Understanding the genetics of complex traits II

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

BA Human Sciences

8th March 2024

Lecture plan

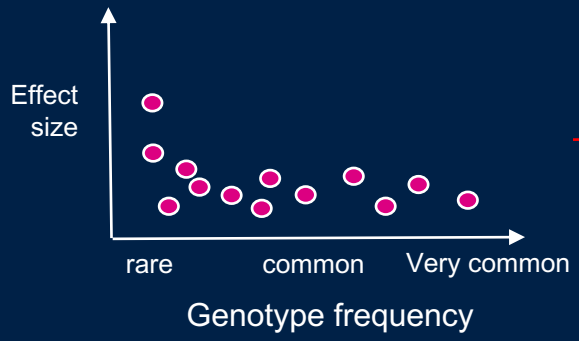
- Recap from last lecture – GWAS and the common variant / common trait hypothesis
- How polygenic are traits anyway?
- The challenge of fine-mapping

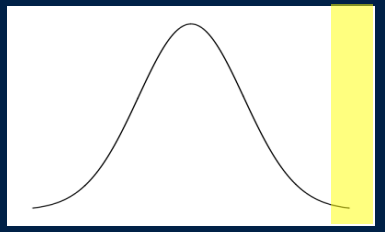
Recap

For many ‘complex’ traits, heritability seems to be due to lots of variants across the genome with small effects

i.e. this:

“polygenic” “multifactorial” “complex”

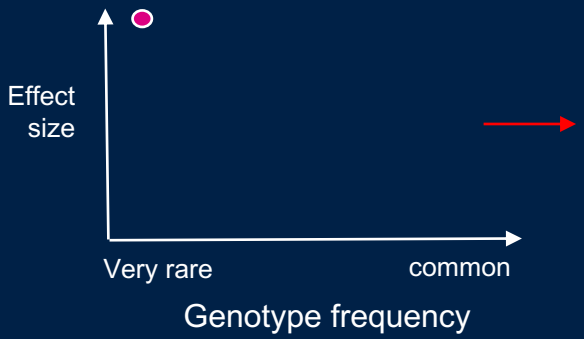


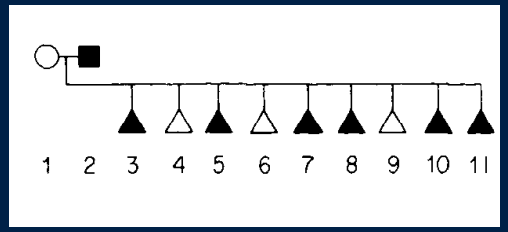


Common trait varying in the population

As opposed to this:

“mendelian” “high penetrance” ”monogenic”

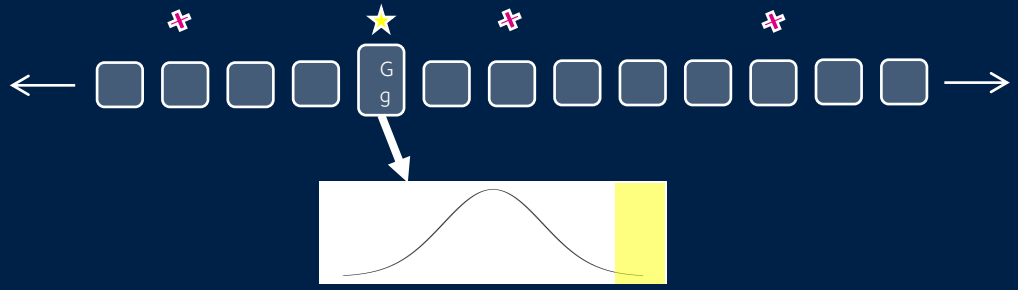




Trait inherited in families

This is the ‘common variant, common disease hypothesis’, first proposed in the 1990s.

We talked about the basic GWAS approach



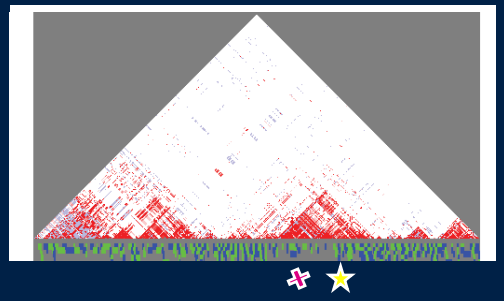
Basic idea: genotype as much genetic variation as possible in disease cases and controls. Then estimate the relative risk of each variant.

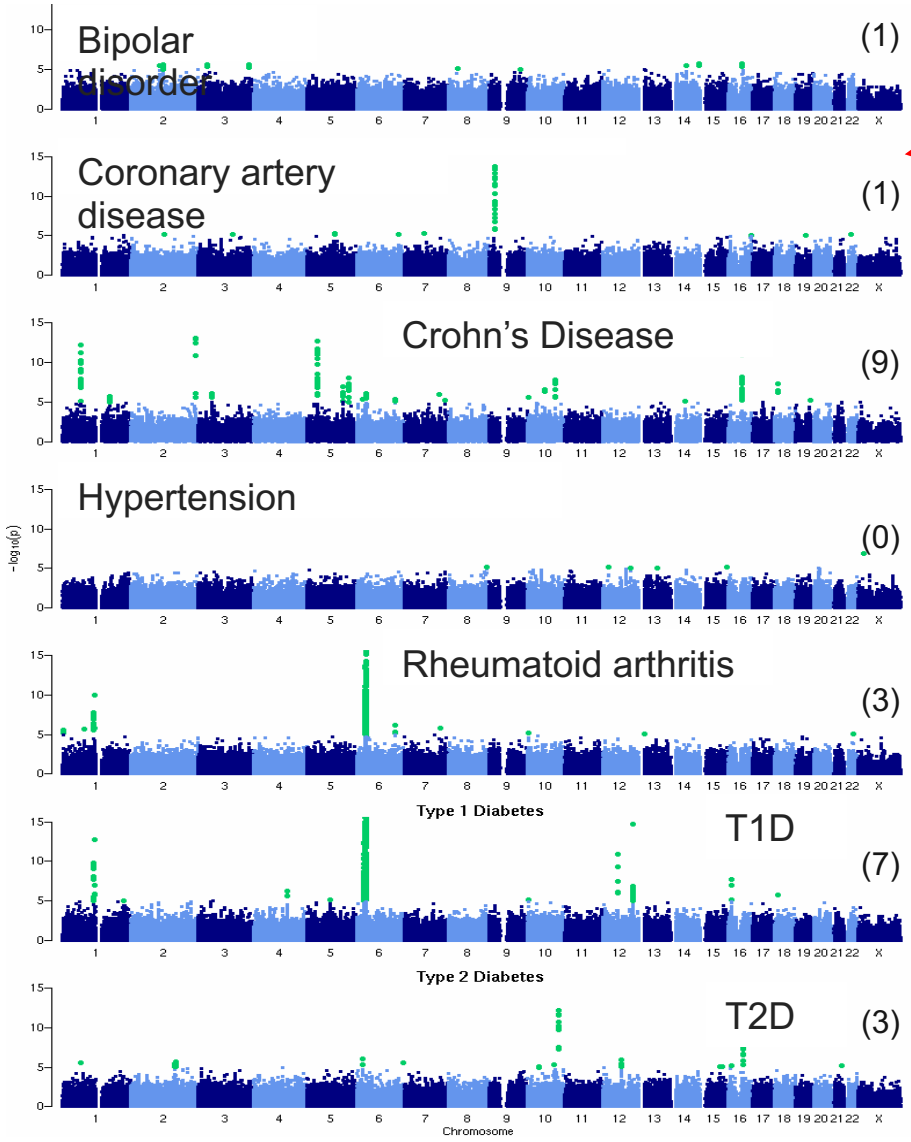
$$\textit{Relative risk} = \frac{P(\text{disease}|\text{genotype } G)}{P(\text{disease}|\text{genotype } g)}$$

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

If statistical evidence is strong that $RR \neq 1$, we may have found an association.

Rely on LD patterns to access associations even if we didn't type the causal genetic variant.





Evidence for association
$(-\log_{10}$ P-value$)$

The Wellcome Trust Case-Control Consortium study (2007)

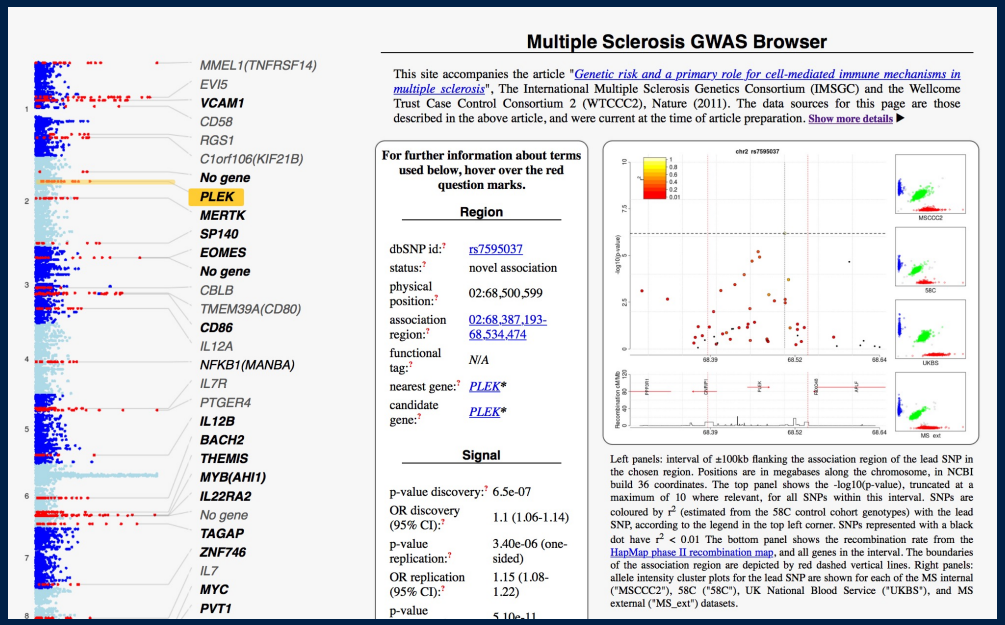
This study proved that the hypothesis really was true! The methodology worked and multiple genetic associations were found across the genome.

Although traits varied a bit in how many associations were seen.

To detect effects we need large samples – here N=5,000 per disease

The genome-wide association study design really can find these common genetic associations

Consolidation question



GWAS of multiple sclerosis (2011)

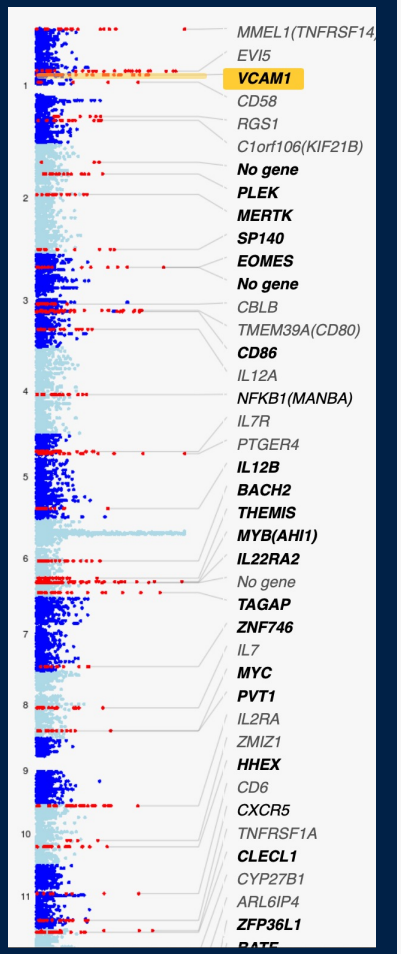
https://www.chg.ox.ac.uk/wtccc2/ms/
(I think this requires the trailing /)

Visit the above site and make sure you understand what is shown. Pick a signal and try to work out

- What is the sample size?
- How strong was the evidence?
- Does the genotyping look accurate?
- Does the association follow LD patterns as you'd expect?
- What is the estimated effect size?
- Did it replicate? How do discovery and replication effect sizes compare?
- What genes are nearby? Can you figure out what they do? (**Warning**: this can be a time sink!)

