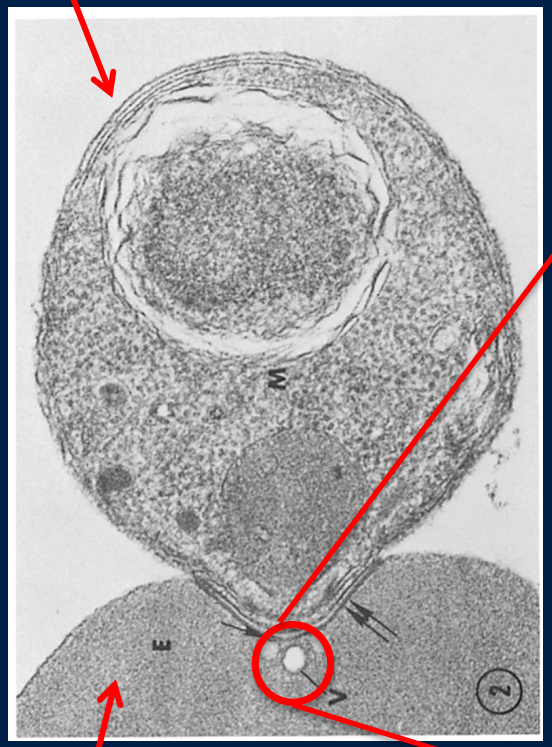
Inside red blood cell



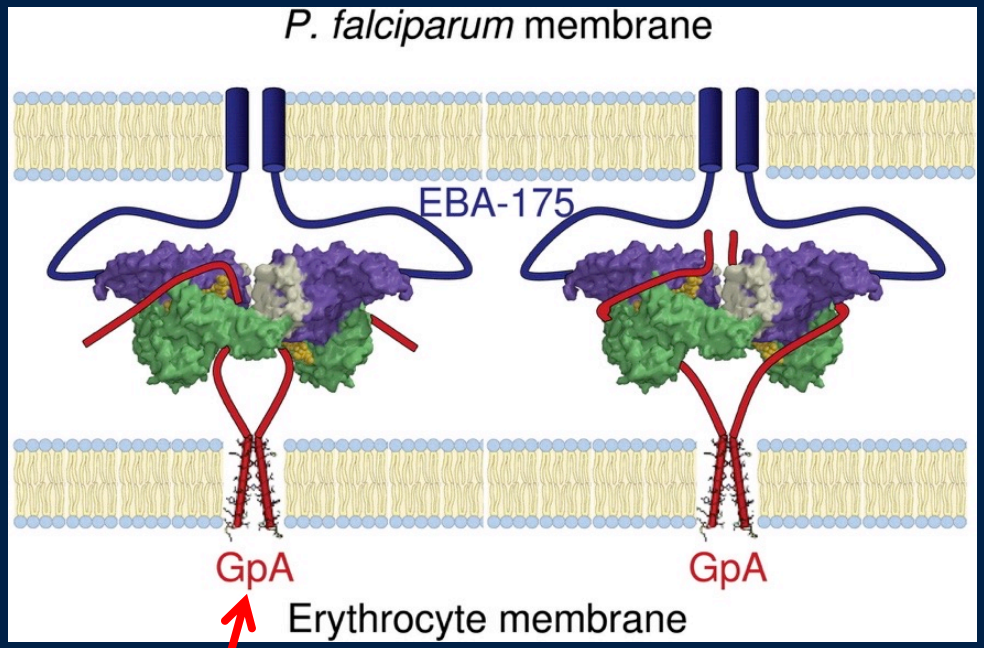
Grimes and Slater, The Inherited Metabolic Diseases, 1994

Glycophorins are receptors for *P.falciparum* during red blood cell invasion

P. Falciparum parasite



red blood cell



Glycophorin A

Miller et al, J. Exp. Med 1979

Tolia et al, Cell 2005

Can we finemap?

We had an exciting association. But fine-mapping has proven to be difficult for many GWAS loci.

To hope for success we might need:

✅ - Good candidates for the functional gene?

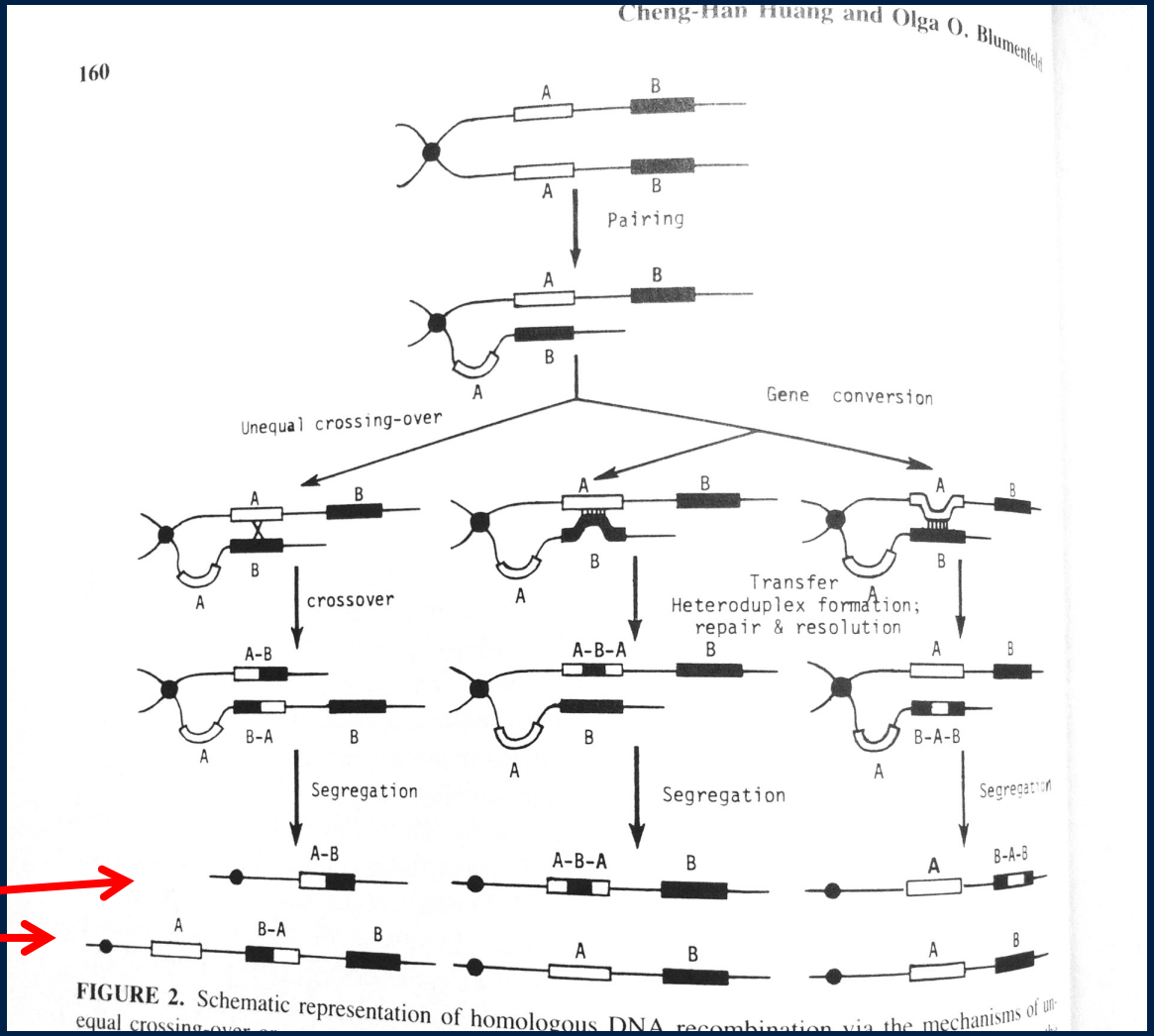
- Good candidates for the causal mutation(s)?

Structural variants create deletions, duplications, and hybrid genes

The MNS blood group is highly diverse, with over 45 known antigens.

Encoded by single nucleotide polymorphisms and structural variants

Deleted / duplicated / hybrid genes



Can we finemap?

We had an exciting association. But fine-mapping has proven to be difficult for many GWAS loci.

To hope for success we might need:

✅ - Good candidates for the functional gene?
✅ - Good candidates for the causal mutation(s)?

Steps to fine-map

Step 1: type or sequence as much of the genetic variation in the region as possible – hope to catch the causal mutation.

Step 2: re-analyse the association.

Step 3: look for functional mutations

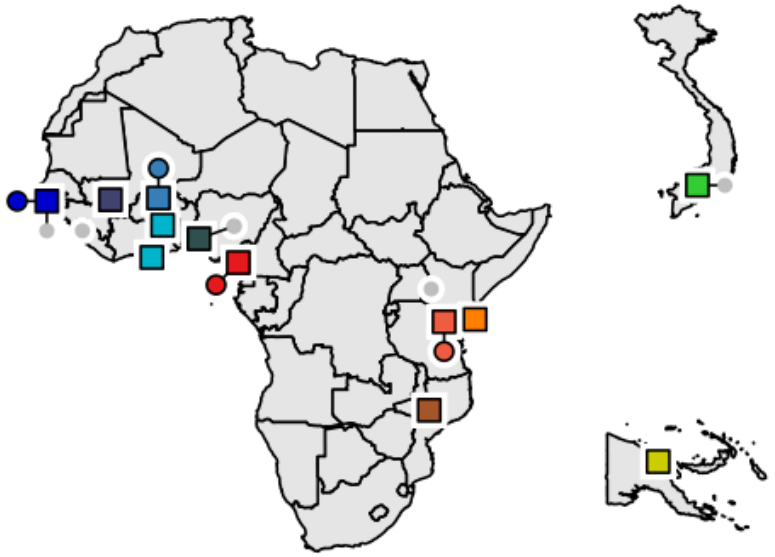
A regional reference panel capturing structural variation

We used the 1000 Genomes Project Phase III reference panel, plus:

a

Whole–genome sequences

group	trios	duos	other	TOTAL
Gambia				
FULA	31	1	5	**100**
JOLA	32	1	2	**100**
MANDINKA	33	0	1	**100**
WOLLOF	32	1	3	**98**
Burkina Faso				
MOSSI	0	0	57	**57**
Cameroon				
BANTU	5	3	11	**31**
SEMIBANTU	8	0	7	**32**
Tanzania				
CHAGGA	21	2	13	**80**
PARE	22	2	7	**77**
WASAAMBA	23	6	9	**90**

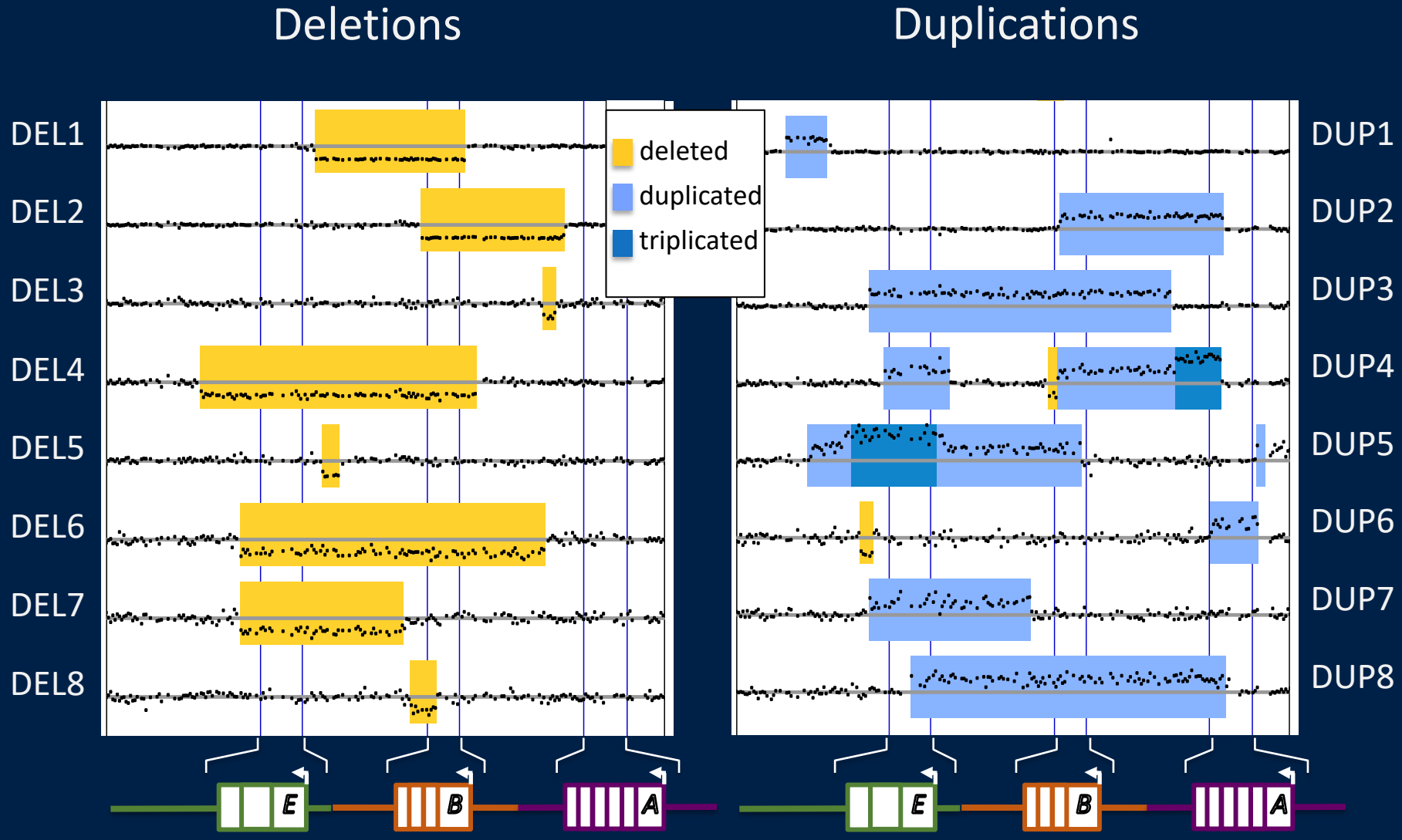


Use whole-genome sequencing from over 3,600 individuals worldwide.
Discover genetic variation (including structural variants).

MalariaGEN
GENOMIC EPIDEMIOLOGY NETWORK

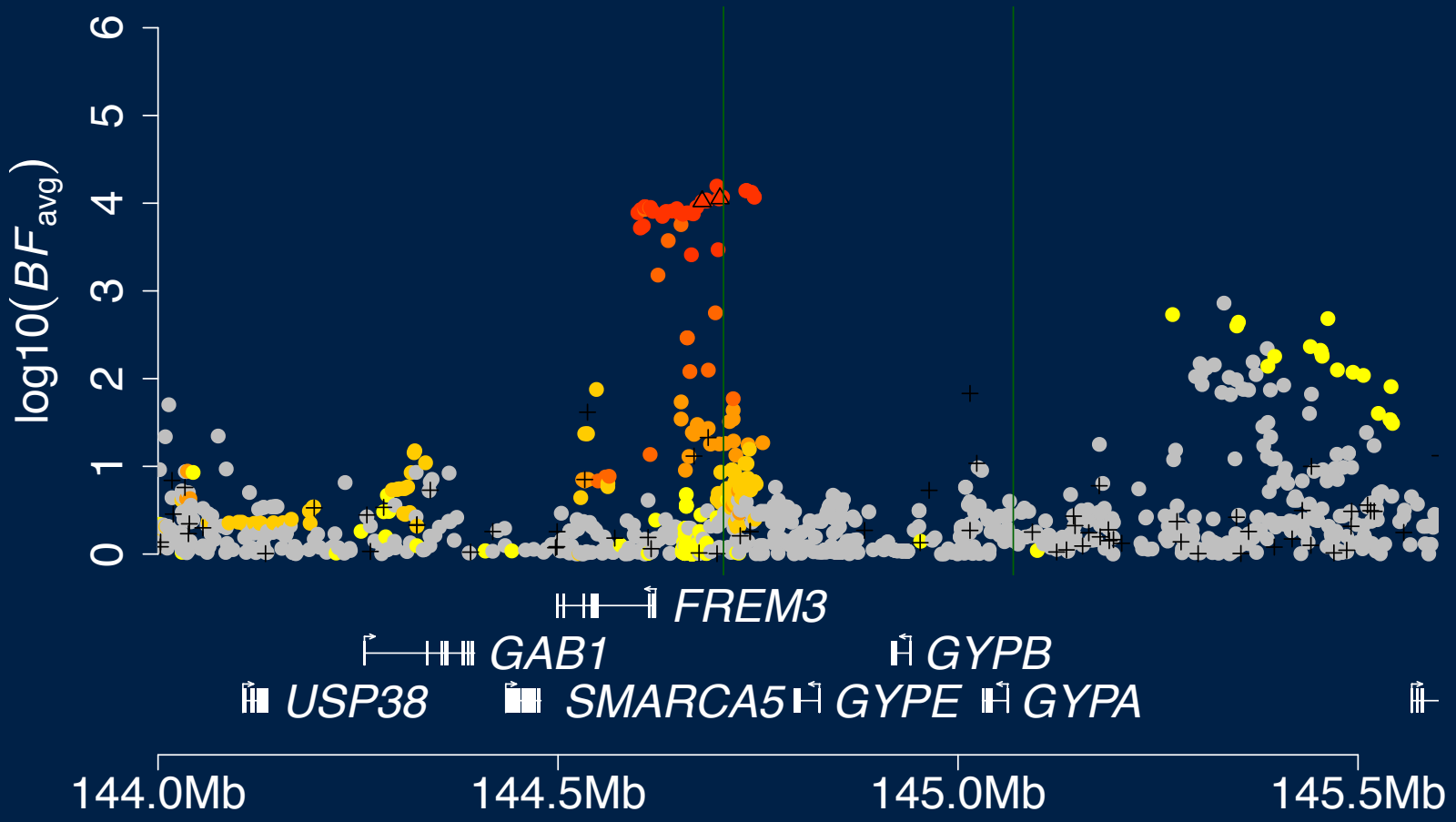
www.malariagen.net

Structural variants from sequencing data



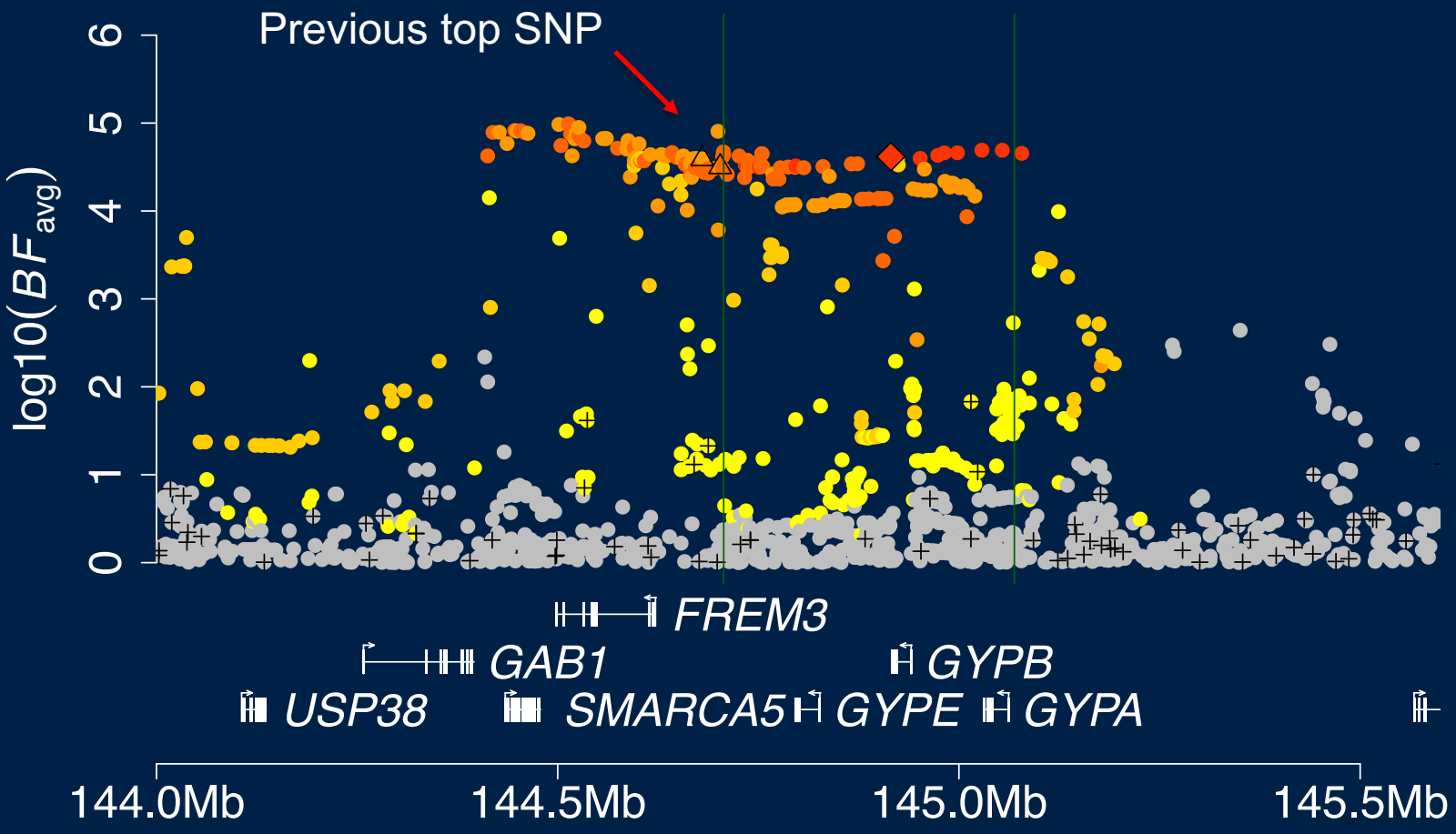
14% of Africans carry a CNV affecting these genes

Before fine-mapping



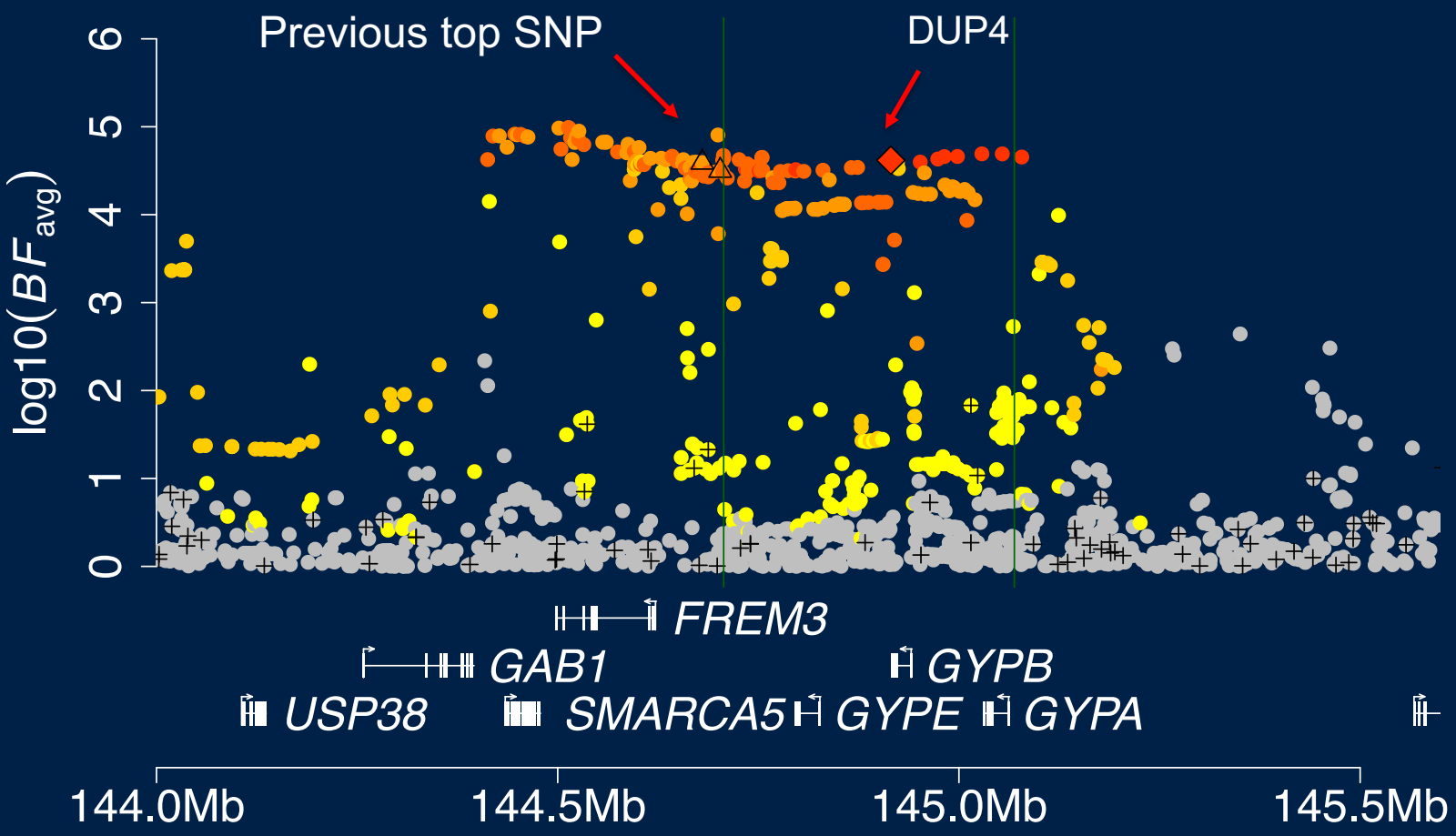
Original result before adding information from new African sequenced genomes

After fine-mapping

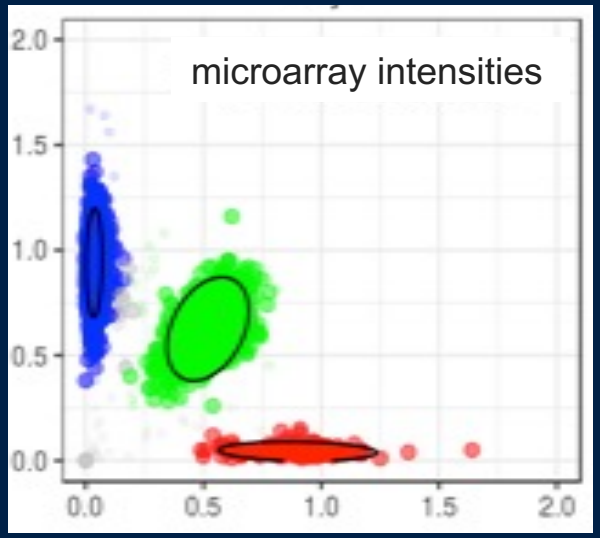


Result after incorporating genetic variation discovered in sequenced samples

After fine-mapping



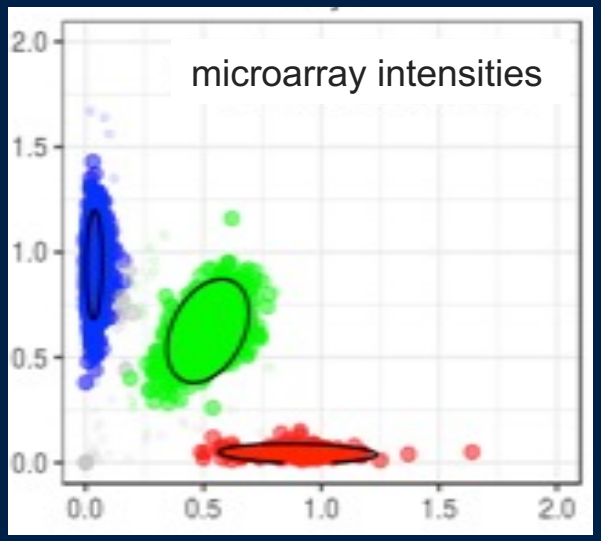
Confirming structural variants using cluster plots



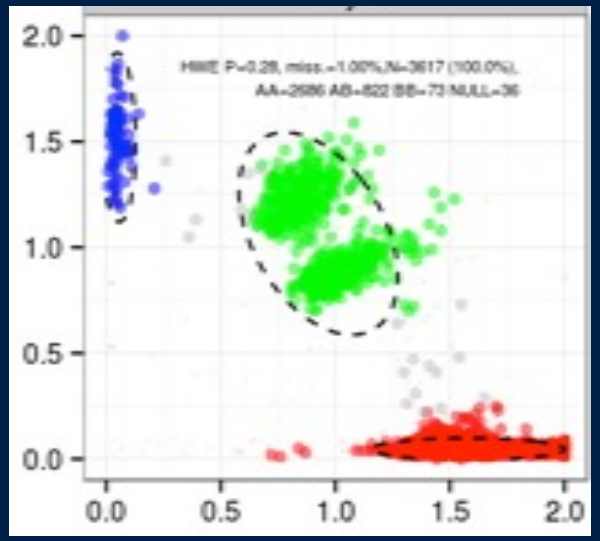
This is how a microarray cluster plot should look: 3 clusters for AA / AB / BB genotypes

Confirming structural variants using cluster plots

Actually this signal was evident in our cluster plots



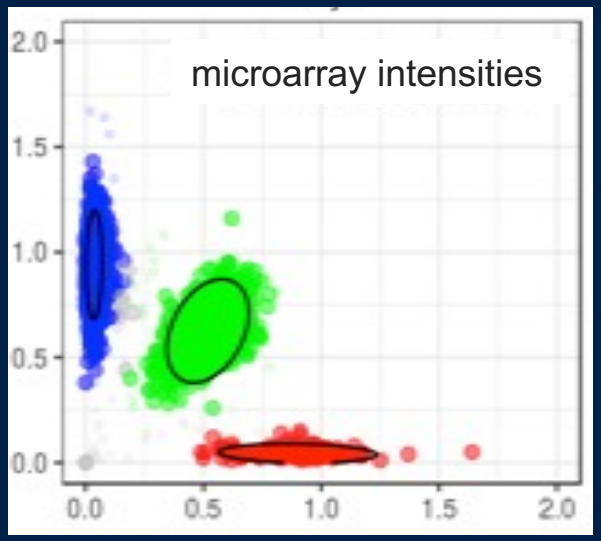
This is how a microarray cluster plot should look: 3 clusters for AA / AB / BB genotypes



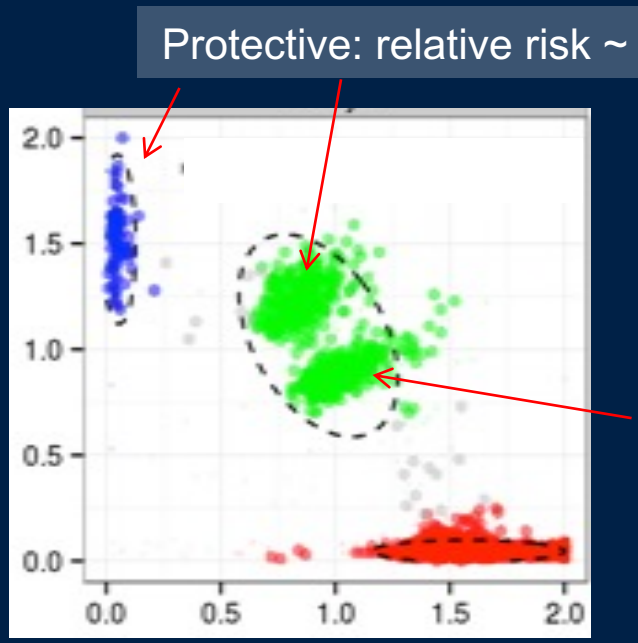
What we saw in this region

Confirming structural variants using cluster plots

Still true that nothing seemed to be functional. What next?