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

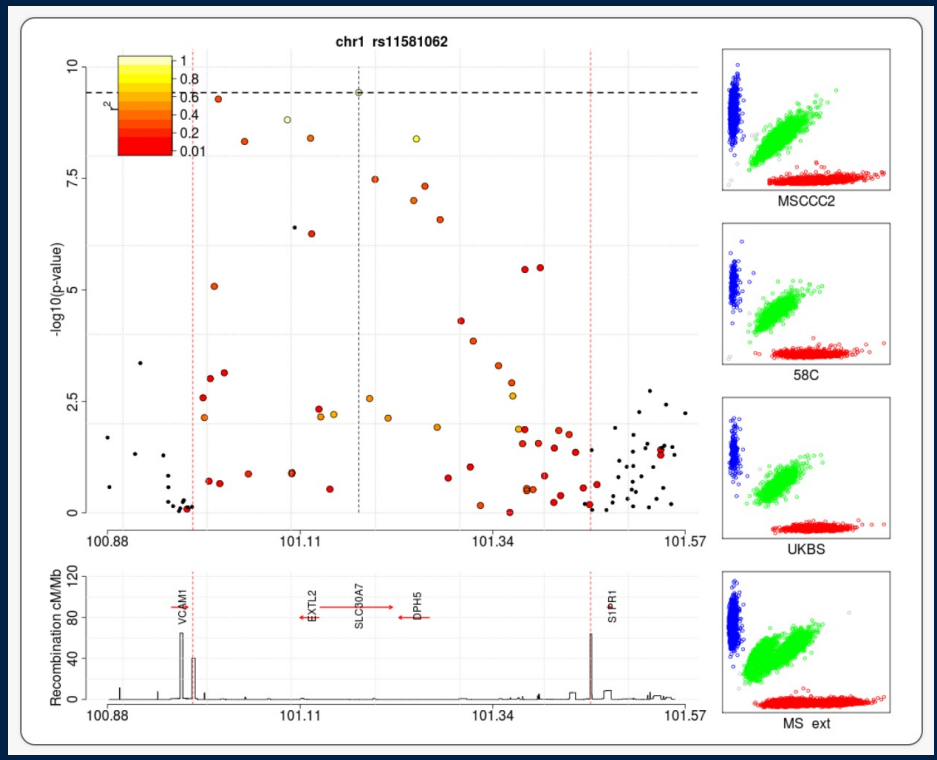
p-value discovery:?	3.7e-10
OR discovery (95% CI):?	1.13 (1.09-1.18)
p-value replication:?	4.20e-02 (one-sided)
OR replication (95% CI):?	1.07 (0.99-1.15)
p-value combined:?	2.50e-10
OR combined (95% CI):?	1.12 (1.1-1.13)
Risk (non-risk) allele:	G(A)

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

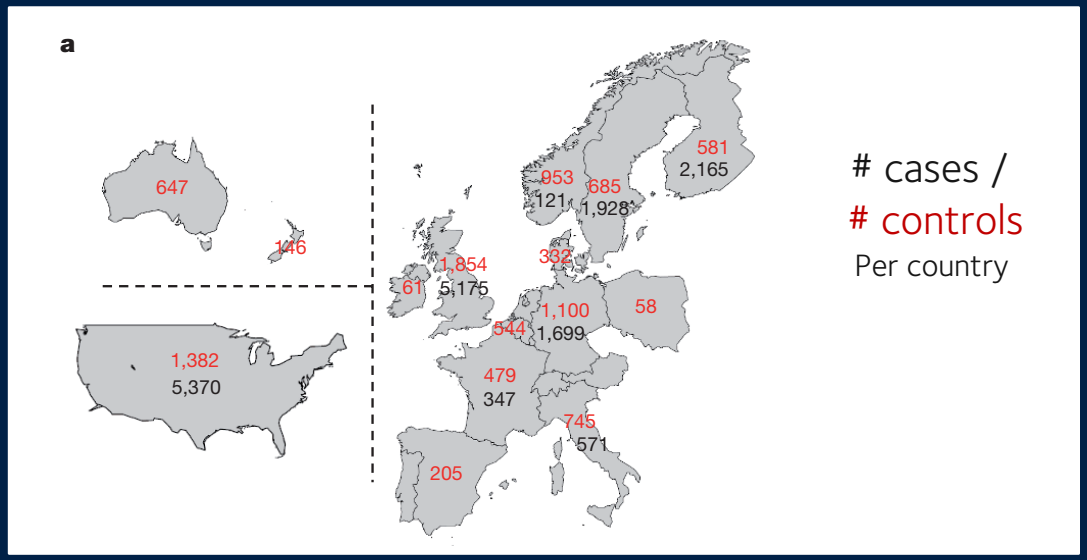
Proximal genes?

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

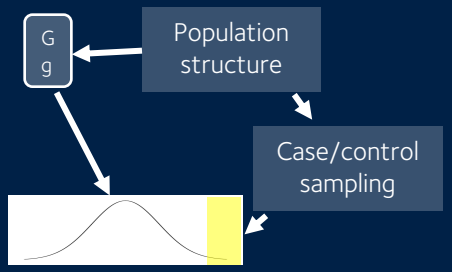


Can you explain?

Dealing with population structure



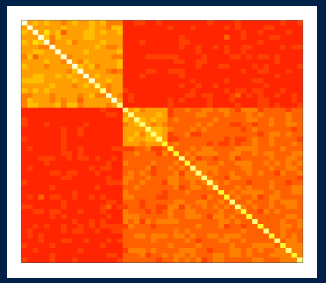
Answer: very strong confounding by population structure / sampling



Solution:

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

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

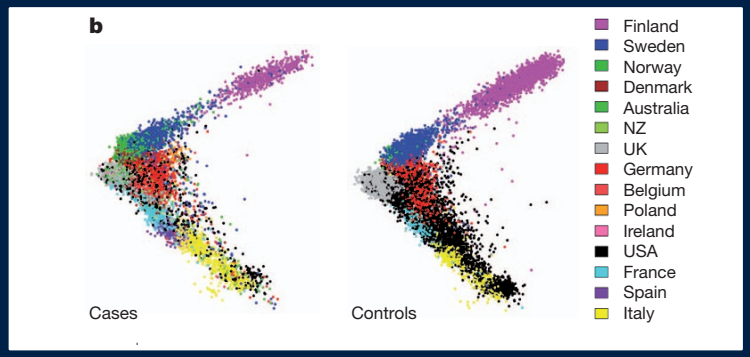


Matrix of relationships between samples

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

1. Logistic regression including principal components

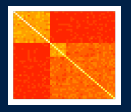
$$\text{outcome} \sim \text{genotype} + PCs$$



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

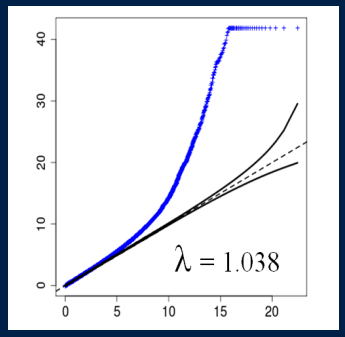
Uses just the strongest directions of variation in relatedness (population structure)

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

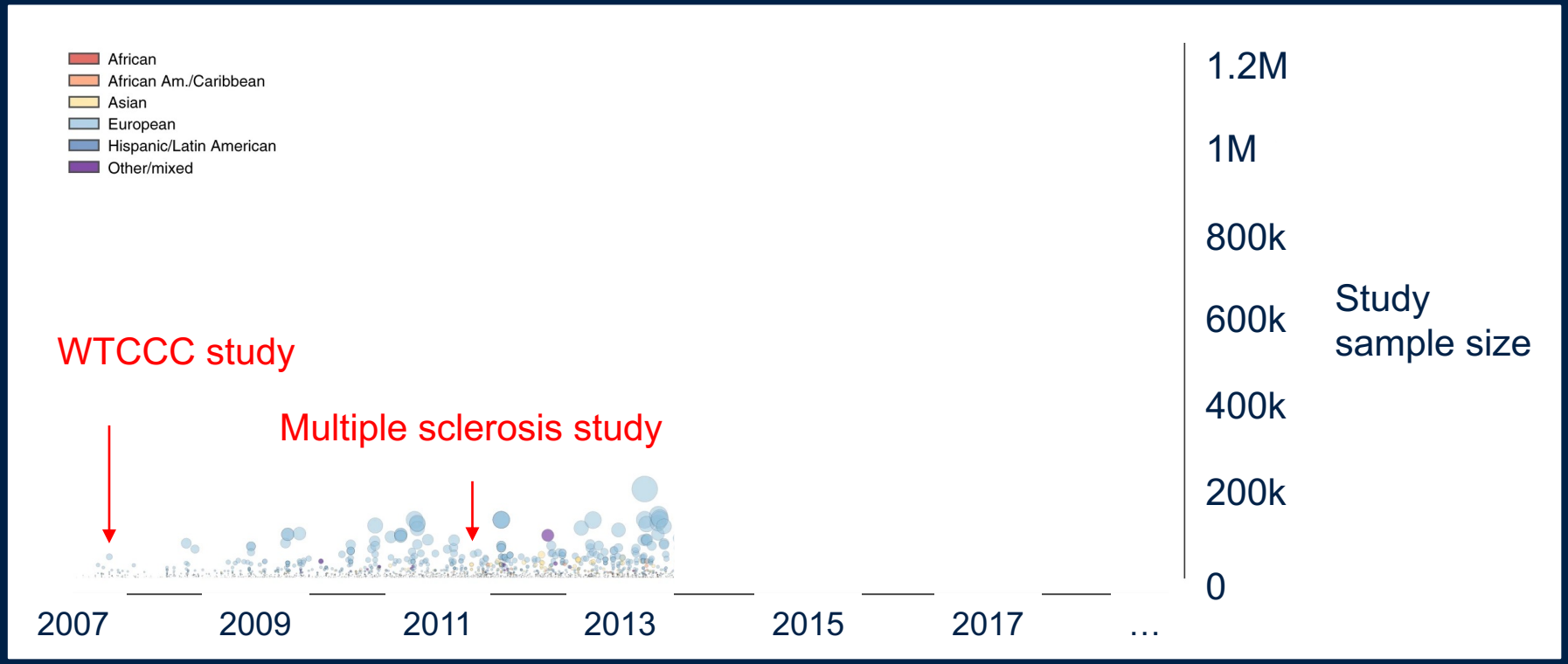


Most p-values are now not inflated

Lecture plan

- Recap from last lecture – GWAS and the common variant / common trait hypothesis
- How polygenic are traits anyway?
- The challenge of fine-mapping

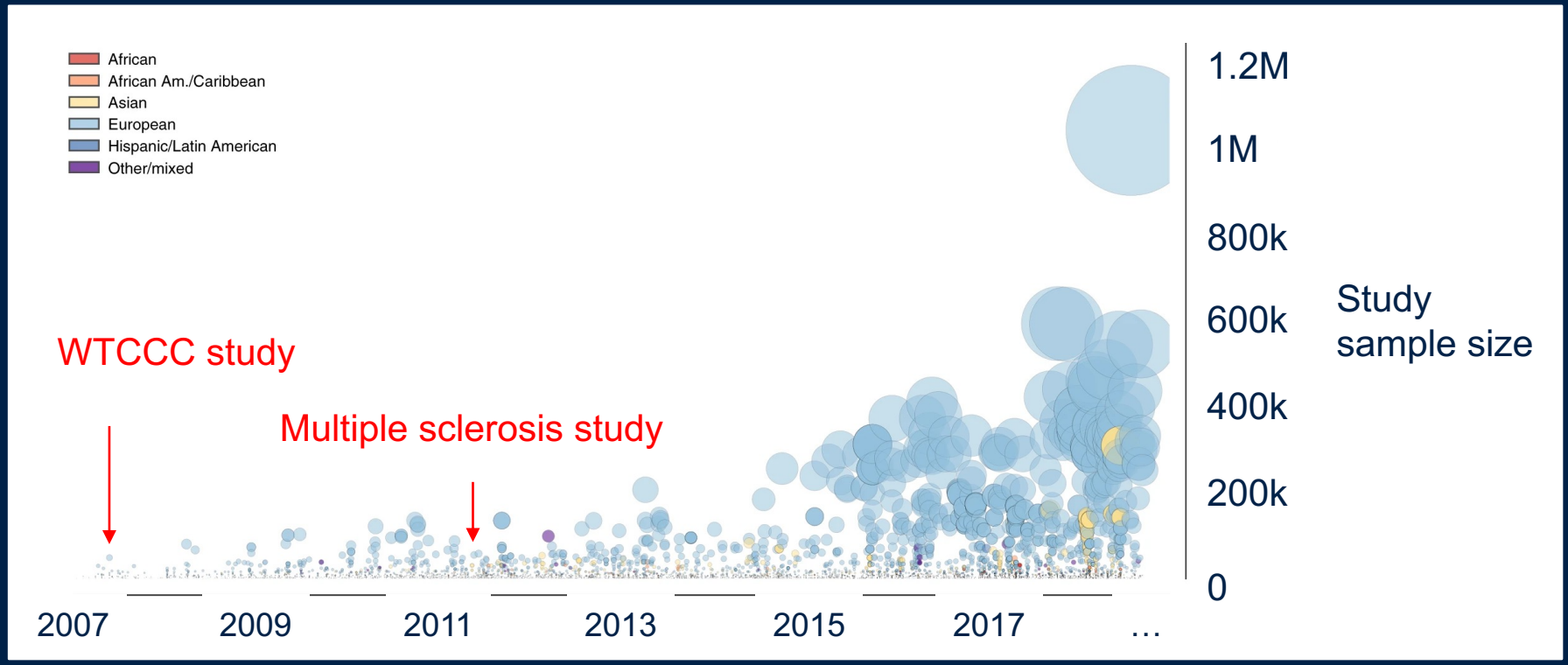
What happened next?