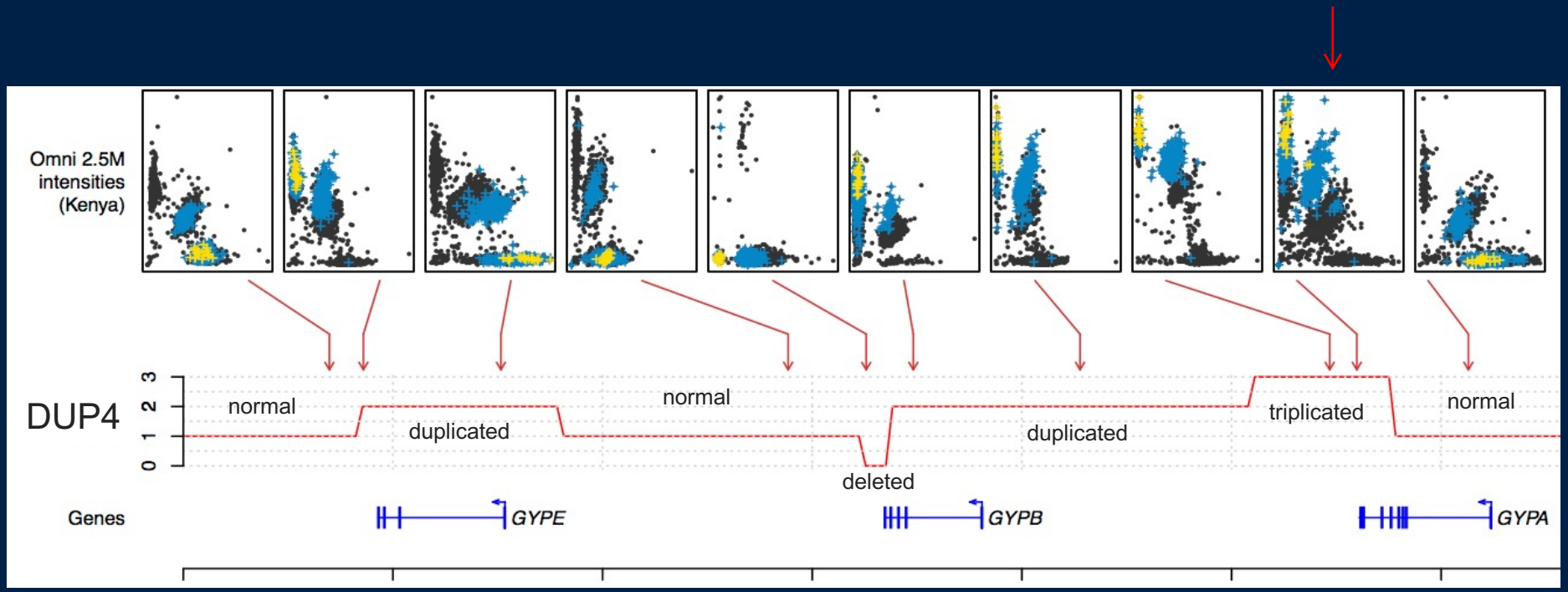


This is how a microarray cluster plot should look: 3 clusters for AA / AB / BB genotypes



What we saw in this region

Confirming structural variants using cluster plots

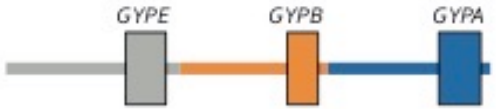


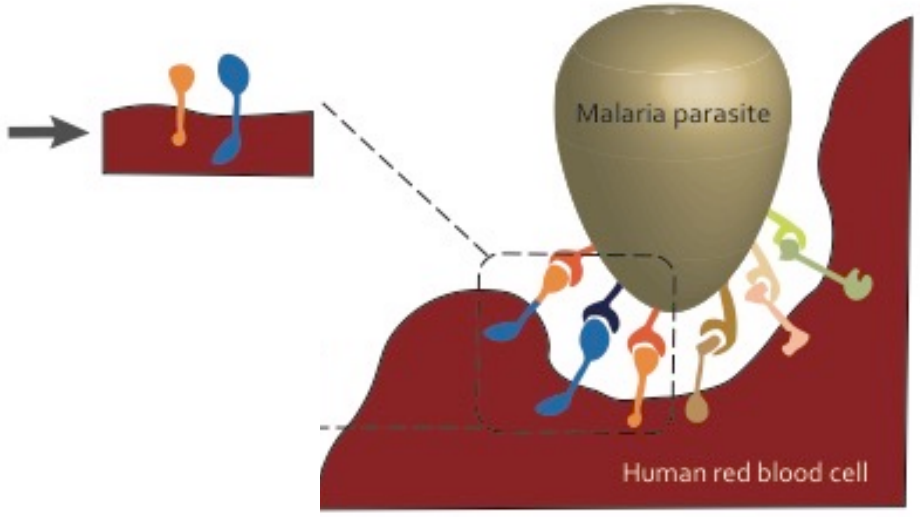
We were able to use cluster plots to confirm individuals in our GWAS really do carry the complicated structural variant “DUP4”.

DUP4 is pretty complicated – what could it be?

What is DUP4?

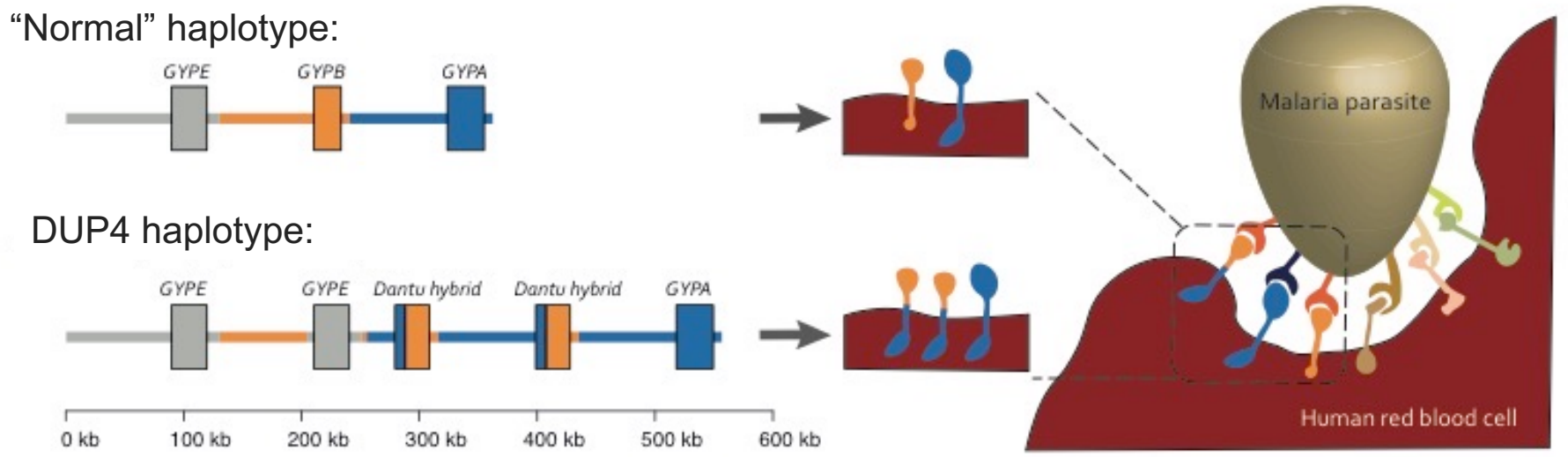
“Normal” haplotype:





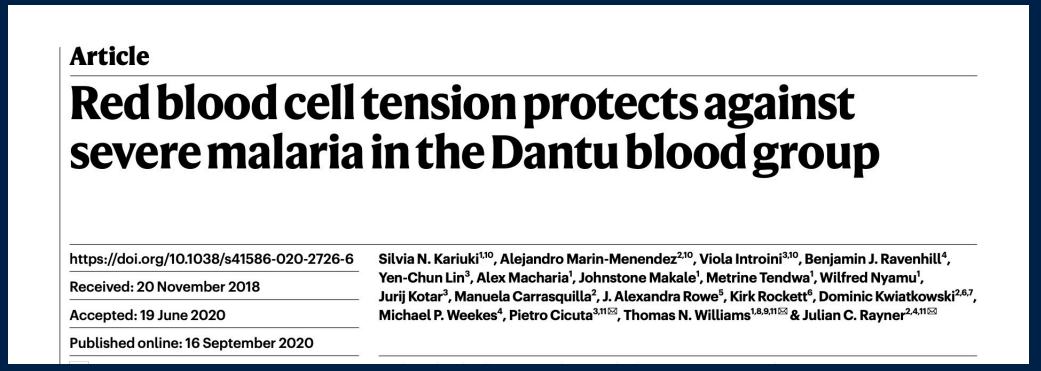
https://doi.org/10.1126/science.aam6393

What is DUP4?



https://doi.org/10.1126/science.aam6393

Functional followup study →



Article

Red blood cell tension protects against severe malaria in the Dantu blood group

https://doi.org/10.1038/s41586-020-2726-6
Received: 20 November 2018
Accepted: 19 June 2020
Published online: 16 September 2020

Silvia N. Kariuki[1,10], Alejandro Marin-Menendez[2,10], Viola Introini[3,10], Benjamin J. Ravenhill[4], Yen-Chun Lin[3], Alex Macharia[1], Johnstone Makale[1], Metrine Tendwa[1], Wilfred Nyamu[1], Jurij Kotar[3], Manuela Carrasquilla[2], J. Alexandra Rowe[5], Kirk Rockett[6], Dominic Kwiatkowski[2,6,7], Michael P. Weekes[4], Pietro Cicuta[3,11✉], Thomas N. Williams[1,8,9,11✉] & Julian C. Rayner[2,4,11✉]

https://doi.org/10.1038/s41586-020-2726-6

Dantu is globally rare…

The Dantu blood group has been found in:

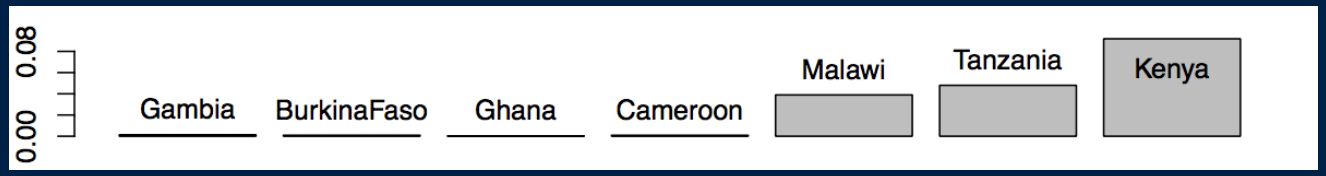
1 in 44,112	Londoners*
0 in 1,000	Germans†
1 in 320	African Americans†
0 in 2870	Gambians‡

...but found at high frequency in east Africa

The Dantu blood group has been found in:

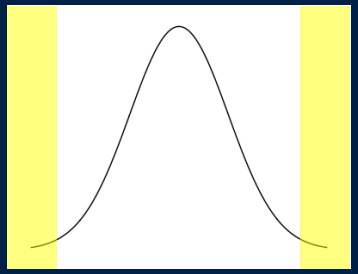
1 in 44,112	Londoners*
0 in 1,000	Germans†
1 in 320	African Americans†
0 in 2870	Gambians‡
1 in 12	Malawians‡
1 in 6	Kenyans (from the Kilifi region)‡

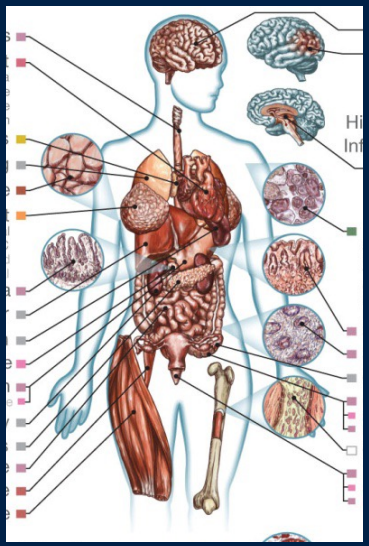
Allele frequency:



West Africa ⟷ East Africa

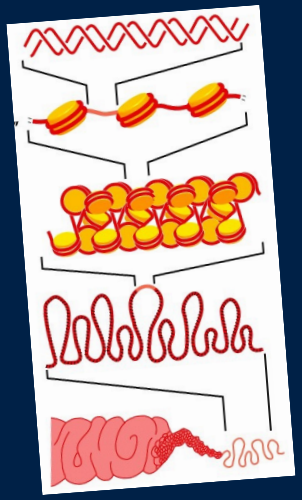
The circle of genetic causation

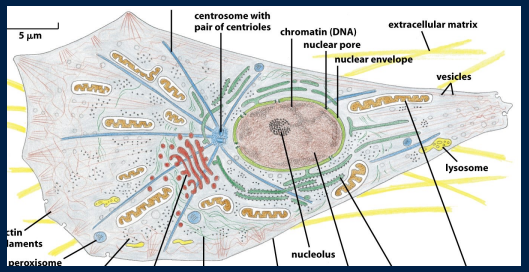




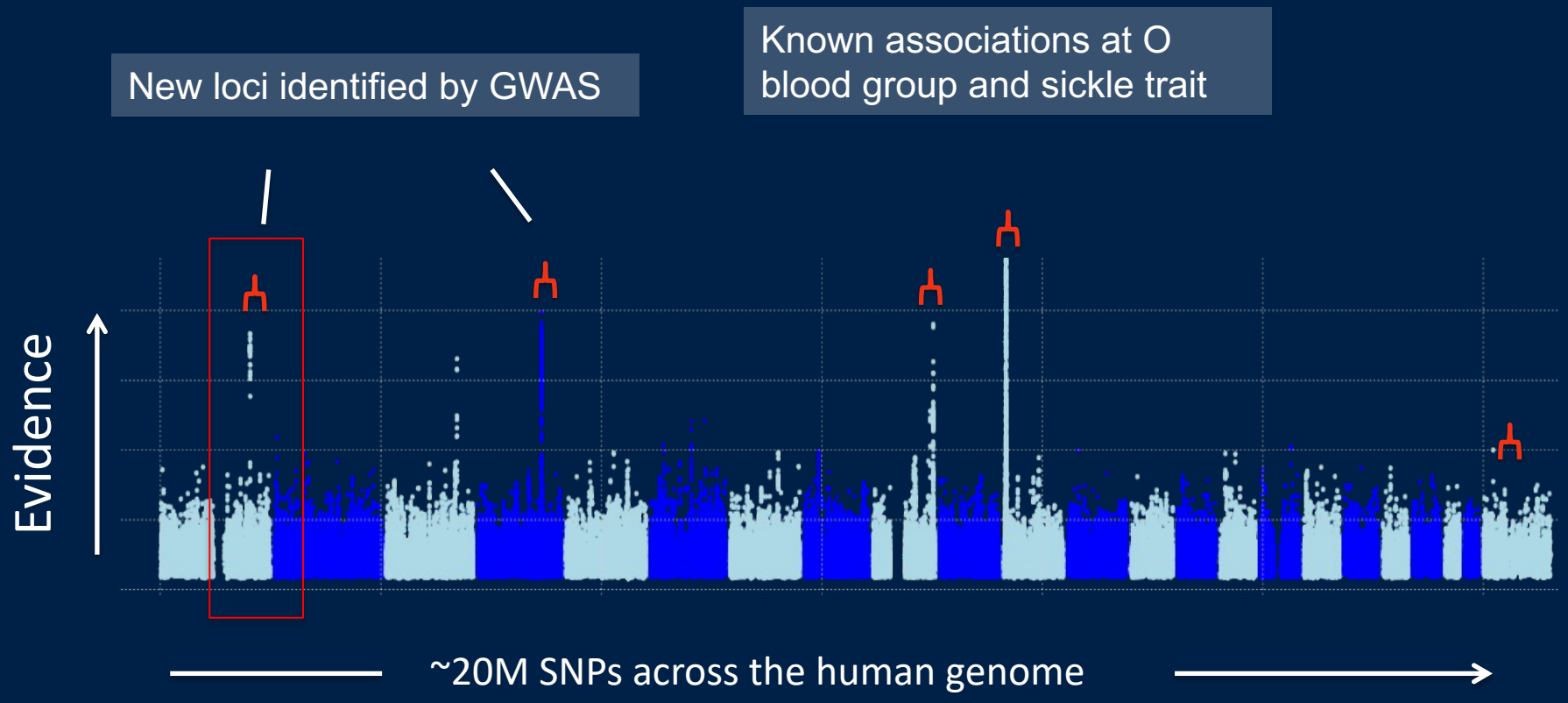
...that combine to make individuals...

Example 3: more fine-mapping

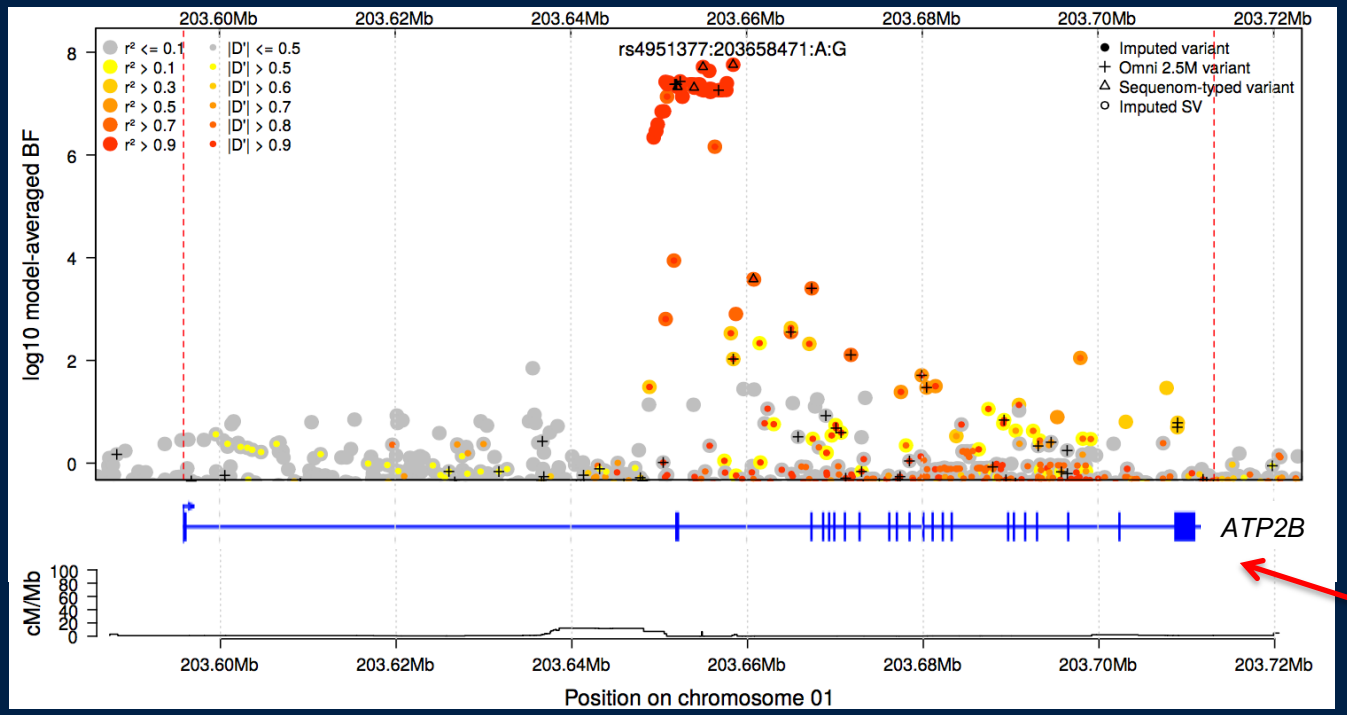




Natural resistance is driven by red blood cell variation



Association near 2nd exon of *ATP2B4*

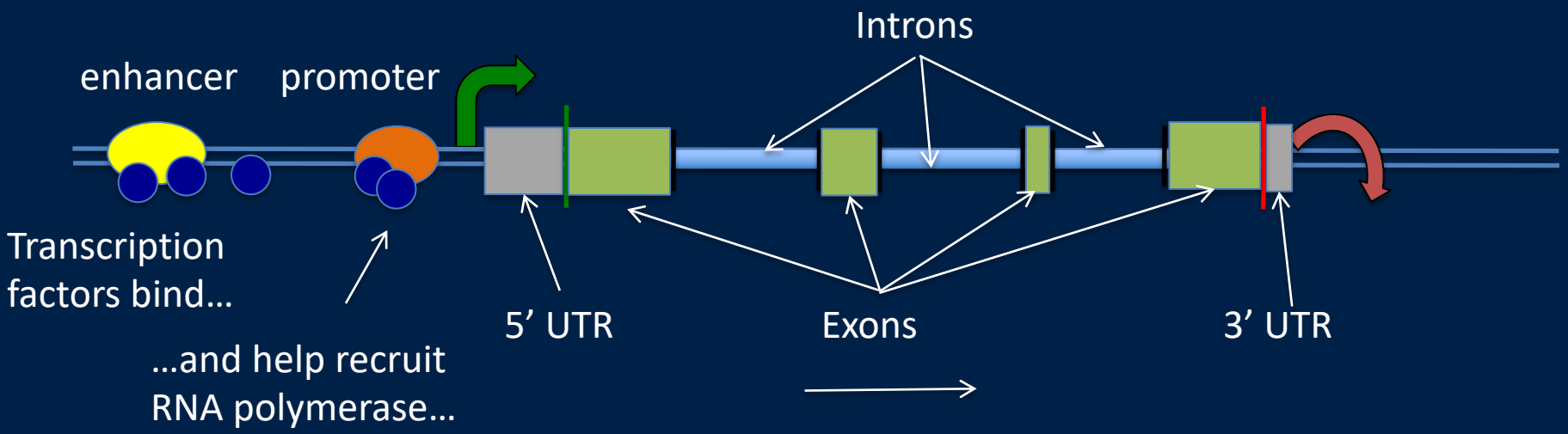


"Canonical" gene model for ATP2B4

ATP2B4 = a red cell "calcium pump"

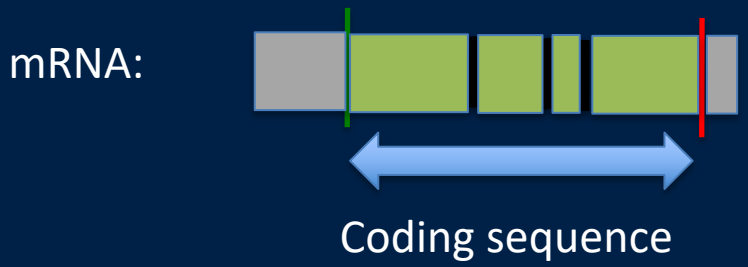
The associated SNPs cover a region around the second exon.
None of these SNPs make changes to the protein.
What could be going on?

Cartoon of a gene

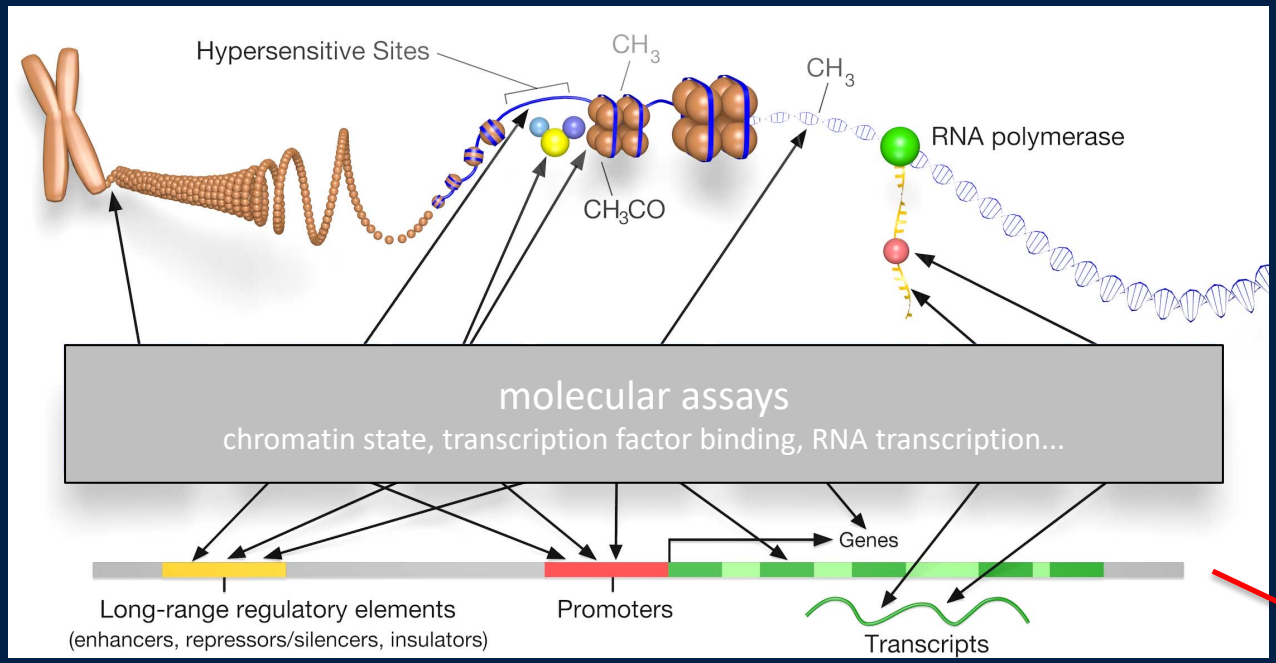


…which transcribes the gene into “pre-mRNA”.

The pre-mRNA is then typically further postranscriptionally modified to remove introns.



Two ways to look at transcription



Can look at chromatin state

RNA expression

ENCODE

ROADMAP epigenomics PROJECT

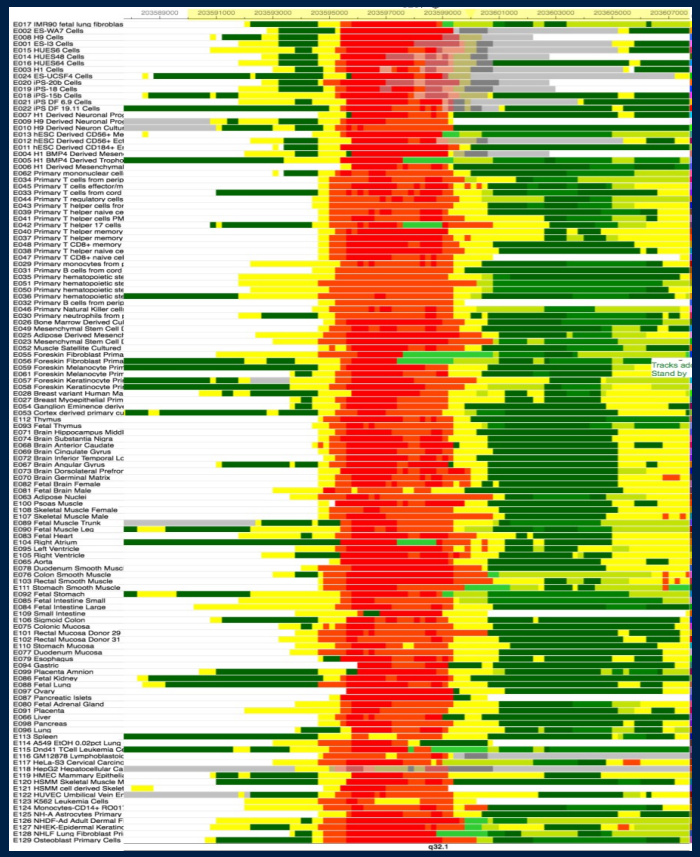
GTEx Portal

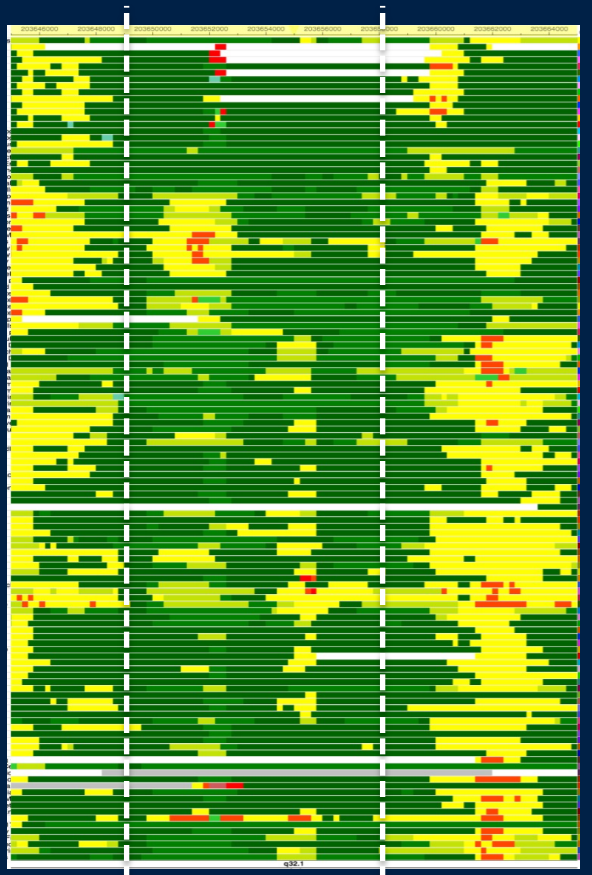
ATP2B4 is widely expressed...