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

NHGRI GWAS Catalog: https://www.ebi.ac.uk/gwas/

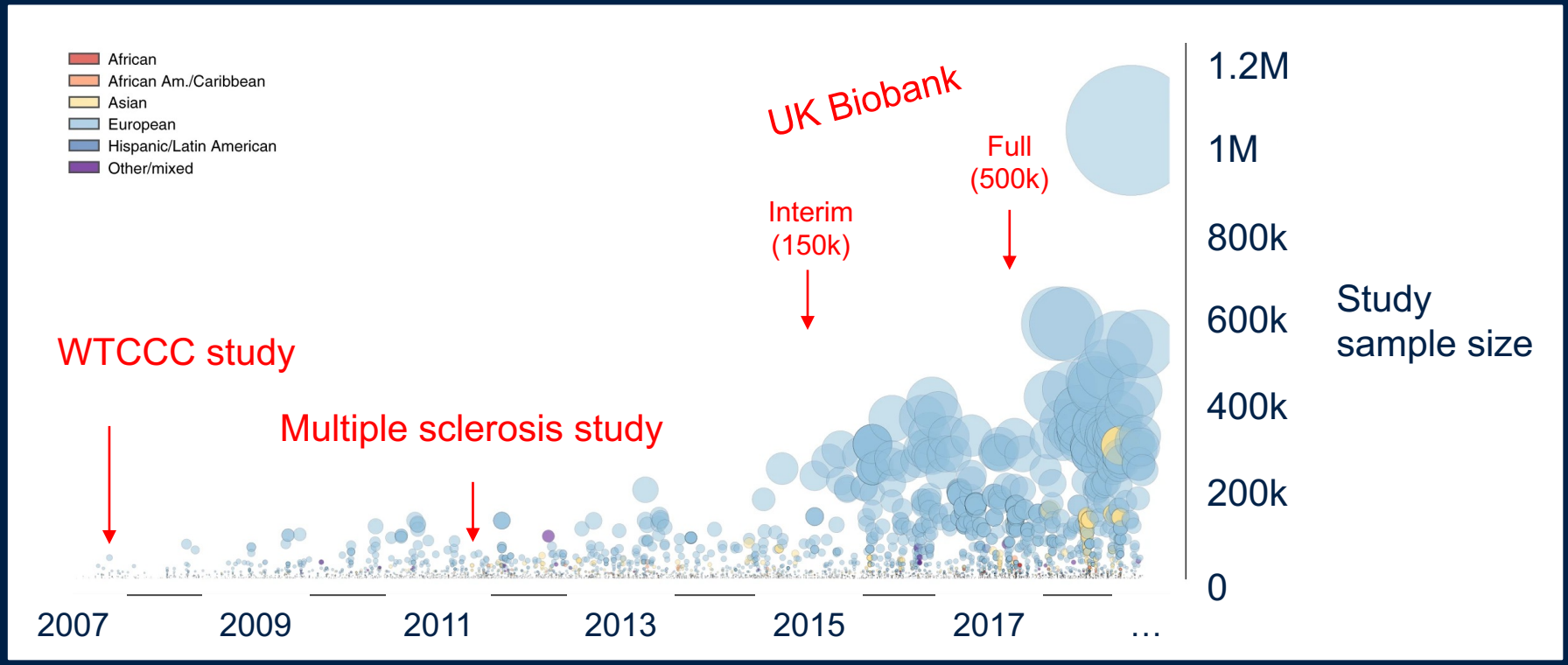
GWAS went large scale



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

NHGRI GWAS Catalog: https://www.ebi.ac.uk/gwas/

GWAS went large scale



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

NHGRI GWAS Catalog: https://www.ebi.ac.uk/gwas/

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)

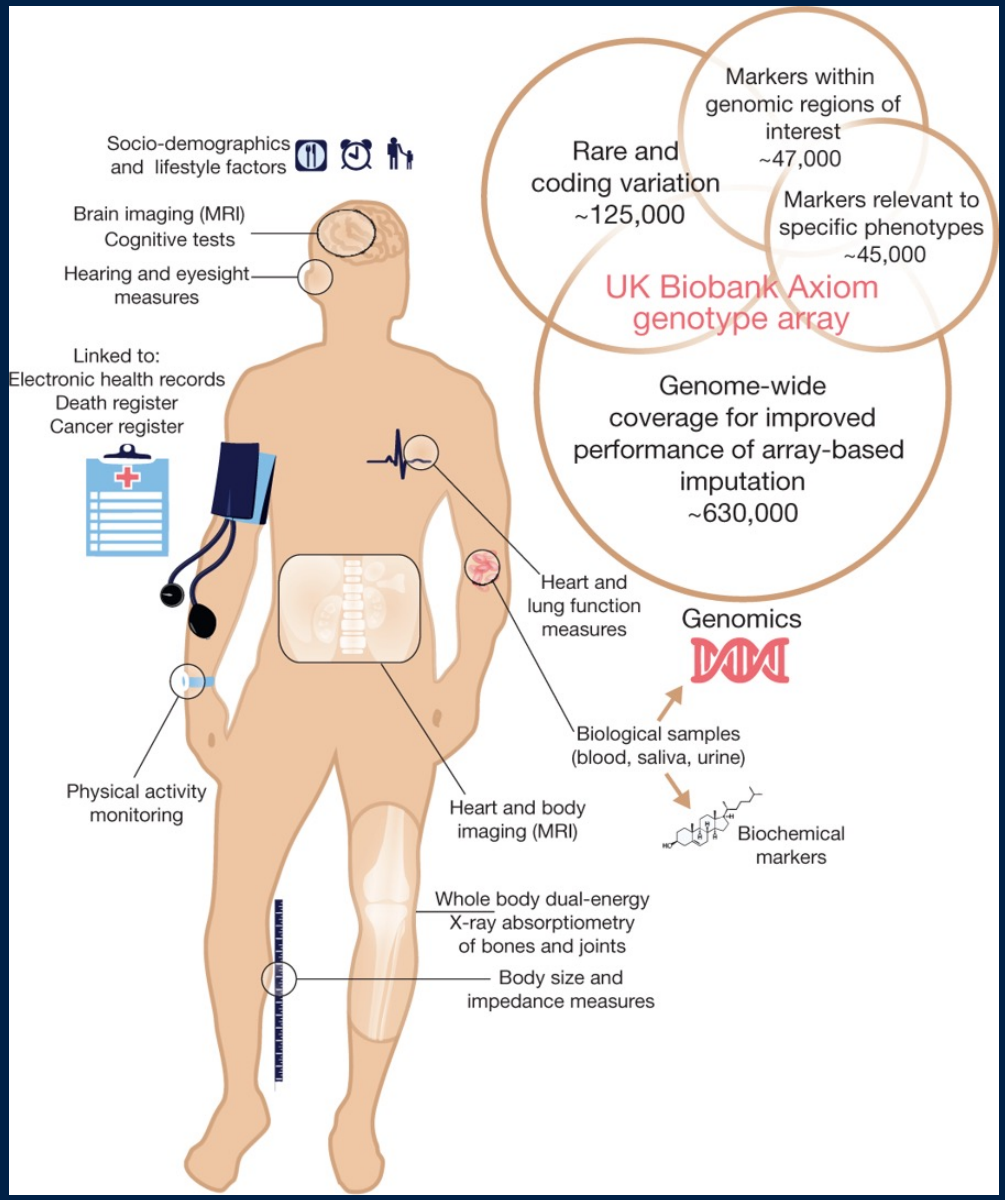
CartaGene (Canada)

China Kadoorie Biobank

UK Biobank

The 100,000 genomes project (UK)

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

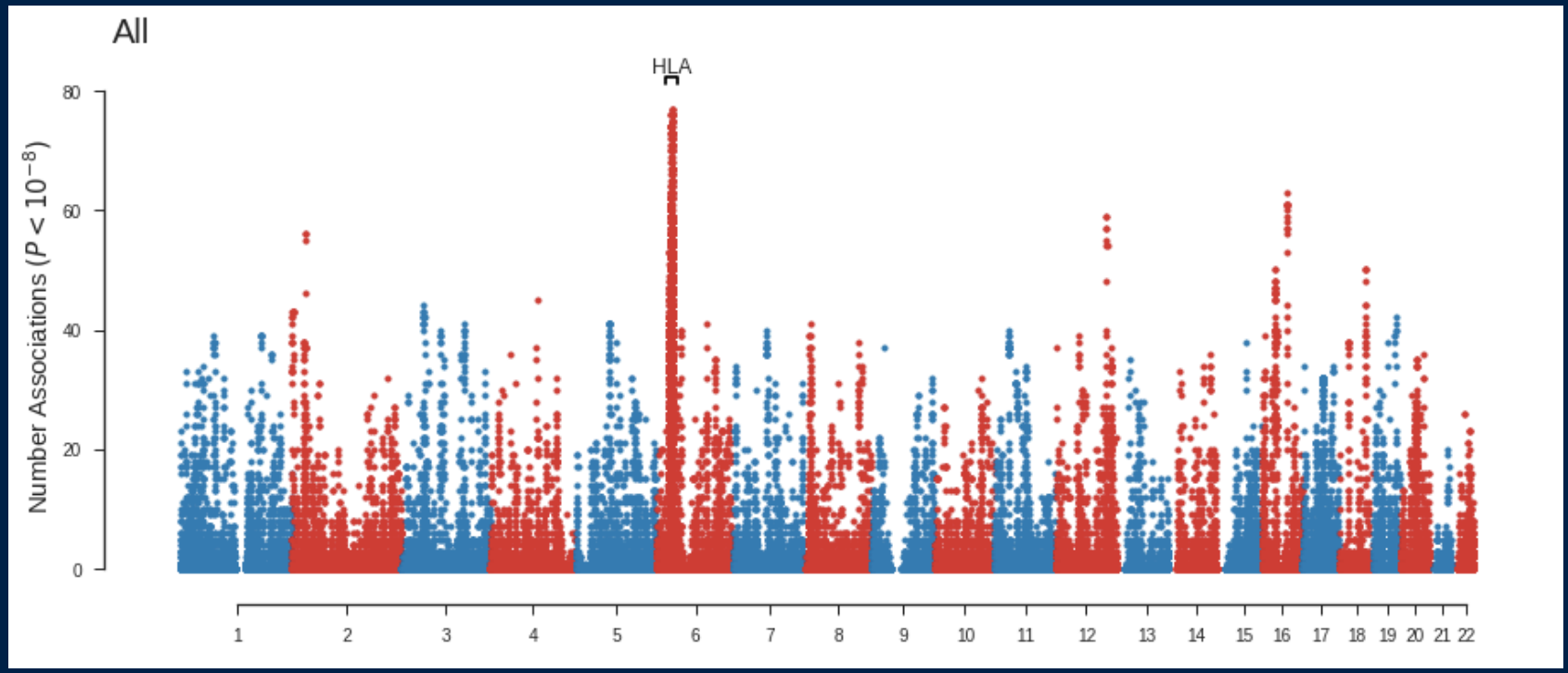


"As of May 2018, there were over 14,000 deaths, 79,000 participants with cancer diagnoses, and 400,000 participants with at least one hospital admission. Considerable efforts are now underway to incorporate data from a range of other national datasets including primary care, screening programmes, and 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 discover 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/

… so how polygenic do traits get?

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

To discover this we would need a large sample size!

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

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.

Height is the epitome of polygenicity

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

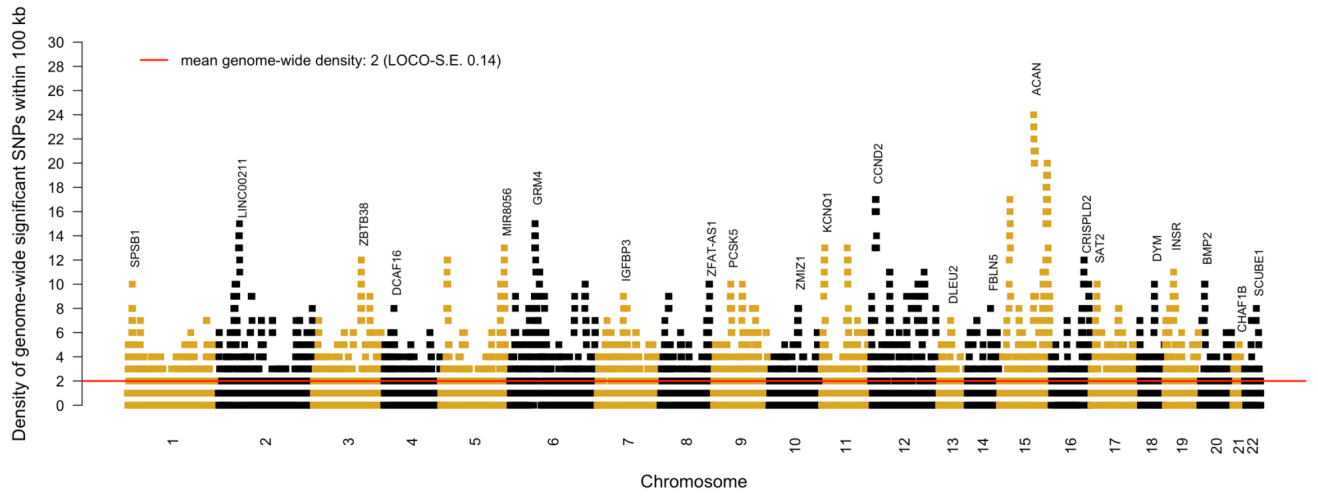
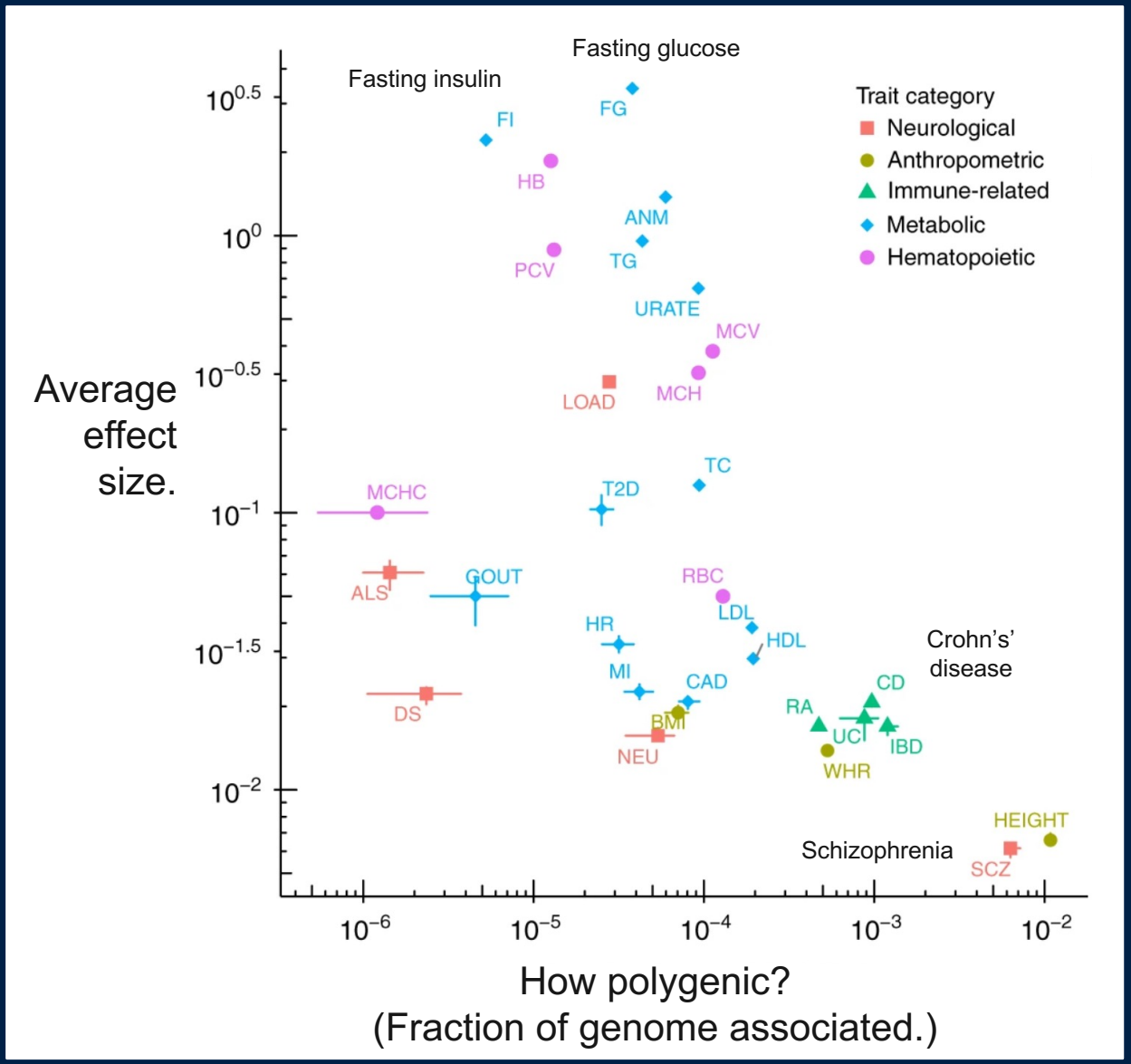


Fig. 1. Brisbane plot showing the genomic density of independent genetic associations with height. Each dot represents one of the 12,111 quasi-independent genome-wide significant (GWS; $P<5\times10^{-8}$) height-associated SNPs identified using approximate conditional and joint multiple-SNP (COJO) analyses 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



The wealth of GWAS data allows studies that estimate the genetic architecture of different traits.