1st exon

2nd exon

Chromatin states in 130 cell types

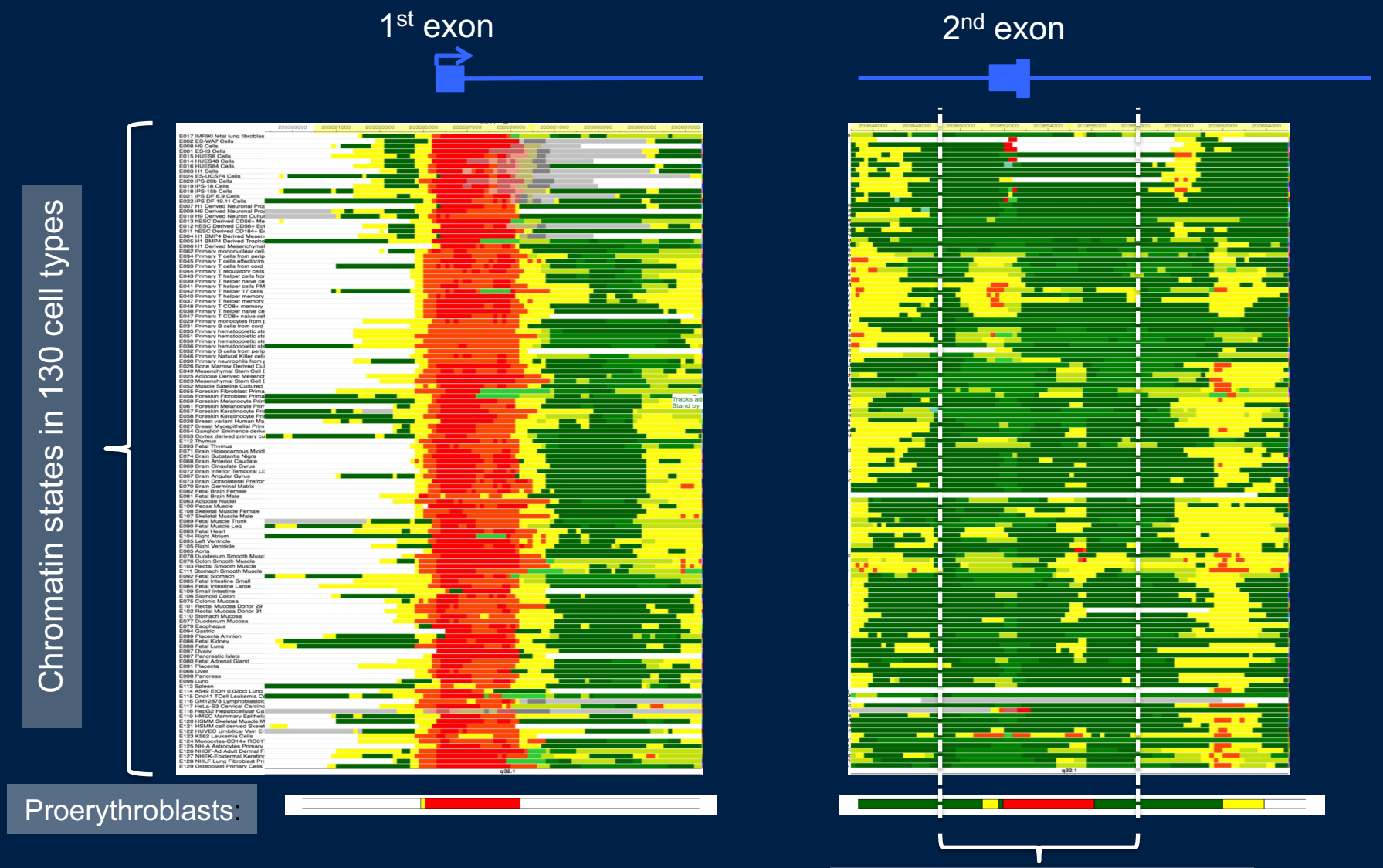




Data from ENCODE / Roadmap

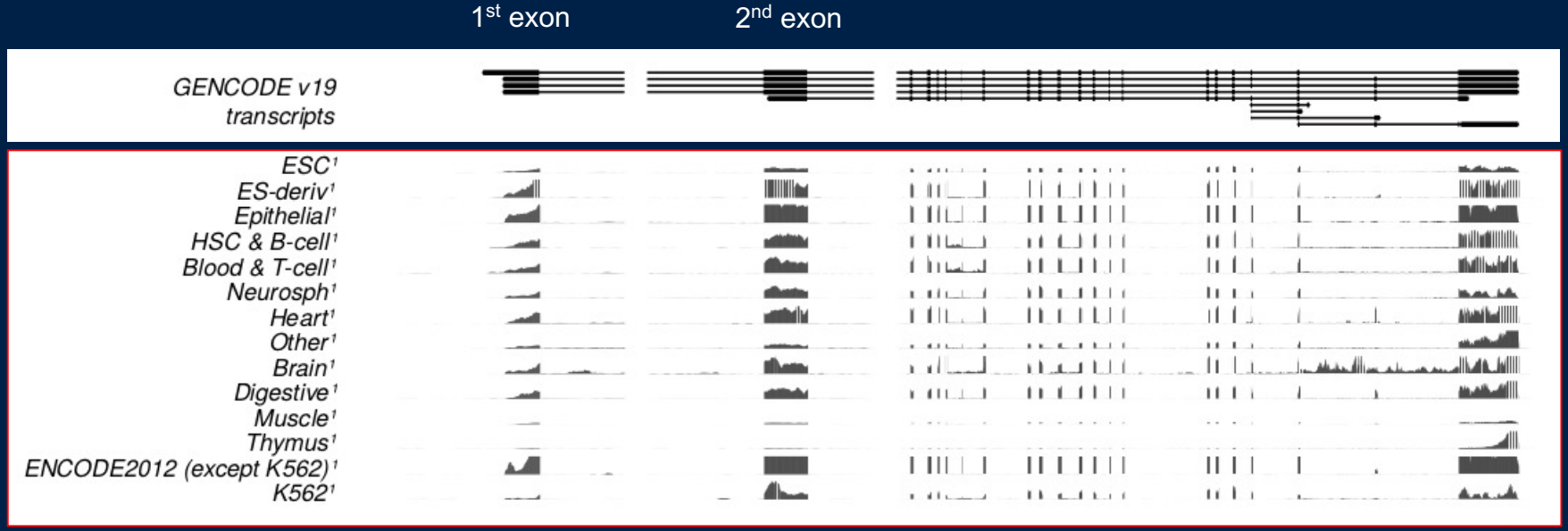
Malaria-associated region

...but shows chromatin differences in RBCs



ATP2B4 is widely expressed…

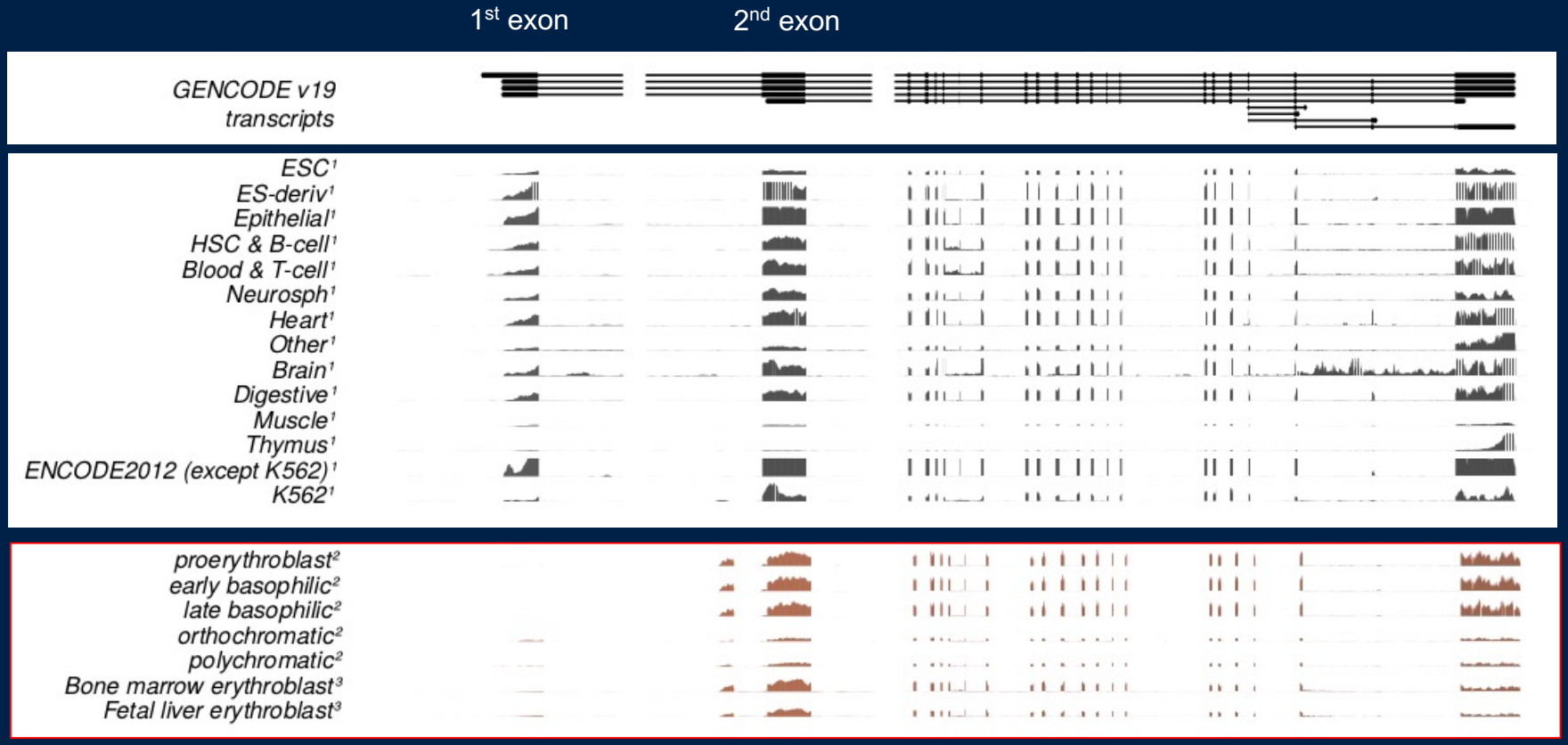
Measured RNA transcription (RNA-seq)



Non-erythroid cells (i.e. no red blood cells)

ATP2B4 has an erythroid-specific transcript

Measured RNA transcription (RNA-seq)

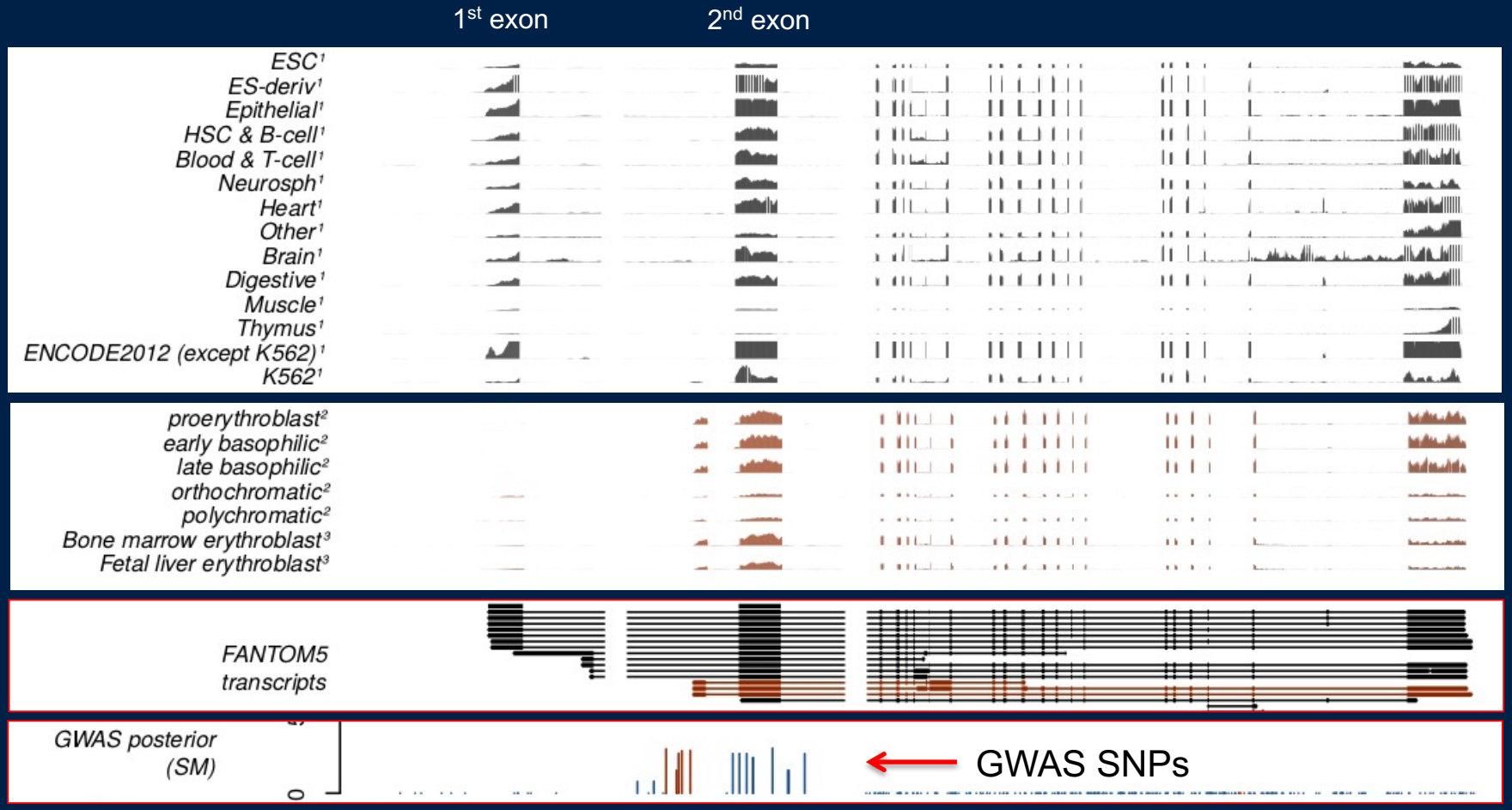


Erythroid cells show a different expression pattern.

Red cells do not have nuclei, so to capture mRNA expression in red cells, these studies experimentally differentiated stem cells into the erythroid lineage, and measured transcription before enucleation.

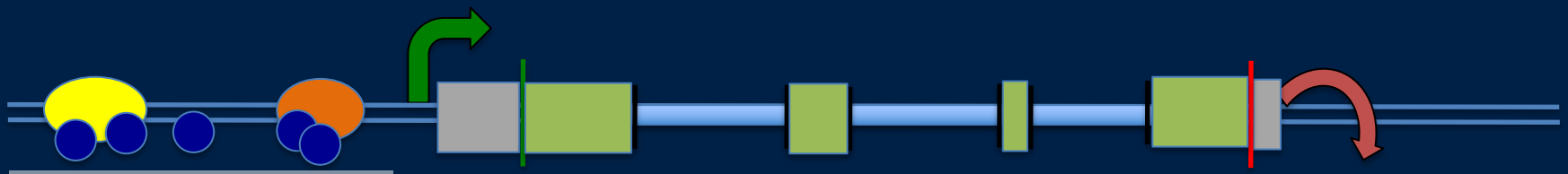
ATP2B4 has an erythroid-specific transcript

Measured RNA transcription (RNA-seq)



Putting together data from a variety of sources suggests the existence of an *alternative transcription start site* near the GWAS signal, but only active in erythrocytes. How can this be?

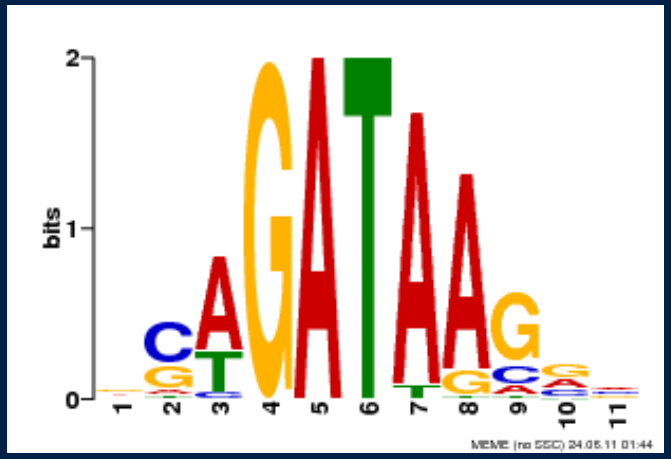
What is different about RBCs?



Transcription factor binding

The transcription of genes in red blood cells is controlled by a particular set of transcription factors – a key one is GATA1.

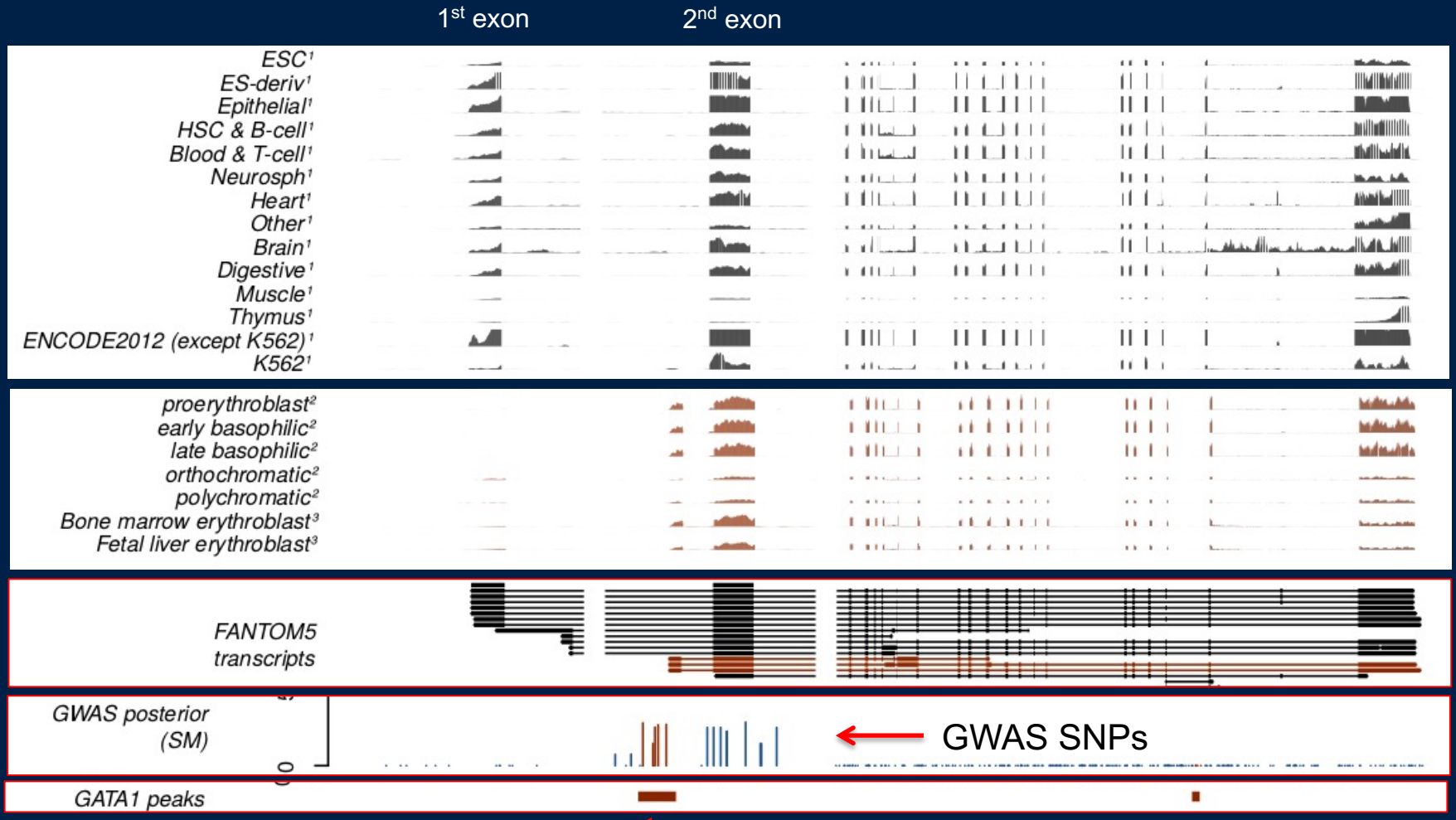
GATA1 is named after the DNA motif it recognises:



v1.factorbook.org

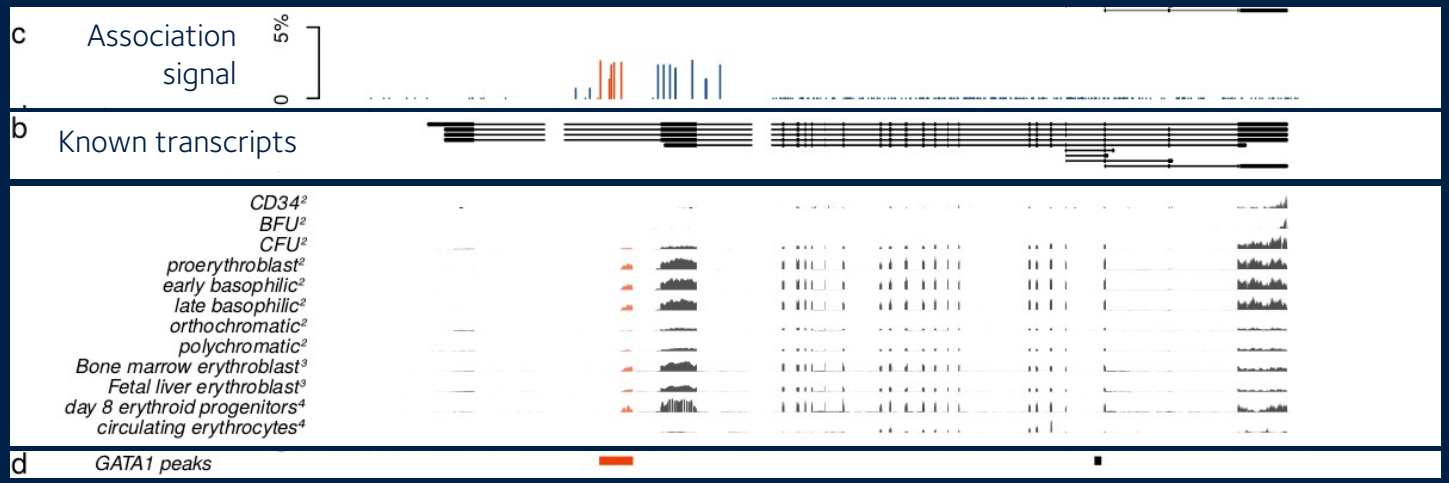
GATA1 binds just upstream of 2nd exon

Measured GATA1 binding



ChIP-seq experiments show GATA1 binds just upstream of our new exon. Moreover, one of the associated SNPs disrupts the GATA1 motif.

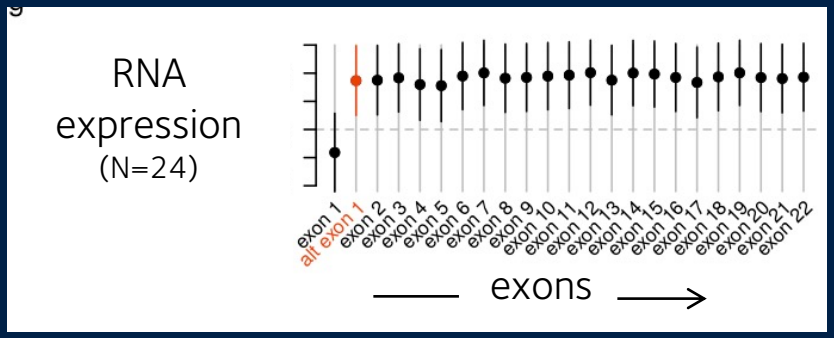
One of the malaria-associated SNPs disrupts the GATA site



rs10715451

...GGAGCG**G**TAAGATA... (malaria-protective allele)

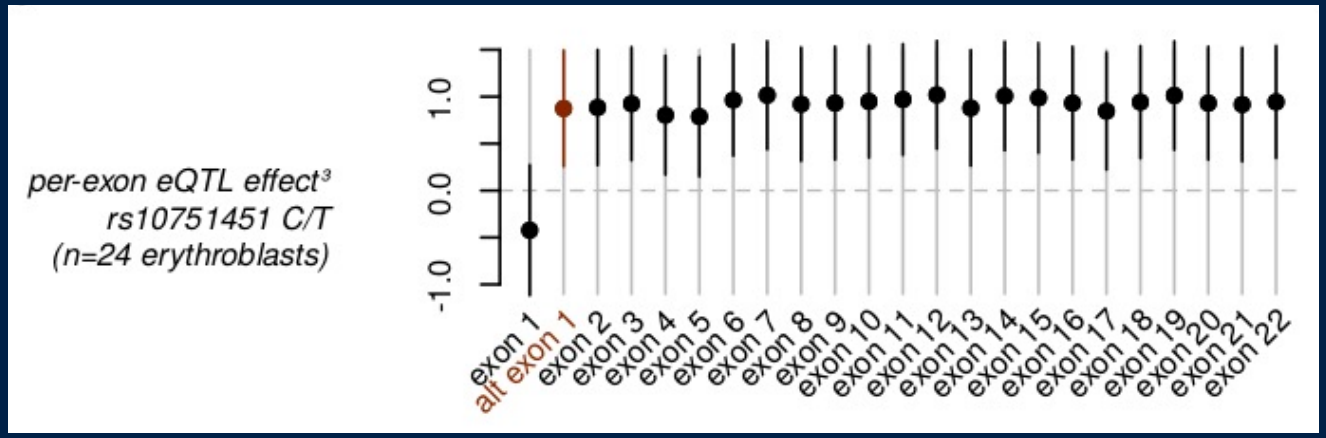
...GGAGCG**A**TAAGATA... (malaria risk allele)



Risk allele creates GATA motif and is associated with increased *ATP2B4* expression – of the erythroid transcript

Does this really hold up?

Prediction: the alternative (=risk) allele creates a GATA1 site. It would increase expression of *ATP2B4* starting at the new exon. But it wouldn't affect expression of the 'usual' 1st exon.



N = 24 experimentally differentiated erythrocyte precursor cells

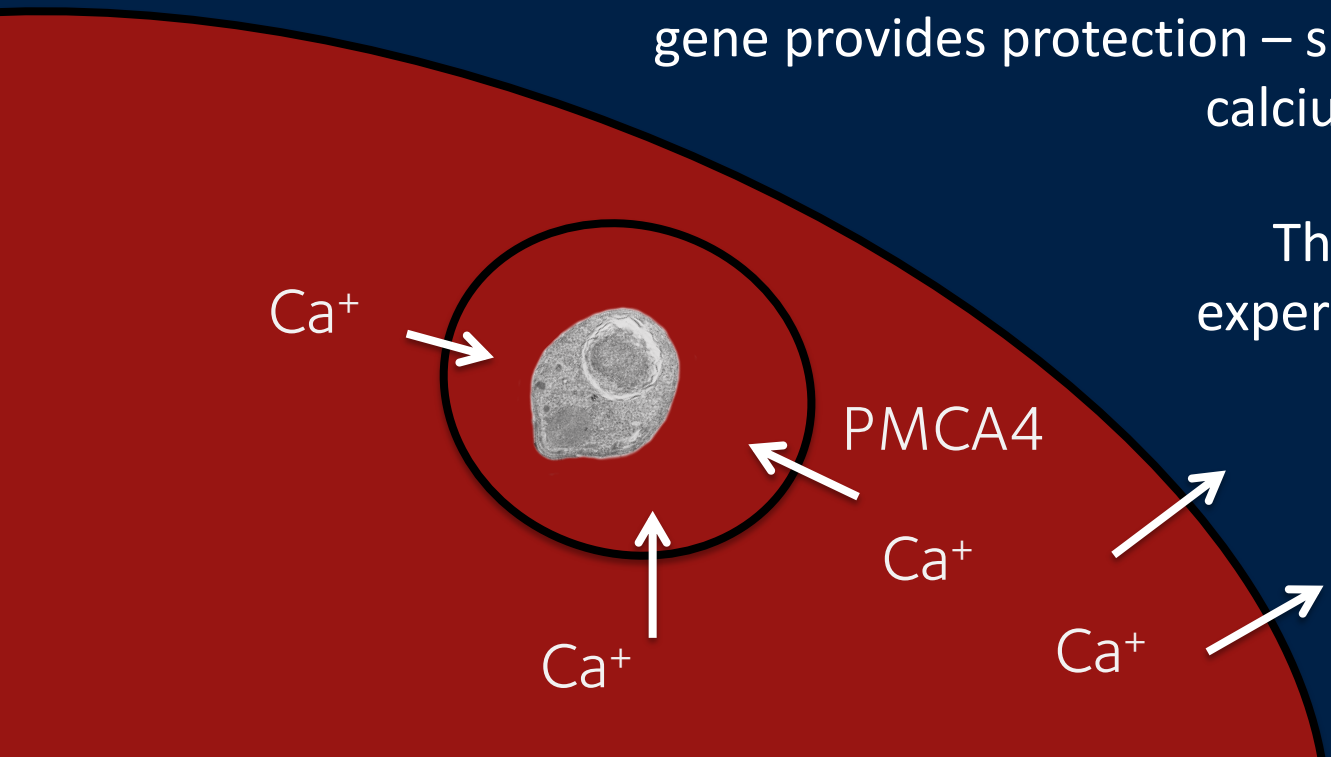
Functional hypothesis

ATP2B4 encodes a calcium pump (called PMCA4) in the RBC membrane. It acts to remove calcium from the cell.