Here – "polygenicity" (x axis) versus average effect size (y axis)

Zhu and Stephens, Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes (2018) https://doi.org/10.1038/s41467-018-06805-x

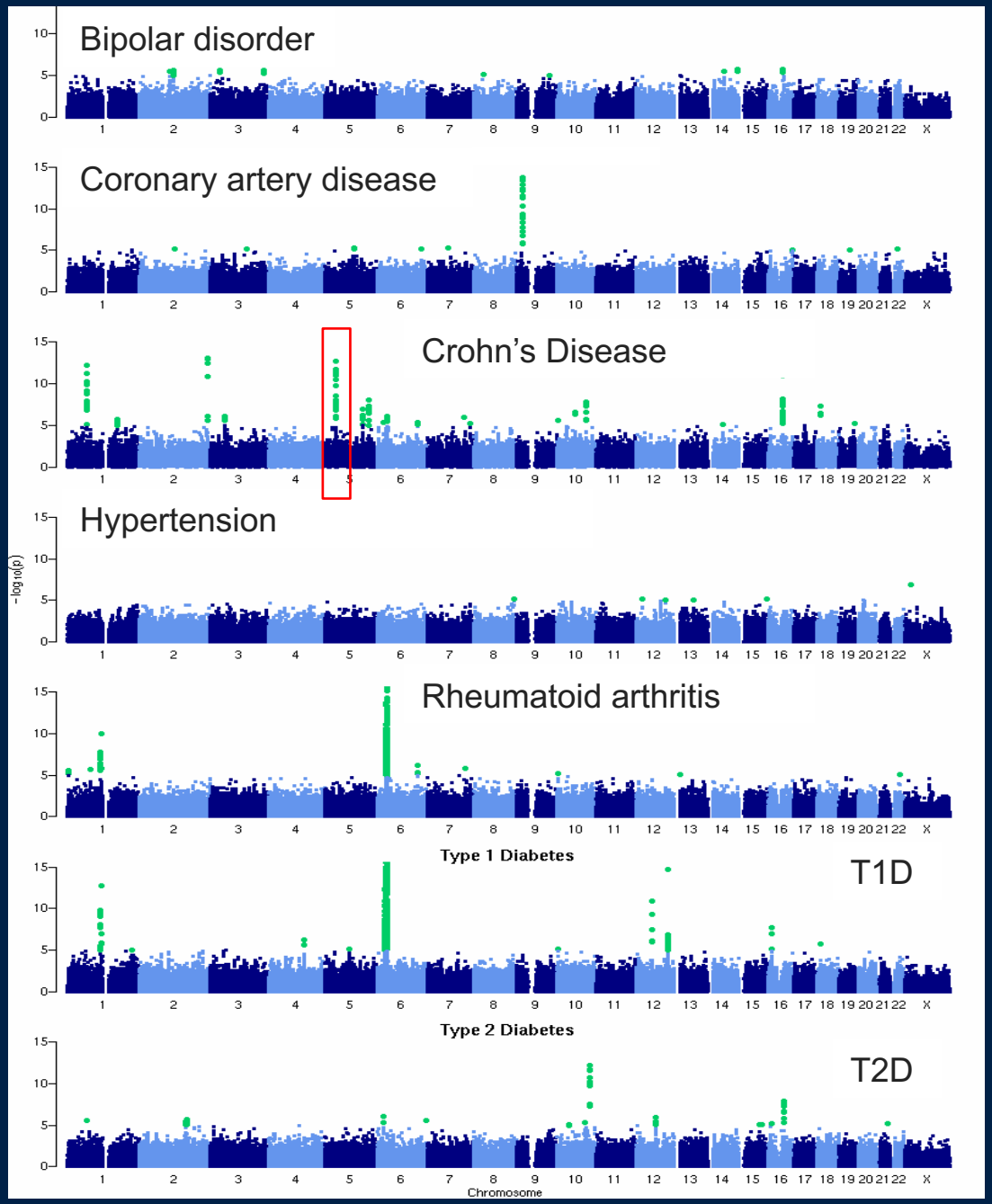
Lecture plan

- Recap from last lecture – GWAS and the common variant / common trait hypothesis
- How polygenic are traits anyway?
- The challenge of fine-mapping

GWAS have clearly told us a great deal about the genetic architecture of complex traits.

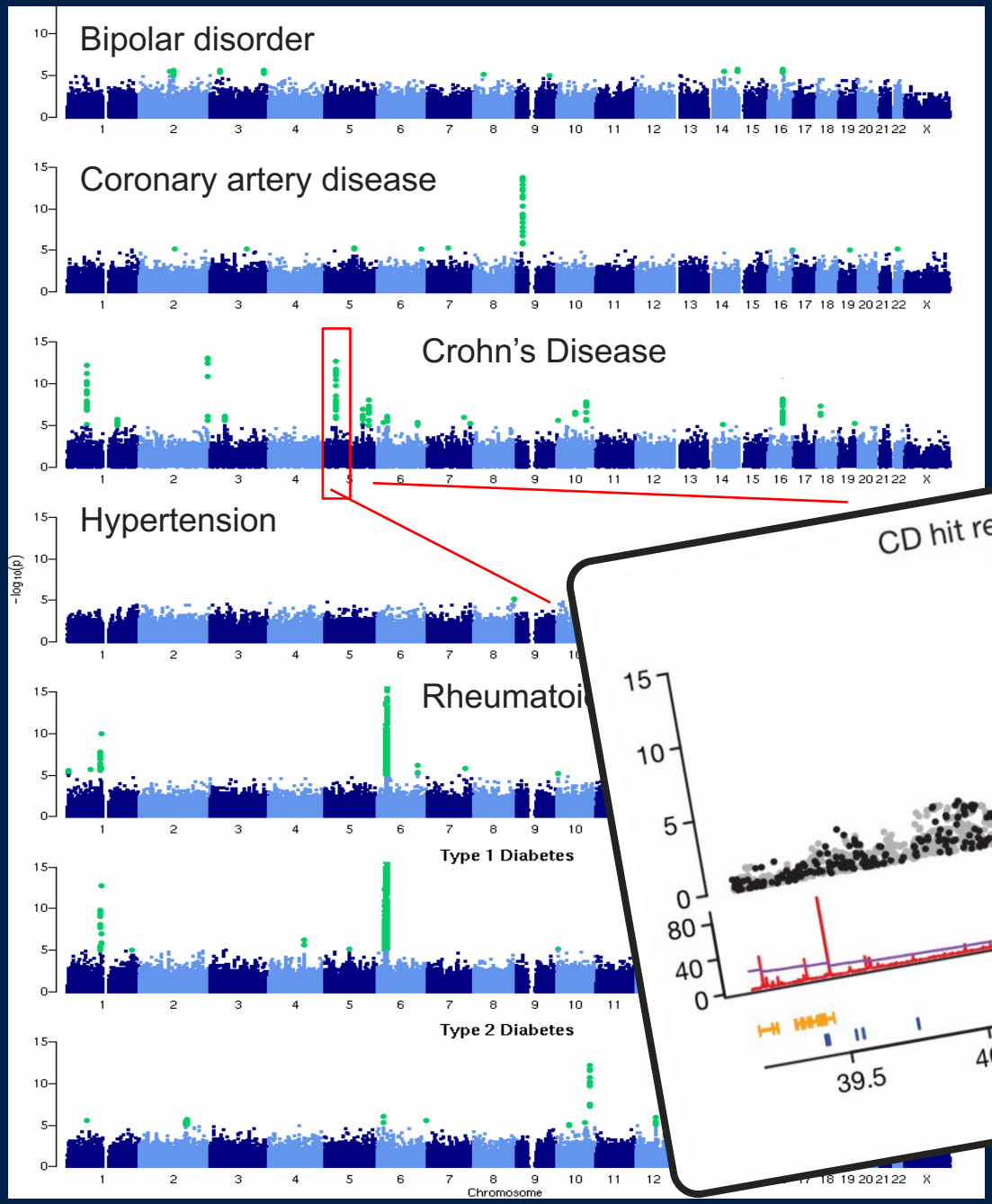
However, with some exceptions there has been less progress in turning GWAS associations into concrete information about biological processes, that can inform new therapies.

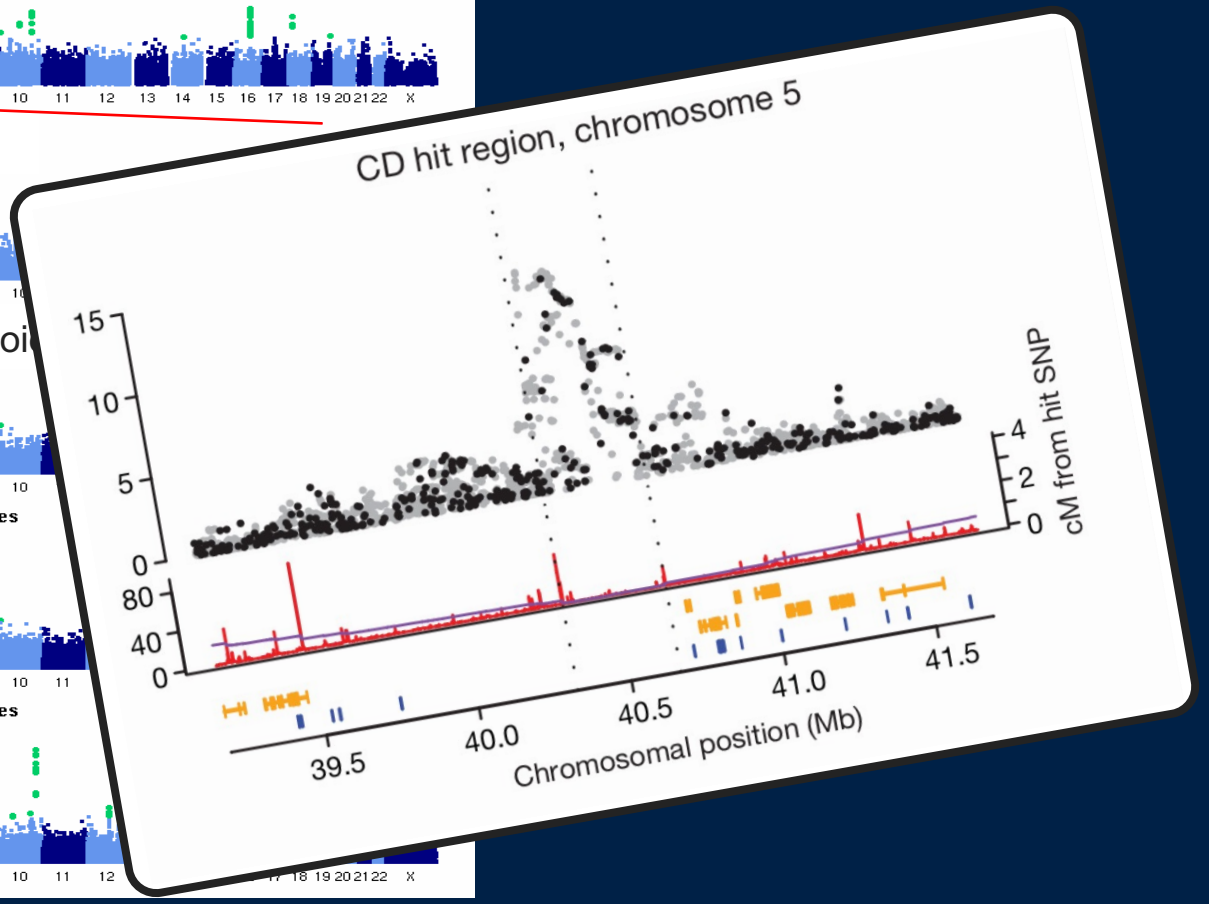
"**Fine-mapping**" = the process of narrowing down an association to a single causal mutation linked to biological mechanism.



E.g. this SNP associated with Crohn's disease :

- Is common (about 63% allele frequency in European populations)
- Has a modest effect size ($RR \approx 1.2$, i.e. about a 20% increase in risk)
- Is strongly associated (this association is now well replicated).





Not clear how this works biologically. E.g. there's no gene under the association signal!

Fine mapping is hard!


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 (Crohn's disease and Ulcerative Colitis), using genotype data on 67,852 individuals, and data on the functional state in relevant cell types.