When the parasite invades, the membrane gets **inverted** around the parasite, so presumably PMCA4 must also get inverted.

This might explain why lower expression of the gene provides protection – since parasites require calcium to grow effectively.

This is a hypothesis - not experimentally tested (yet)!



Zambo et al, Cell Calcium 2017

Biology from GWAS - summary

Non-coding variants

Long-distance interactions in the genome

Changes to gene expression

Polygenic effects (lots of variants involved)

Cell-type / tissue heterogeneity

Pleiotropy (a variant affects lots of phenotypes at once)

Genetic interactions

Host-pathogen interactions

Repetitive DNA / repeat expansions

Genome structural variation

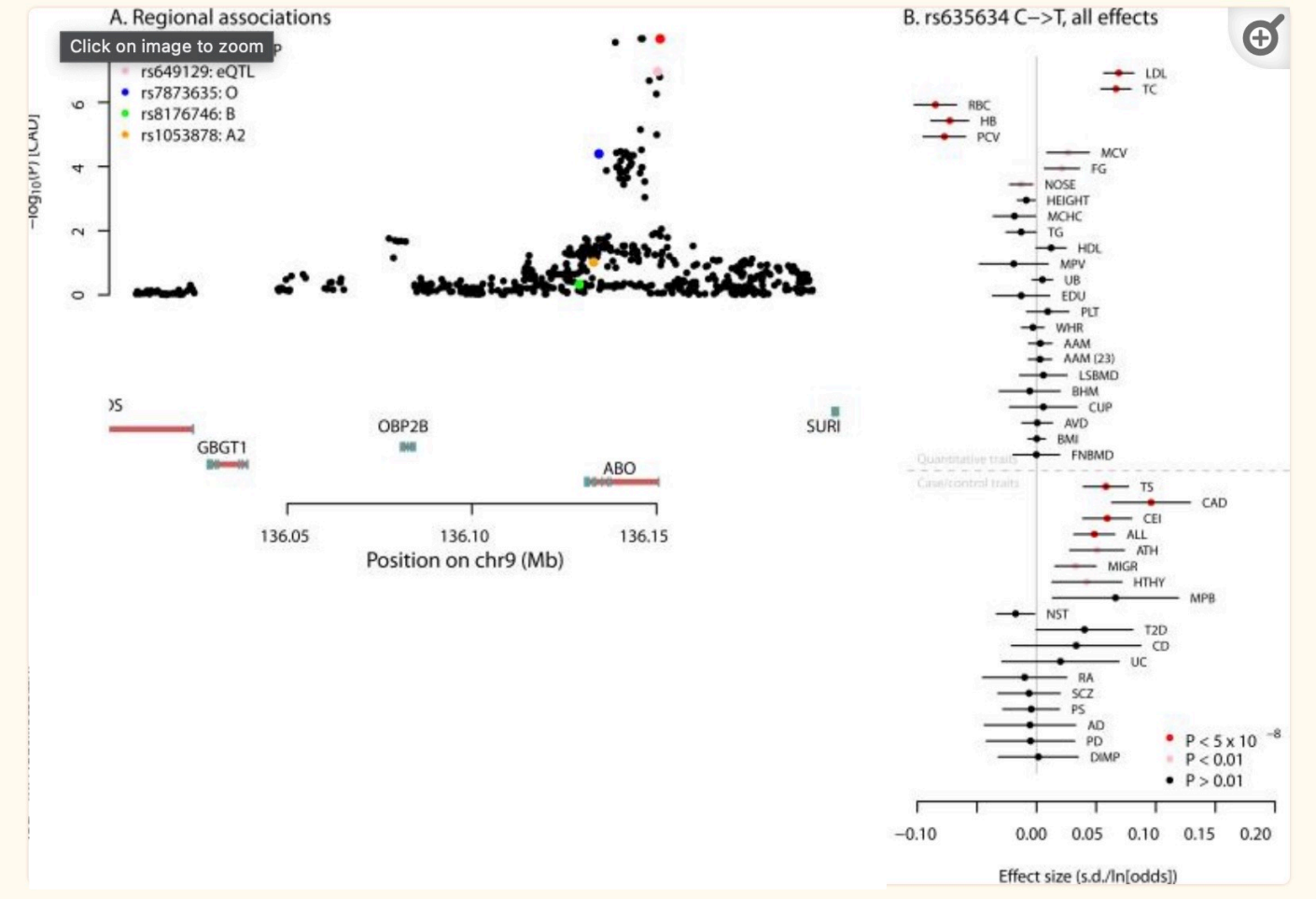
Genome evolution

Anything that can happen, does happen.
...and there is lots of data!

Lecture plan

- Recap & fallout from last lecture
- Gaining biological knowledge from GWAS
- Biological examples
- Pleiotropy, heritability and prediction

We should be looking across traits



Pickrell et al Nat. Genet. 2016

Prospective cohort studies

A new crop of studies aims to create a database of deep genotype, phenotype, and exposure data across large cohorts of individuals sampled from the population or from health services. Examples:

Precision Medicine Initiative (US)

CartaGene (Canada)

China Kadoorie Biobank

UK Biobank

The 100,000 genomes project (UK)

http://www.ukbiobank.ac.uk/

Collected 500,000 UK individuals who were 40-69 years old in 2006-2010.

Participants provided blood, urine and saliva samples. They also provided rich information on health and lifestyle.

Participants have been extensively genotyped and phenotyped

biobank uk

http://www.ukbiobank.ac.uk/

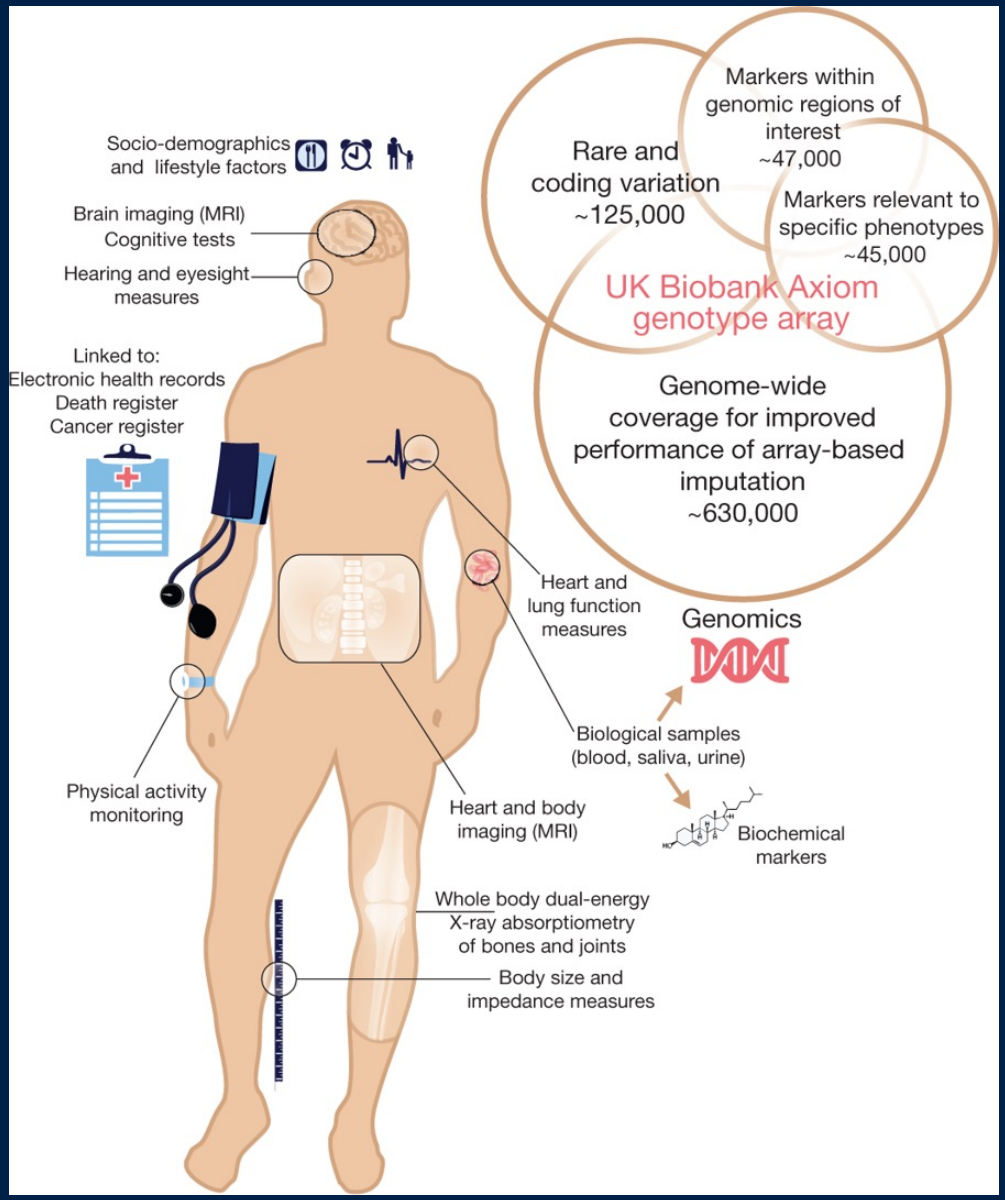
"The UK Biobank... aims to include 500,000 people from all around the UK... aged 40-69. This age group is being studied because it involves people ***at risk*** *over the next few decades of developing a wide range of important diseases (including cancer, heart disease, stroke, diabetes, dementia). The NHS treats the single largest group of people anywhere in the world, and keeps detailed records on all of them from birth to death... This will help researchers to understand the causes of diseases better, and to find new ways to prevent and treat many different conditions"*

biobank uk http://www.ukbiobank.ac.uk/

Genetic data

	N SNPs	N samples
Genotyping on a custom microarray (Affymetrix UK Biobank Axion array)	800,000	500,000
Imputation to almost all common and rare variants	100 million	500,000
Exome sequencing	Everything in gene exons	500,000 in future - by Regeneron
Genome sequencing	Everything	Sequencing is underway

biobank uk http://www.ukbiobank.ac.uk/

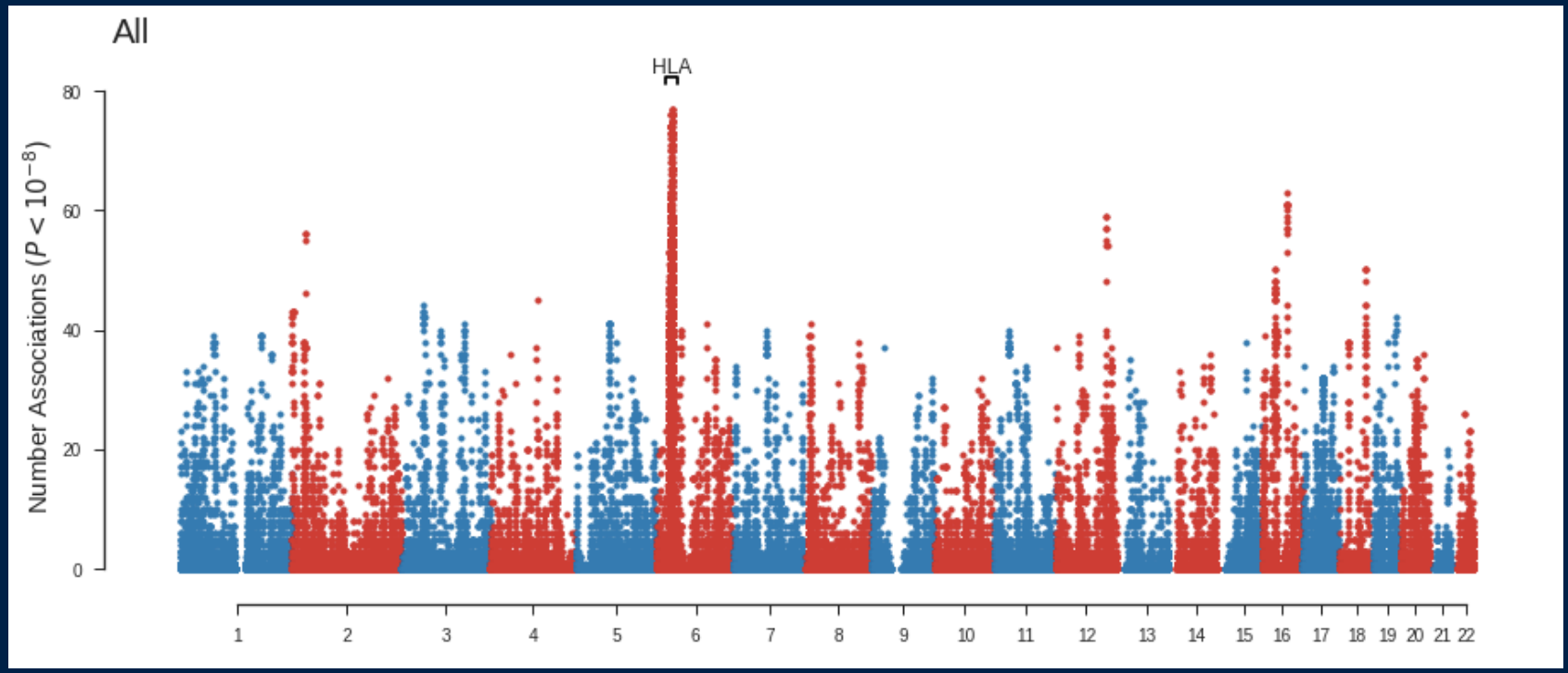


"As of May 2018, there were over 14,000 deaths, 79,000 participants with cancer diagnoses, and 400,000 participants with at least one hospital admission. Considerable efforts are now underway to incorporate data from a range of other national datasets including primary care, screening programmes, and disease-specific registries, as well as asking participants directly about health-related outcomes through online questionnaire. Efforts are also underway to develop scalable approaches that can characterize in detail different health outcomes by cross-referencing multiple sources of coded clinical information"

Bycroft et al Nature 2018

biobank uk

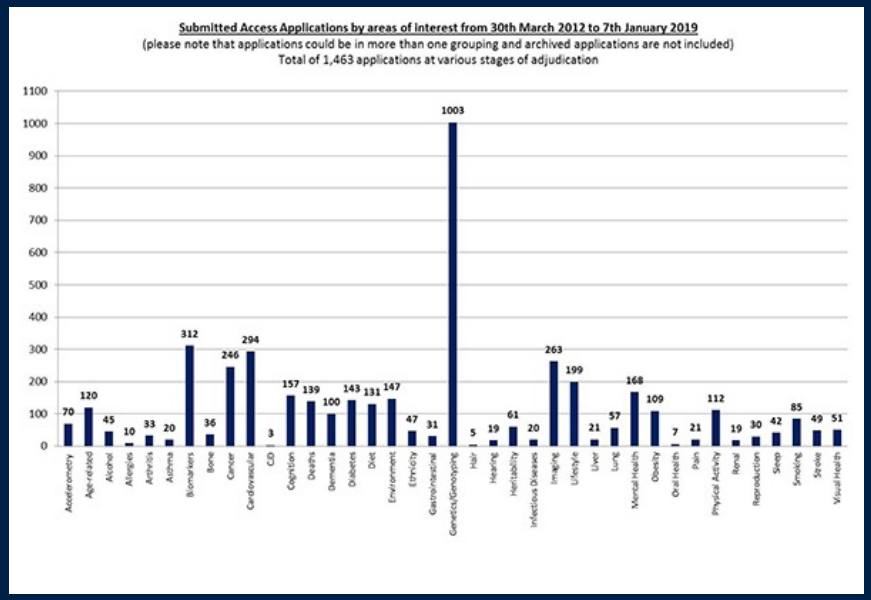
The UK biobank has let us discovery associations with 100s of traits across the whole genome, and indeed many variants are associated with many traits.