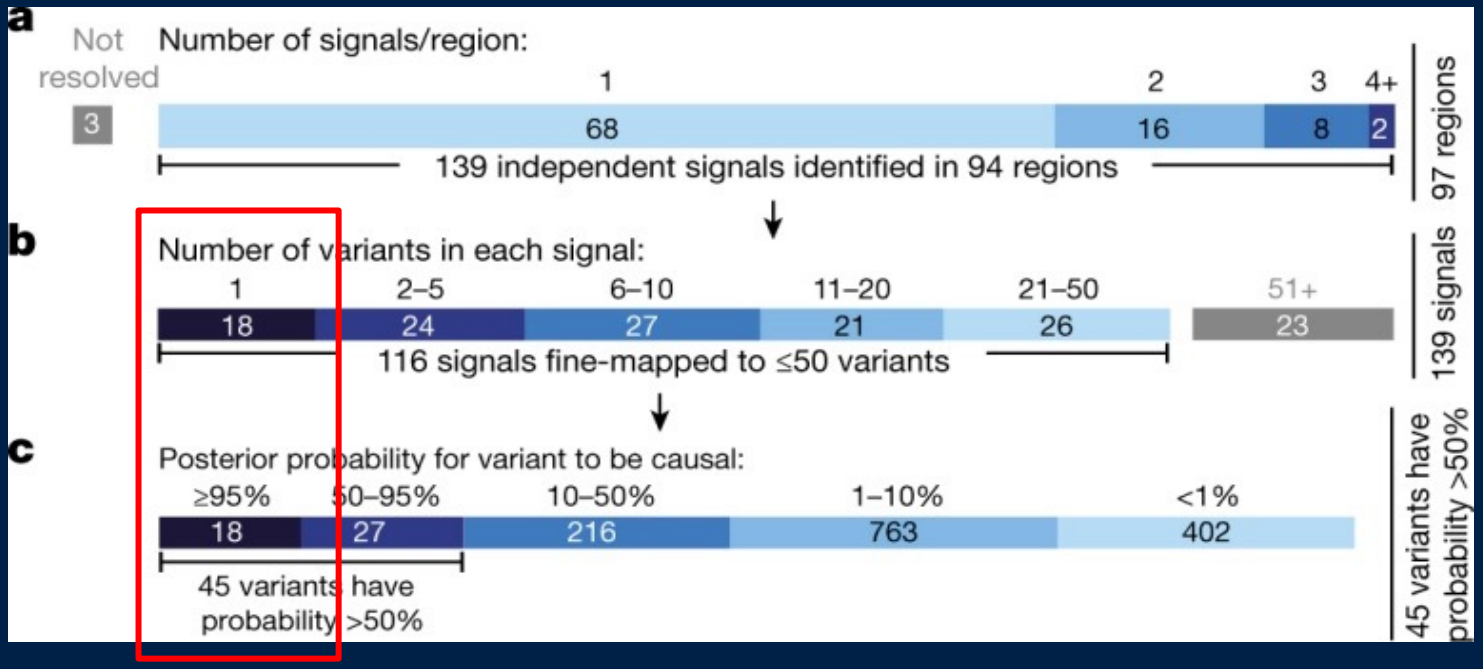
...with mixed success:

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.

Fine mapping is hard!

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

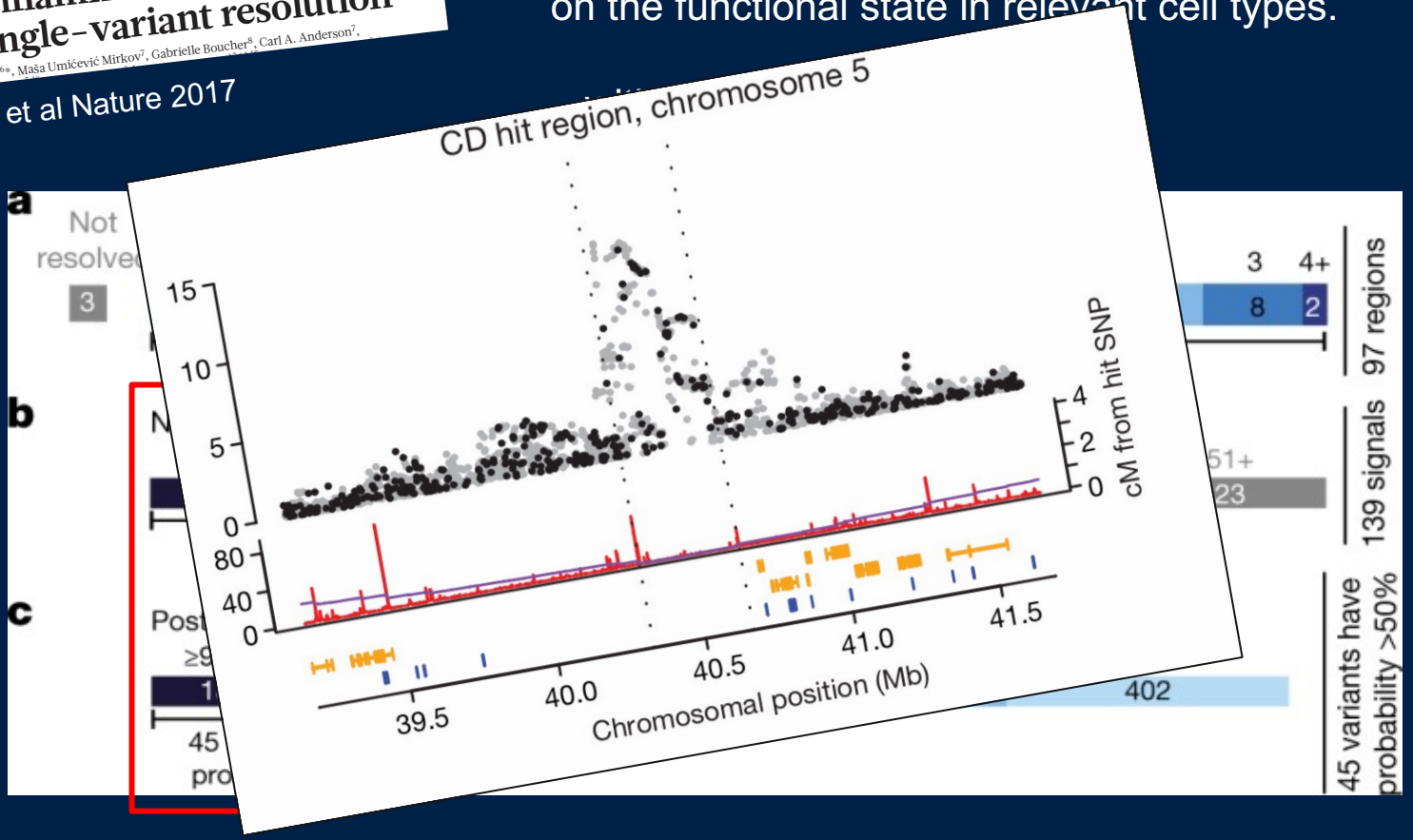
Attempted fine-mapping of 139 signals of association with inflammatory bowel disease (Crohn's disease and Ulcerative Colitis), using genotype data on 67,852 individuals, and data on the functional state in relevant cell types.

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.

Another example - IBD

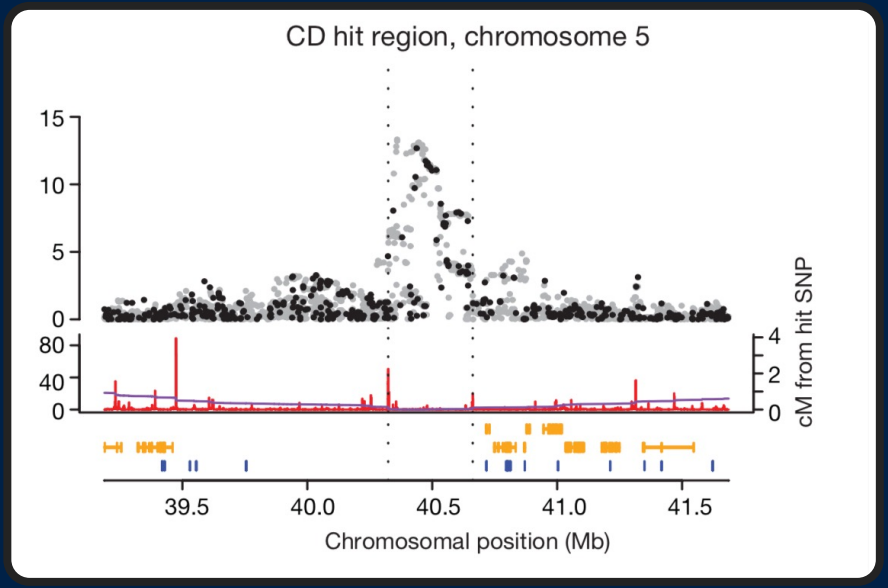

ARTICLE

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],

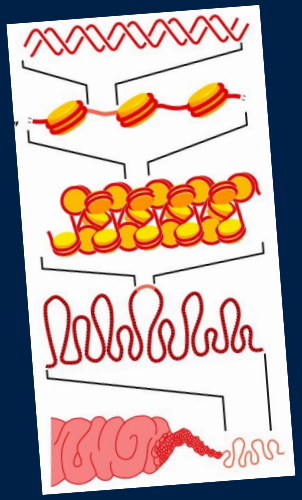
Huang et al Nature 2017



"This analysis [..] leaves 21 non-coding variants, all of which have >50% probabilities of being causal [..] that are not located within known motifs, annotated elements, or in any experimentally determined ChIP-seq peaks or eQTL credible sets[..]. While we have identified a statistically compelling set of genuine associations (often intronic or within 10 kb of strong candidate genes), we can make little inference about function.[…]. That most of the best-refined non-coding associations have no available annotation is perhaps sobering with respect to how well we may currently be able to interpret non-coding variation in medical sequencing efforts. […]

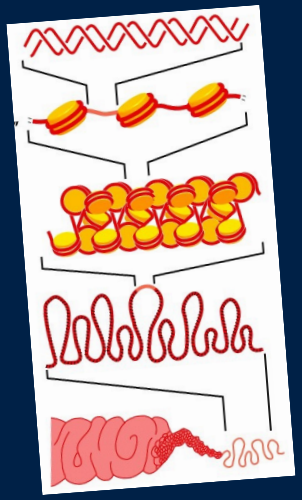
The circle of genetic causation

DNA gets physically packaged up into chromosomes...

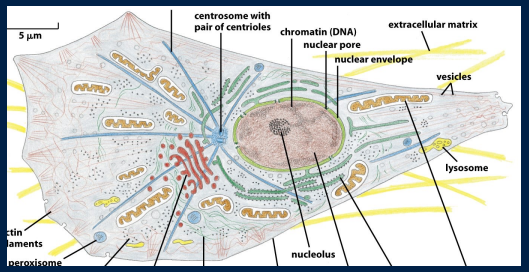


The circle of genetic causation

DNA gets physically packaged up into chromosomes...

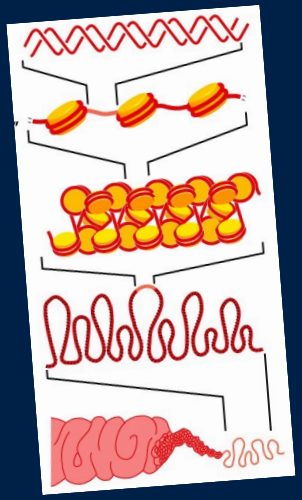


...inside cells, where it is **transcribed** to form proteins and other molecules...

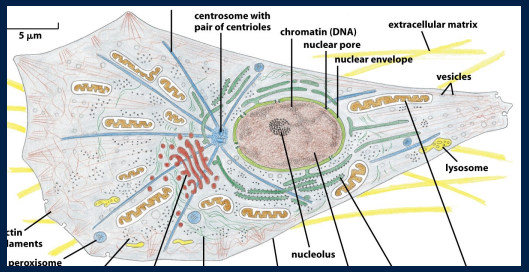


The circle of genetic causation

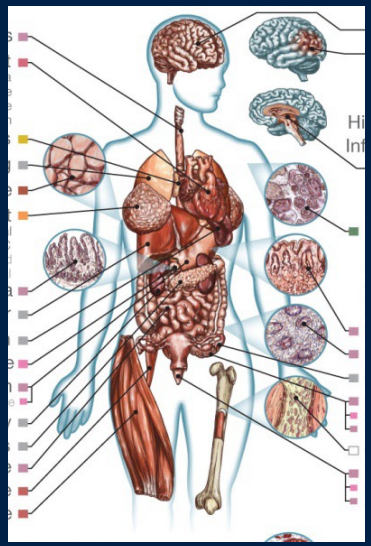
DNA gets physically packaged up into chromosomes...



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



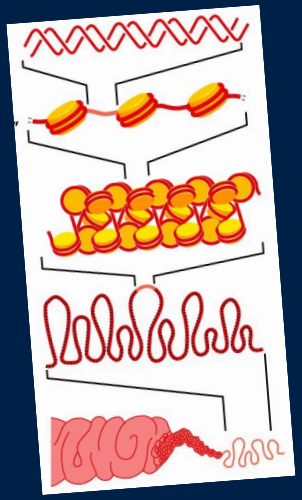
...that affect how the cells behave, forming different organs...



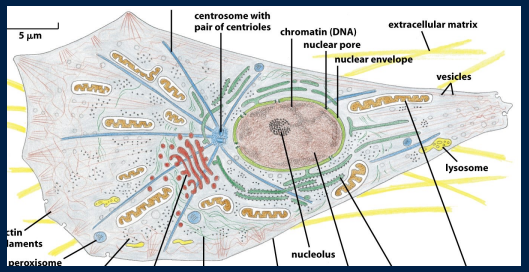
...that combine to make individuals...

The circle of genetic causation

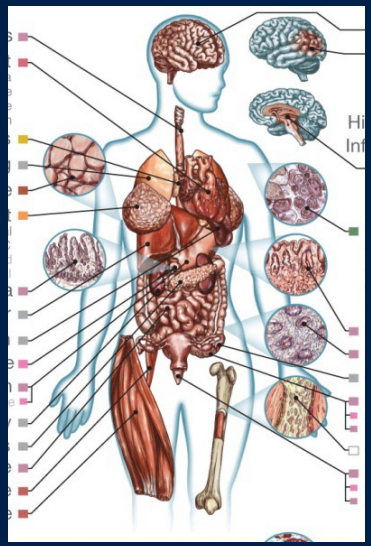
DNA gets physically packaged up into chromosomes...