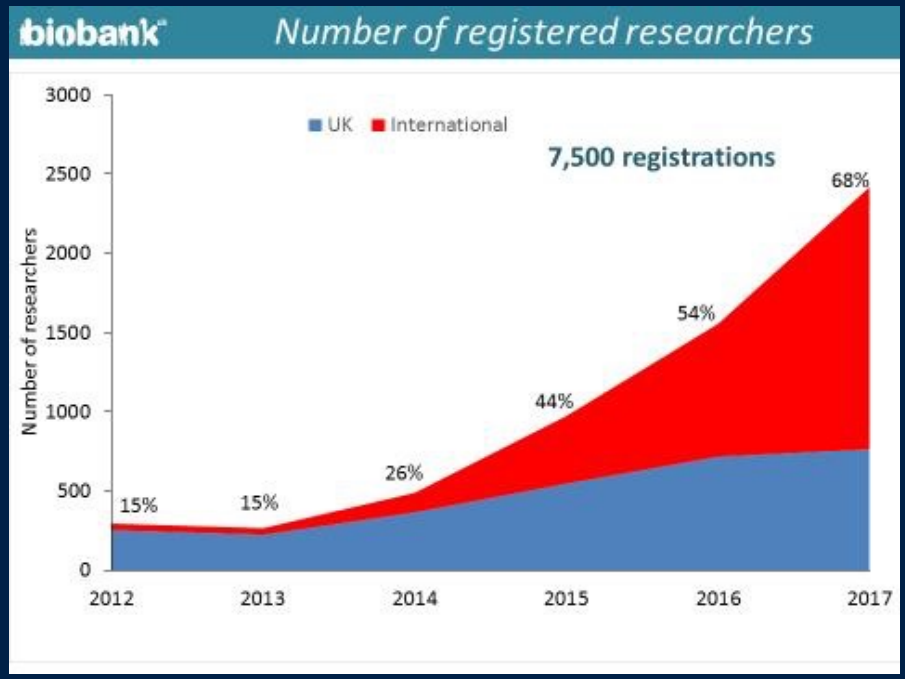


Number of statistically significant assocaitions among 717 traits
Canela-Xandri et al, http://geneatlas.roslin.ed.ac.uk/phewas/

biobank uk

Any researcher can apply for this data.

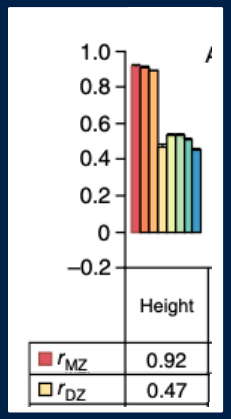




You can browse available data and apply at https://www.ukbiobank.ac.uk

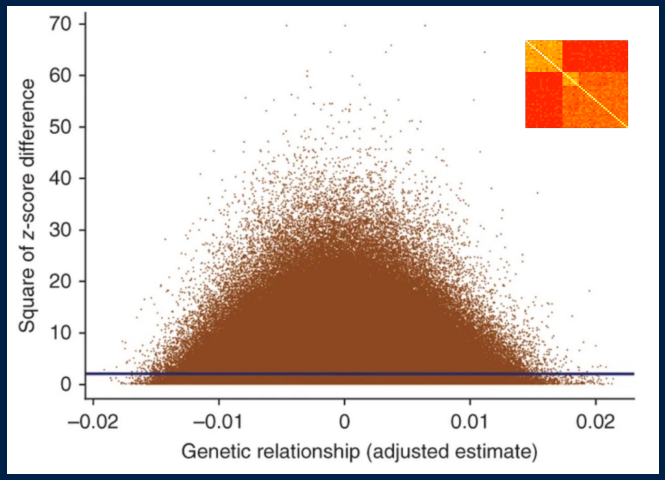
Finally – the largest GWAS conducted to date

Idea: if genetics determines a trait, then *more genetically similar individuals should have more similar phenotypes.* Can estimate how much genetics determines trait variation by comparing trait similarity in monozygotic (identical) and dizygotic twins.



← In twins

From GWAS →



(Adult) height is ~90% heritable

Common SNPs explain a large proportion of the heritability for human height (2010)

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]

About half of the 90% heritability is explained by common SNPs.

GWAS of height in 5.4 million individuals

A Saturated Map of Common Genetic Variants Associated with Human Height from 5.4 Million Individuals of Diverse Ancestries

ABSTRACT

Common SNPs are predicted to collectively explain 40-50% of phenotypic variation in human height, but identifying the specific variants and associated regions requires huge sample sizes. Here we show, using GWAS data from 5.4 million individuals of diverse ancestries, that 12,111 independent SNPs that are significantly associated with height account for nearly all of the common SNP-based heritability. These SNPs are clustered within 7,209 non-overlapping genomic segments with a median size of ~90 kb, covering ~21% of the genome. The density of independent associations varies across the genome and the regions of elevated density are enriched for biologically relevant genes. In out-of-sample estimation and prediction, the 12,111 SNPs account for 40% of phenotypic variance in European ancestry populations but only ~10%-20% in other ancestries. Effect sizes, associated regions, and gene prioritization are similar across ancestries, indicating that reduced prediction accuracy is likely explained by linkage disequilibrium and allele frequency differences within associated regions. Finally, we show that the relevant biological pathways are detectable with smaller sample sizes than needed to implicate causal genes and variants. Overall, this study, the largest GWAS to date, provides an unprecedented saturated map of specific genomic regions containing the vast majority of common height-associated variants.

This very preprint appeared on bioRxiv in January 2022

It claims to map essentially all of the common mutations that determine human height.

There are 12,111 of them and (grouped into regions) they cover 21% of the genome.

Yengo et al bioRxiv (2021) https://doi.org/10.1101/2022.01.07.475305

GWAS of height in 5.4 million indiiduals

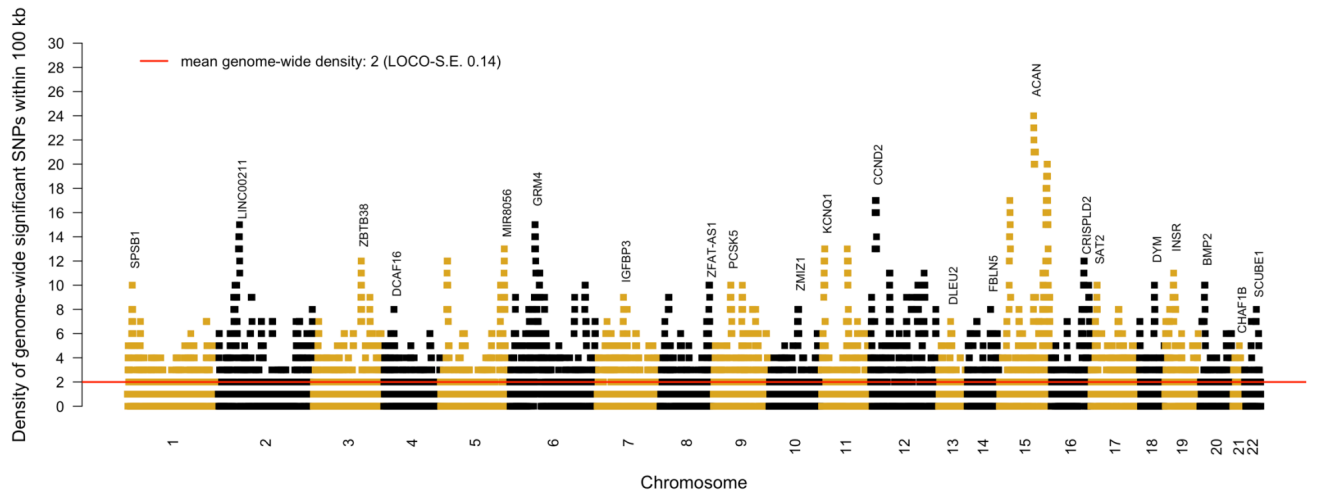
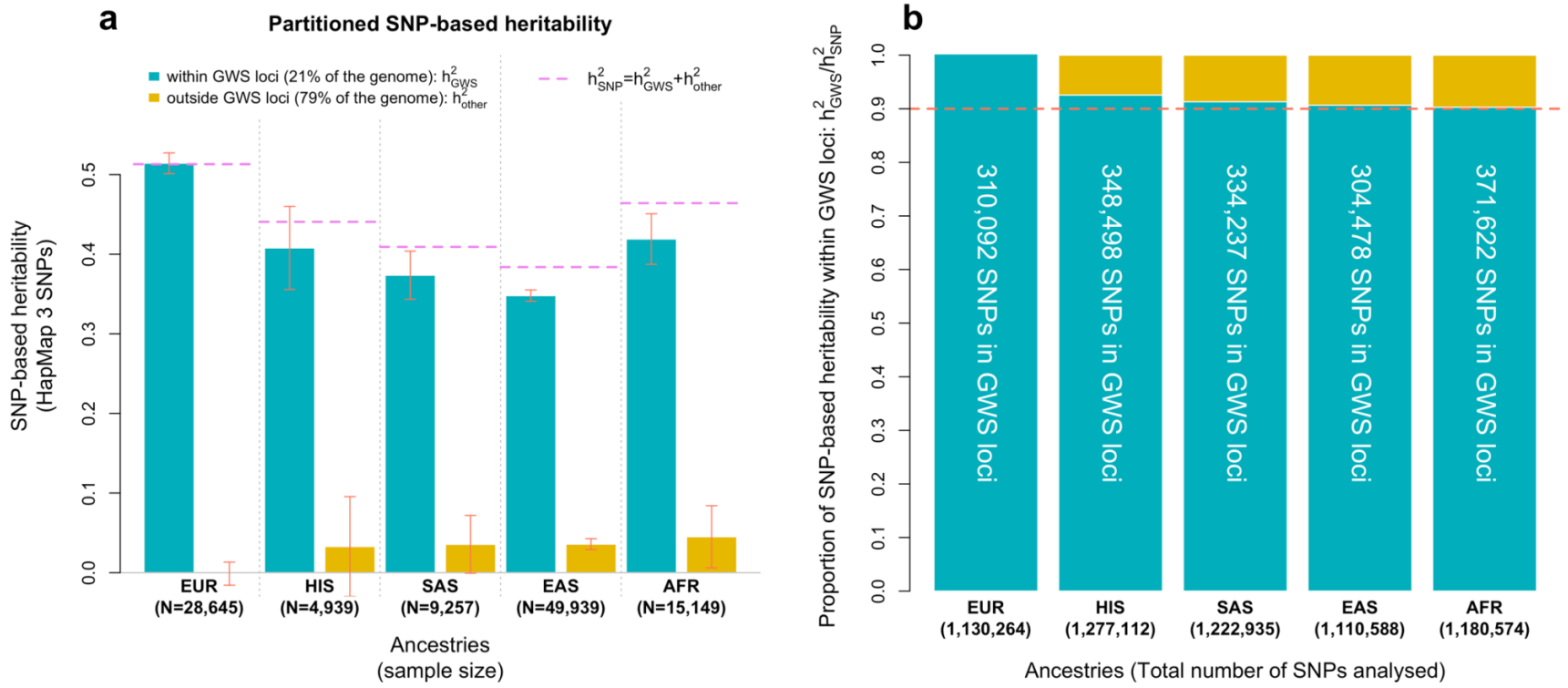


Fig. 1. Brisbane plot showing the genomic density of independent genetic associations with height. Each dot represents one of the 12,111 quasi-independent genome-wide significant (GWS; $P<5\times10^{-8}$) height-associated SNPs identified using approximate conditional and joint multiple-SNP (COJO) analyses of our trans-ancestry GWAS meta-analysis. Density was calculated for each associated SNP as the number of other independent associations within 100 kb. A density of 1 means that a GWS COJO SNP share its location with another independent GWS COJO SNP within <100 kb. The average signal density across the genome is 2 (standard error; S.E. 0.14). S.E. were calculated using a Leave-One-Chromosome-Out jackknife approach (LOCO-S.E.). Sub-significant SNPs are not represented on the figure.

12,111 SNPs in regions covering ~21% of genome

GWAS of height in 5.4 million indiiduals

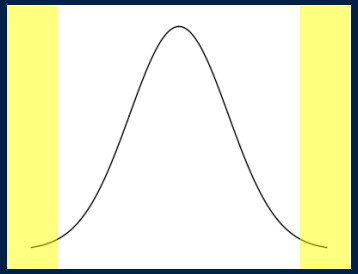


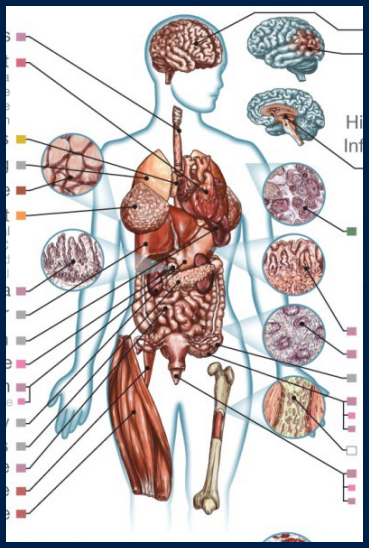
The regions identified explain a very large proportion of the heritability of height – especially in European populations. (The rest of the heritability is probably in rarer variants not accessed by this study).

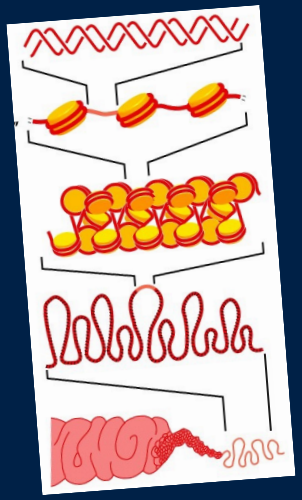
Conclusions and summary

- Most human traits are highly heritable
- For ‘complex’ traits, the effects are made up of many genetic variants often with modest effects
- GWAS study designs can find these variants. 100s of 1000s of trait-associated SNPs have now been identified. They rely on large samples and dense genotyping, and exploit ancestral recombination between samples to narrow down signals.
- A major frontier is to understand the biology and translate these findings into clinically useful insights and predictions.

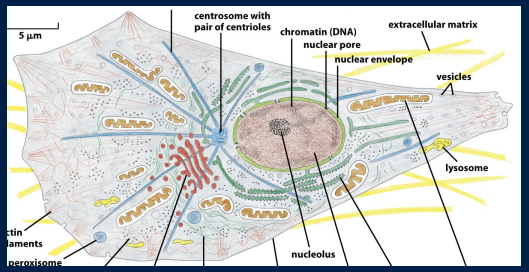
(We need lots of quantitively-minded people to do this!)







Thanks!


5 µm
centrosome with pair of centrioles
chromatin (DNA)
nuclear pore
nuclear envelope
extracellular matrix
vesicles
lysosome
nucleolus