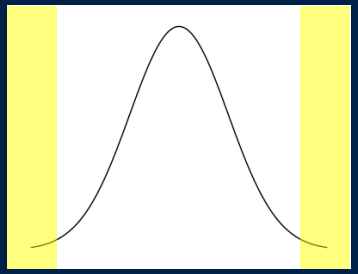
...inside cells, where it is **transcribed** to form proteins and other molecules...



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



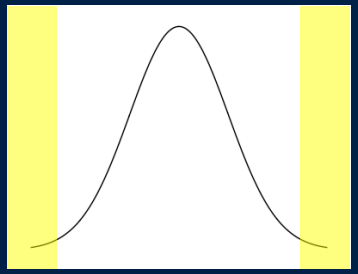
...that combine to make individuals...



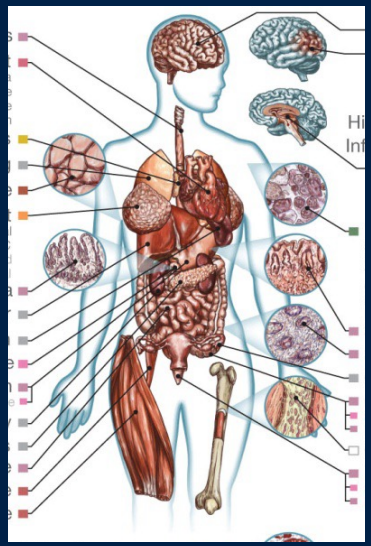
...whose success is affected by the traits they have...

The circle of genetic causation

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



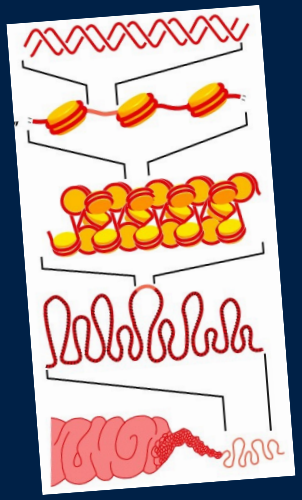
...whose success is affected by the traits they have...



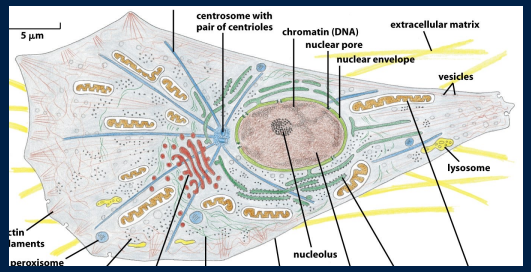
...that combine to make individuals...

There is complex biology at all stages

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



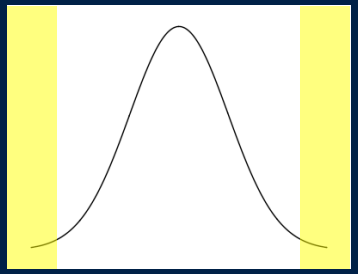
...inside cells, where it is **transcribed** to form proteins and other molecules...



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

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

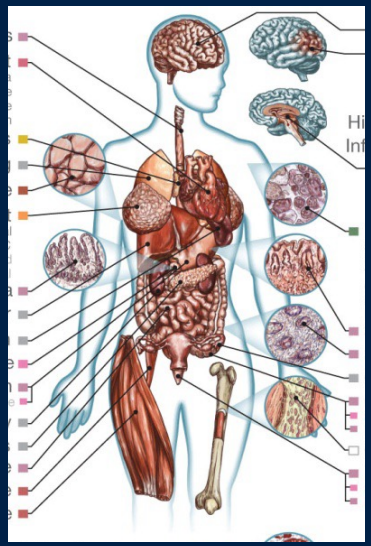
Clinical phenotype measurements

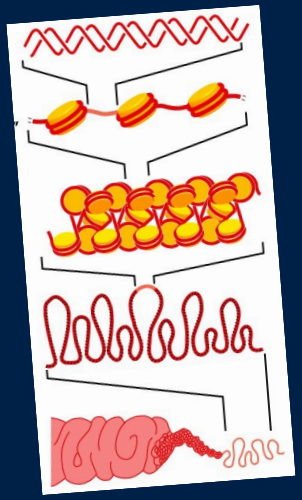
...that gets physically packaged up into chromosomes...

There is complex biology at all stages

And we can measure it.

Biomarker measurements



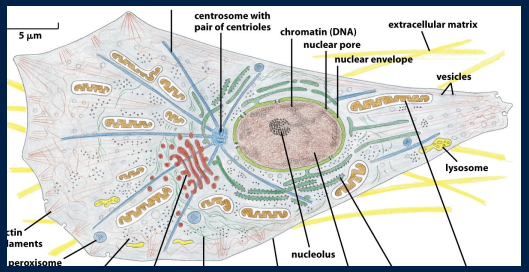


Chromatin state marker assays, ChIP-seq, ...

RNA-seq, spectroscopy, antibody binding

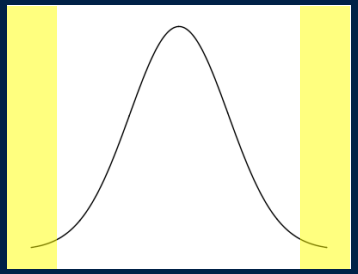
...that combine to make individuals...

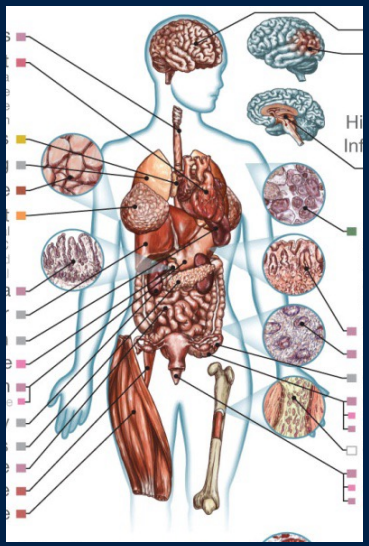
...inside cells, where it is **transcribed** to form proteins and other molecules...



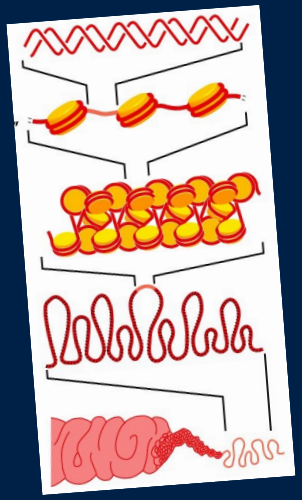
...that affect how the cells behave, forming different organs...

The circle of genetic causation

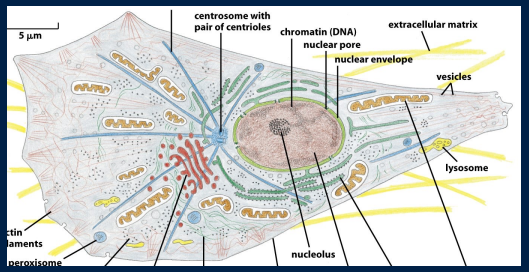


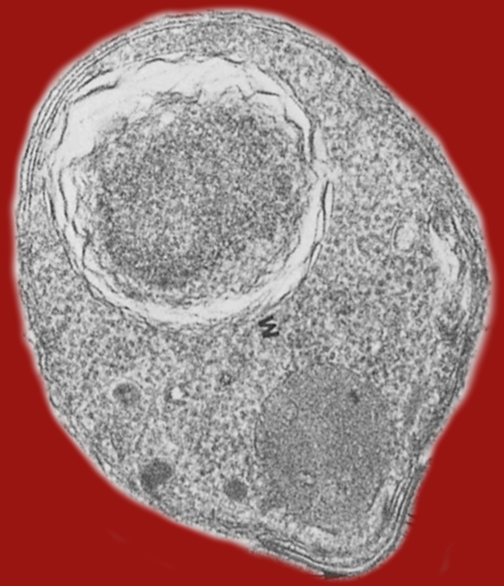


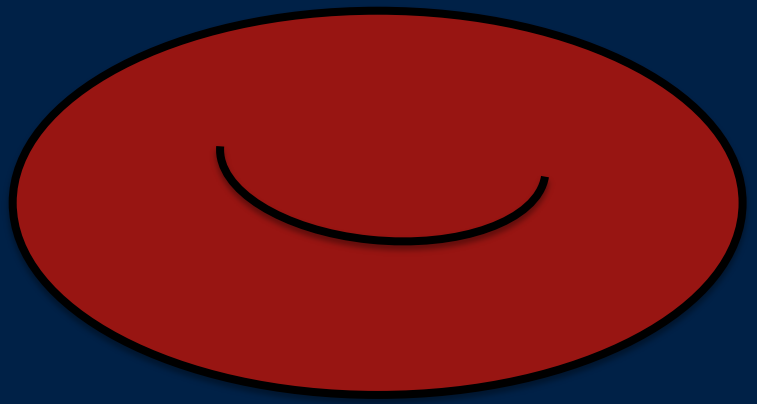
...that combine to make individuals...



Fine-mapping example 1:
genetic complexity





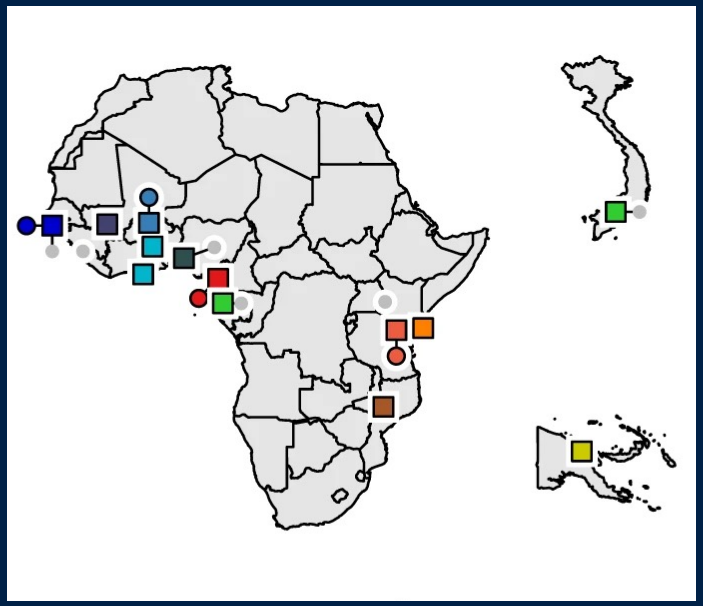


Plasmodium falciparum **VS** humans

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



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

GWAS in 17,000 severe malaria cases and population controls
From 12 sites in Africa, Oceania, and SE Asia.
Genotyped on the Illumina Omni 2.5M array

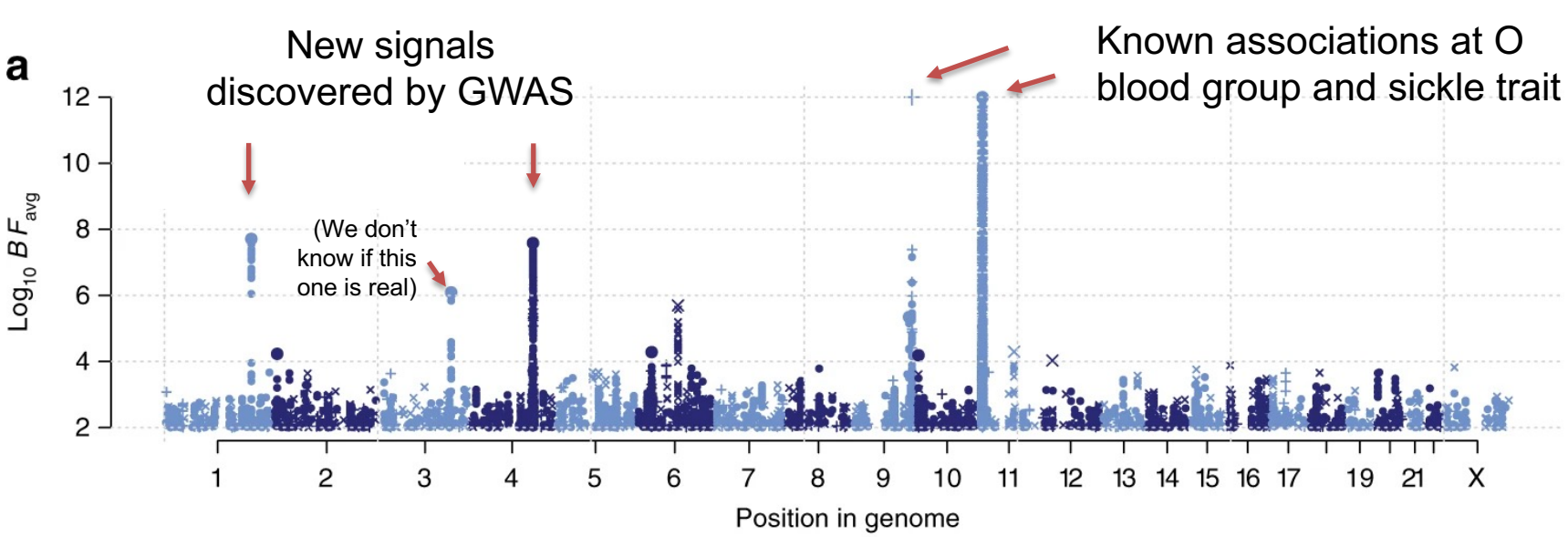
+ whole-genome sequences for imputation

Malaria Genomic Epidemiology Network. "*Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania*".
Nature Communications (2019). https://doi.org/10.1038/s41467-019-13480-z

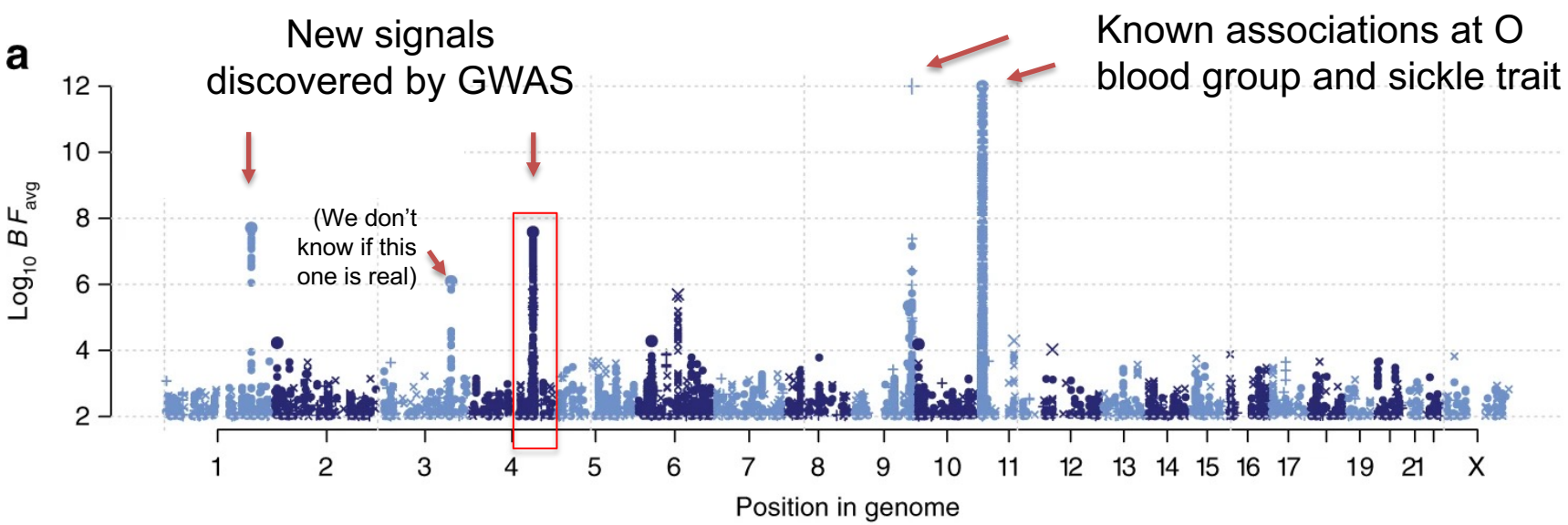
MalariaGEN
GENOMIC EPIDEMIOLOGY NETWORK

www.malariagen.net

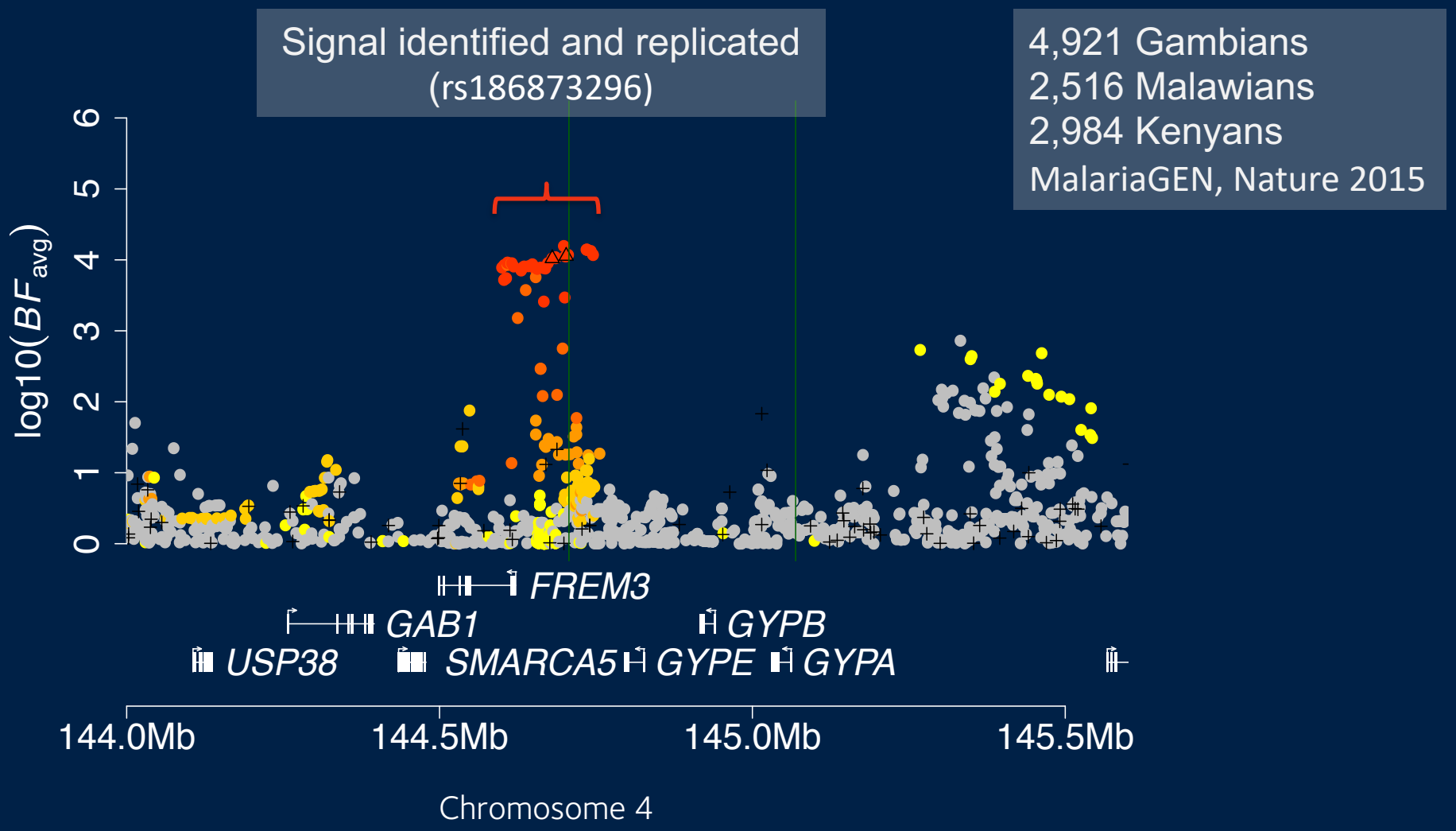
Natural resistance is driven by red blood cell variation



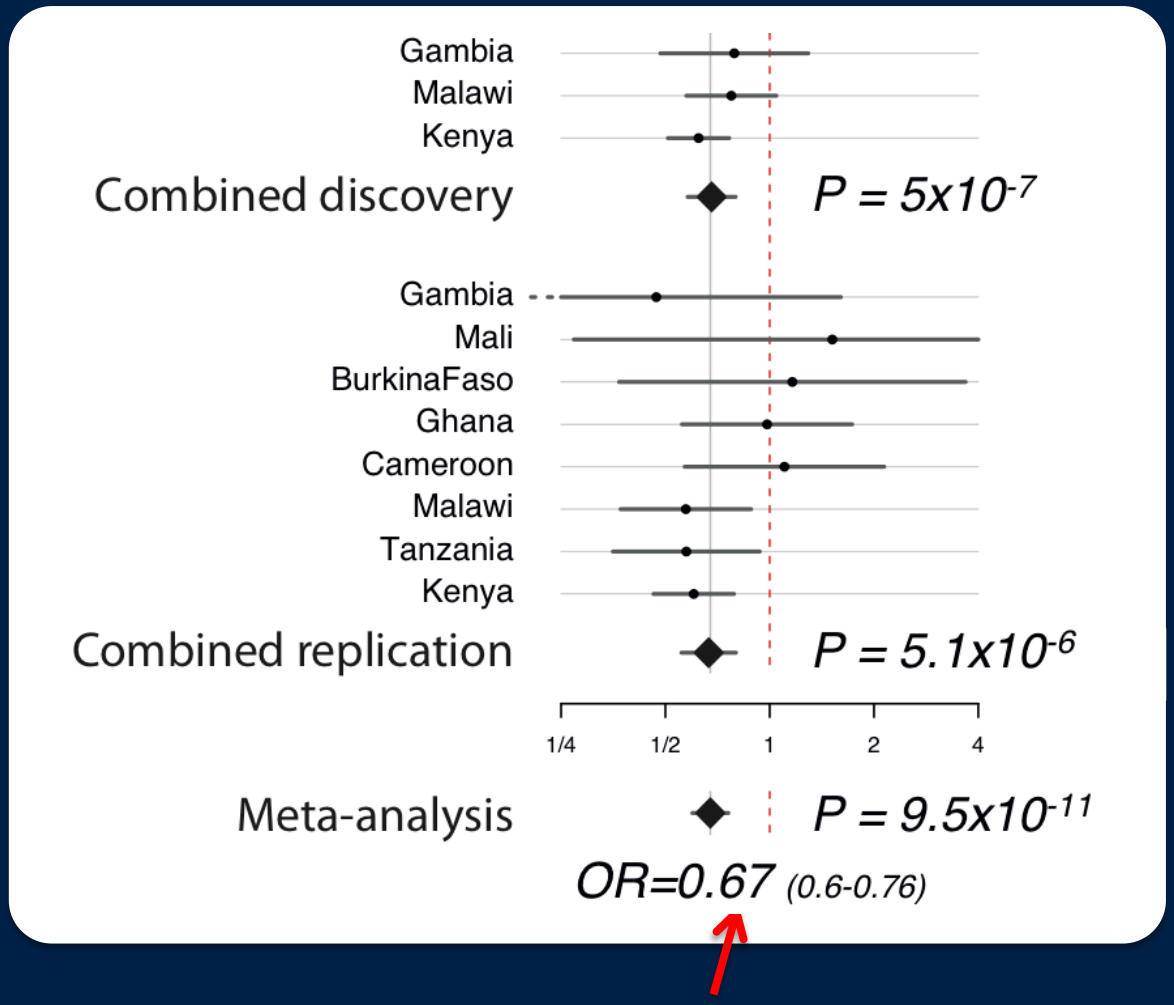
Natural resistance is driven by red blood cell variation



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

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

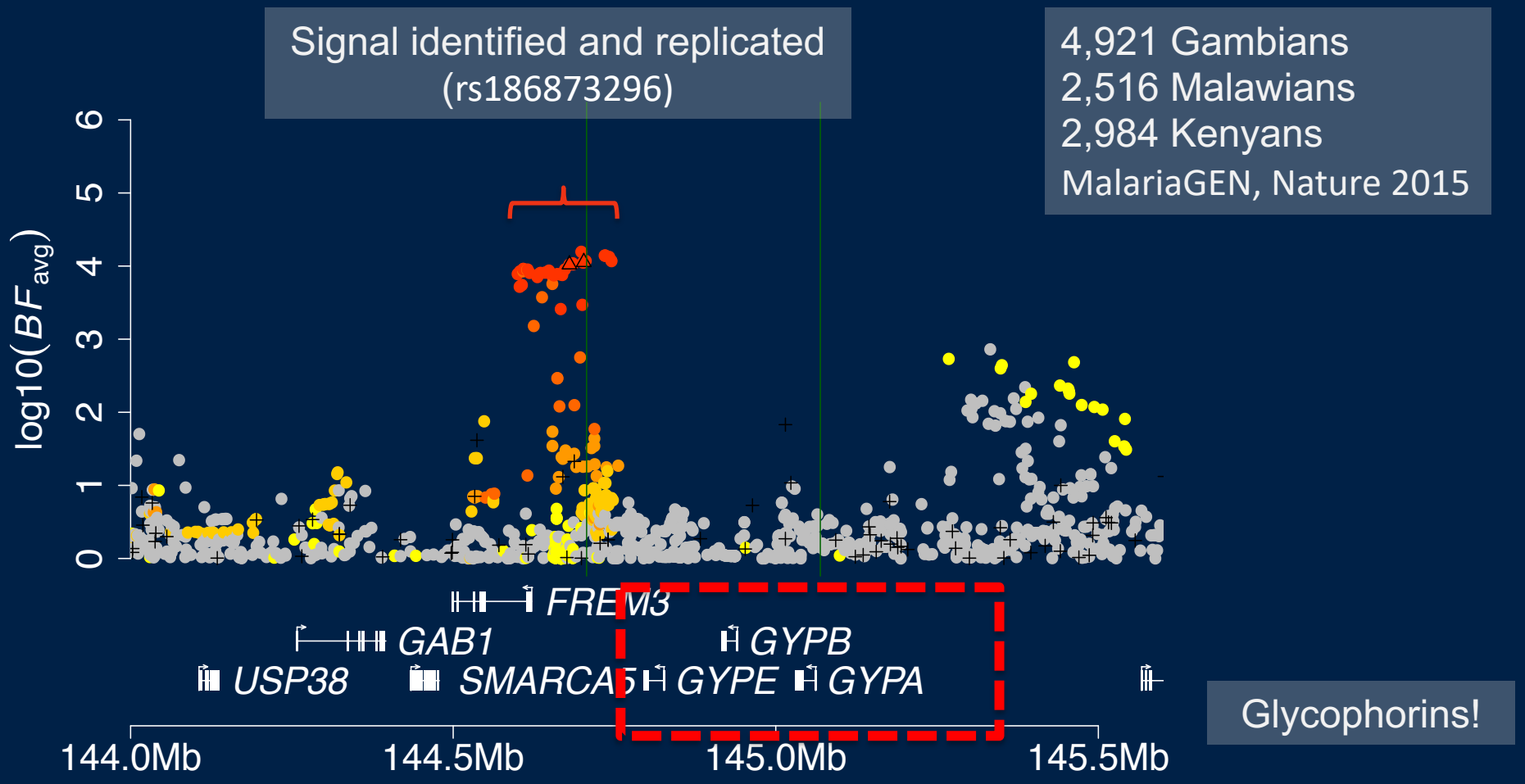
Can we finemap?

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



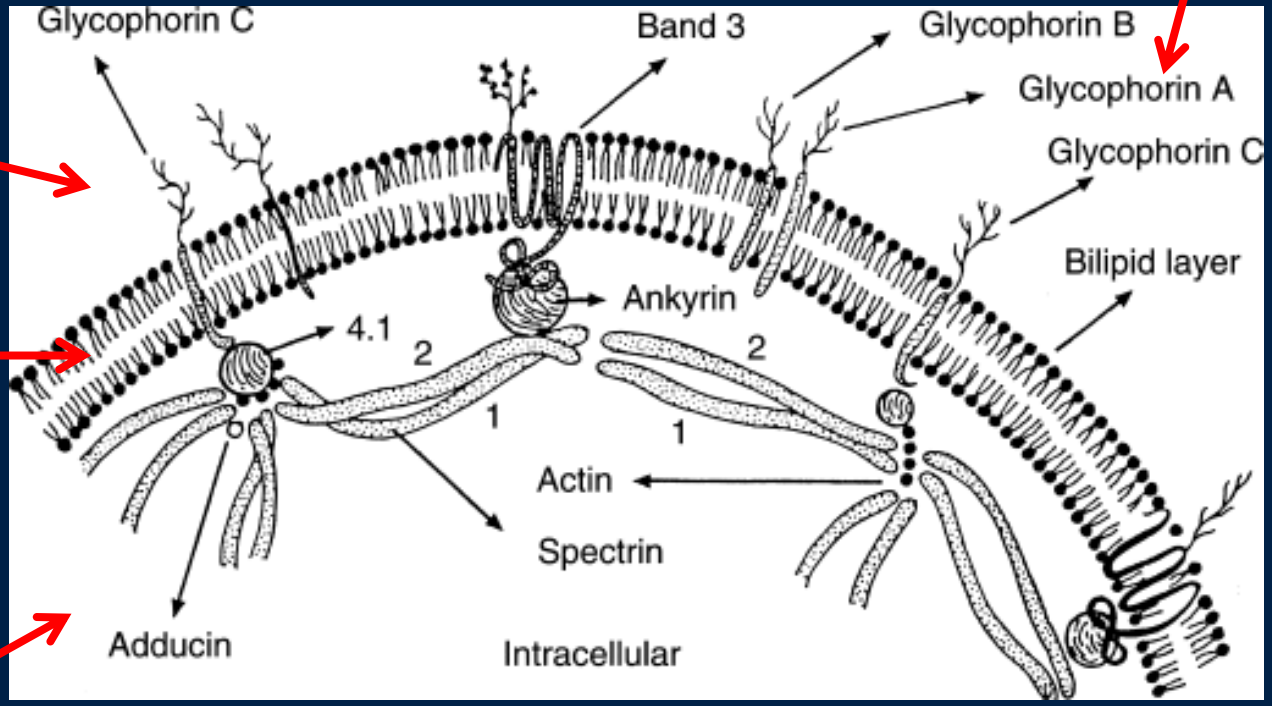
Glycophorins encode the ‘MNS’ blood group
(antigenic molecules on RBC surface)

Glycophorins

Outside red blood cell

Red blood cell membrane

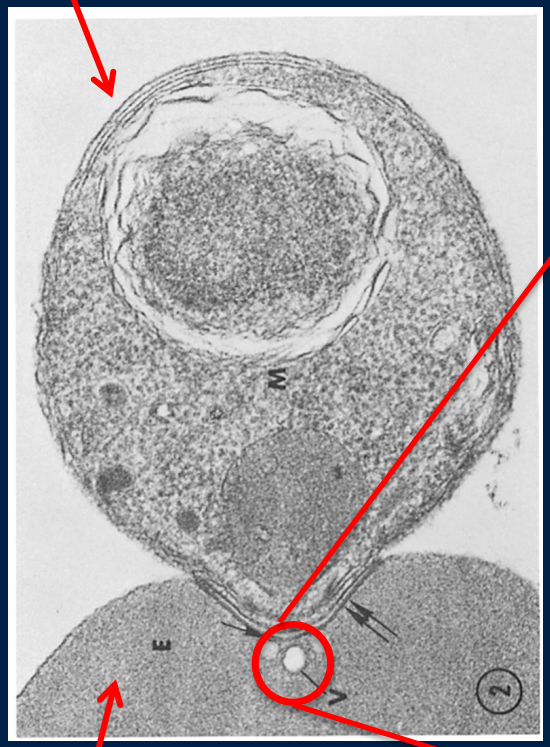
Inside red blood cell



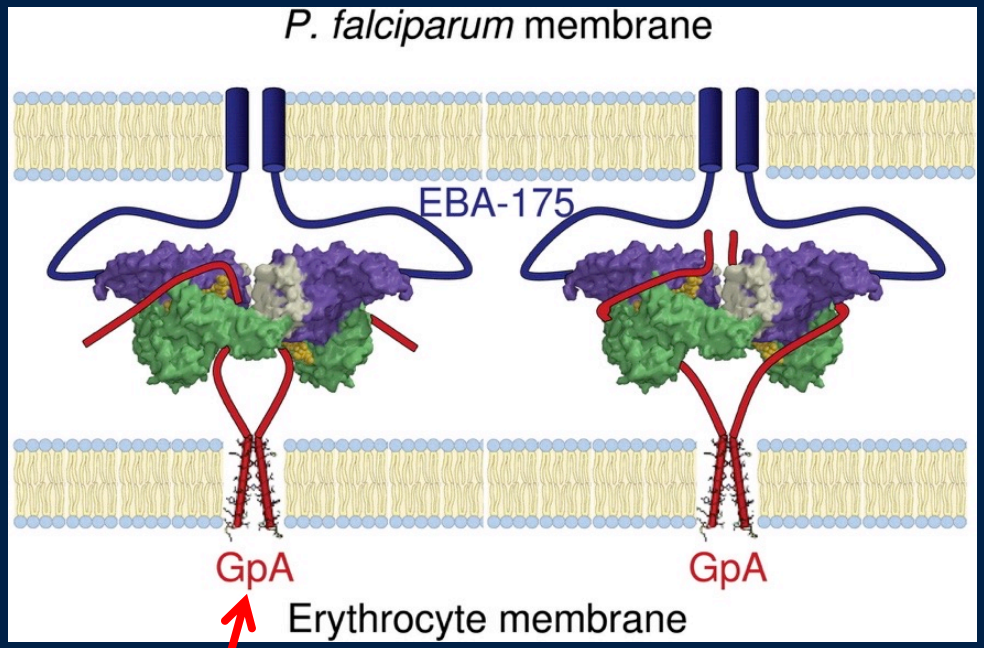
Grimes and Slater, The Inherited Metabolic Diseases, 1994

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)?

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)?

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?

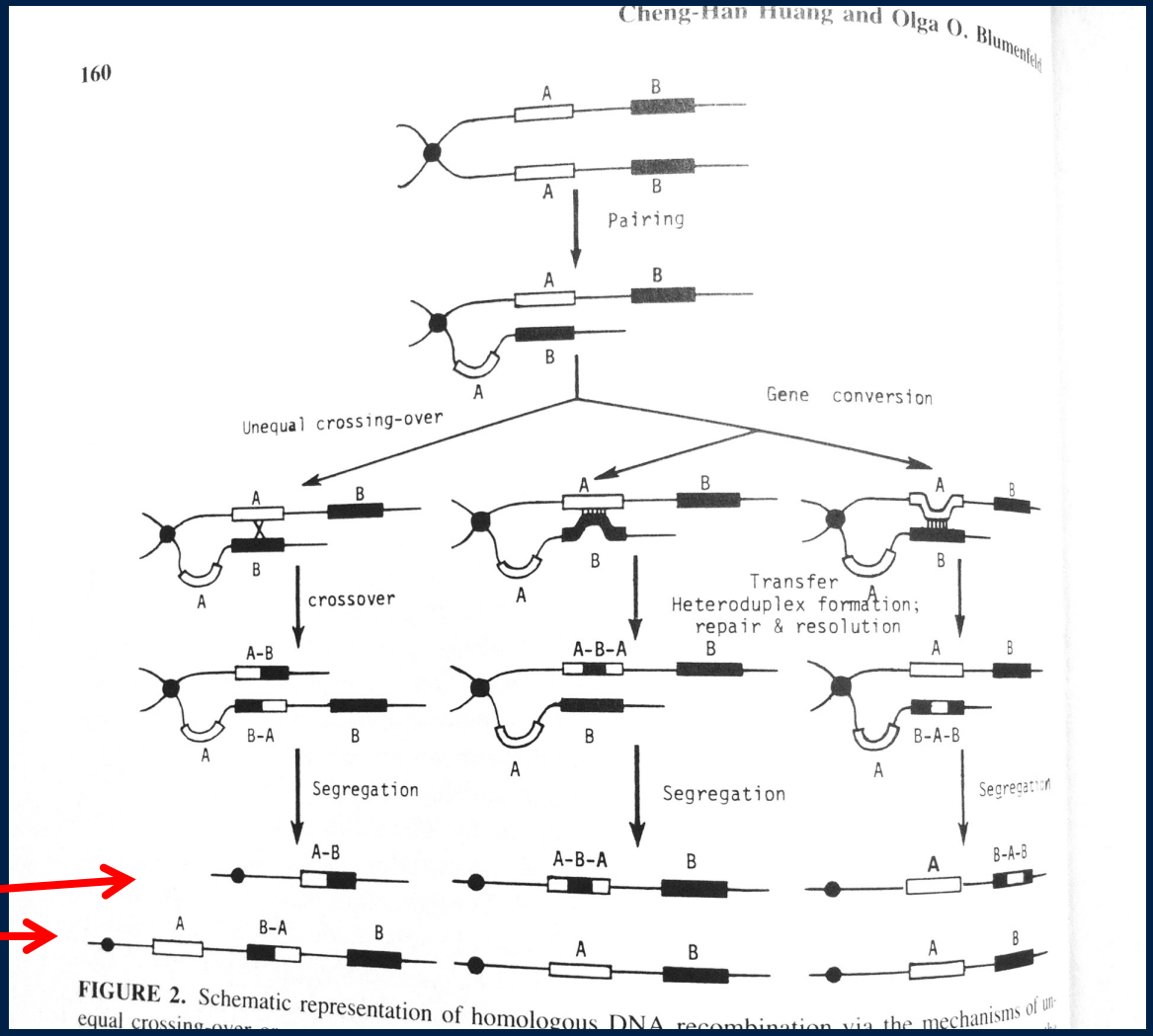
- Good candidates for the causal mutation(s)?

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

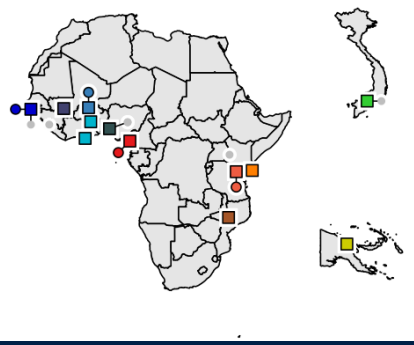
A regional reference panel capturing structural variation

We used the >3,600 samples including
- 1000 Genomes Project Phase III reference panel
- plus our newly-sequenced samples

a

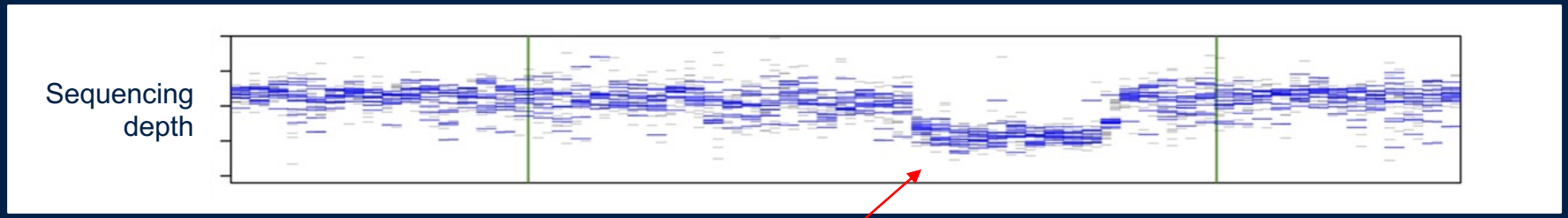
Whole–genome sequences

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**



…to call SNPs and indels and structural variation.

Illustration of structural variant calling:



(this sample has a deletion in this region)

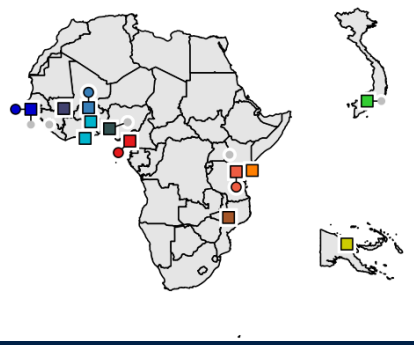
A regional reference panel capturing structural variation

We used the >3,600 samples including
- 1000 Genomes Project Phase III reference panel
- plus our newly-sequenced samples

a

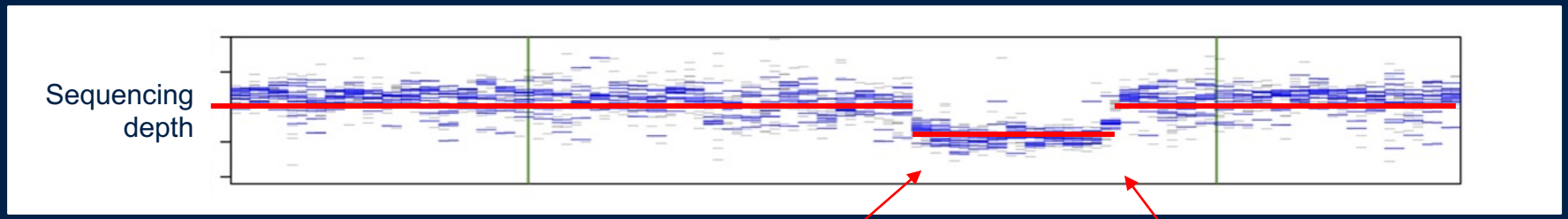
Whole–genome sequences

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**



…to call SNPs and indels and structural variation.

Illustration of structural variant calling:



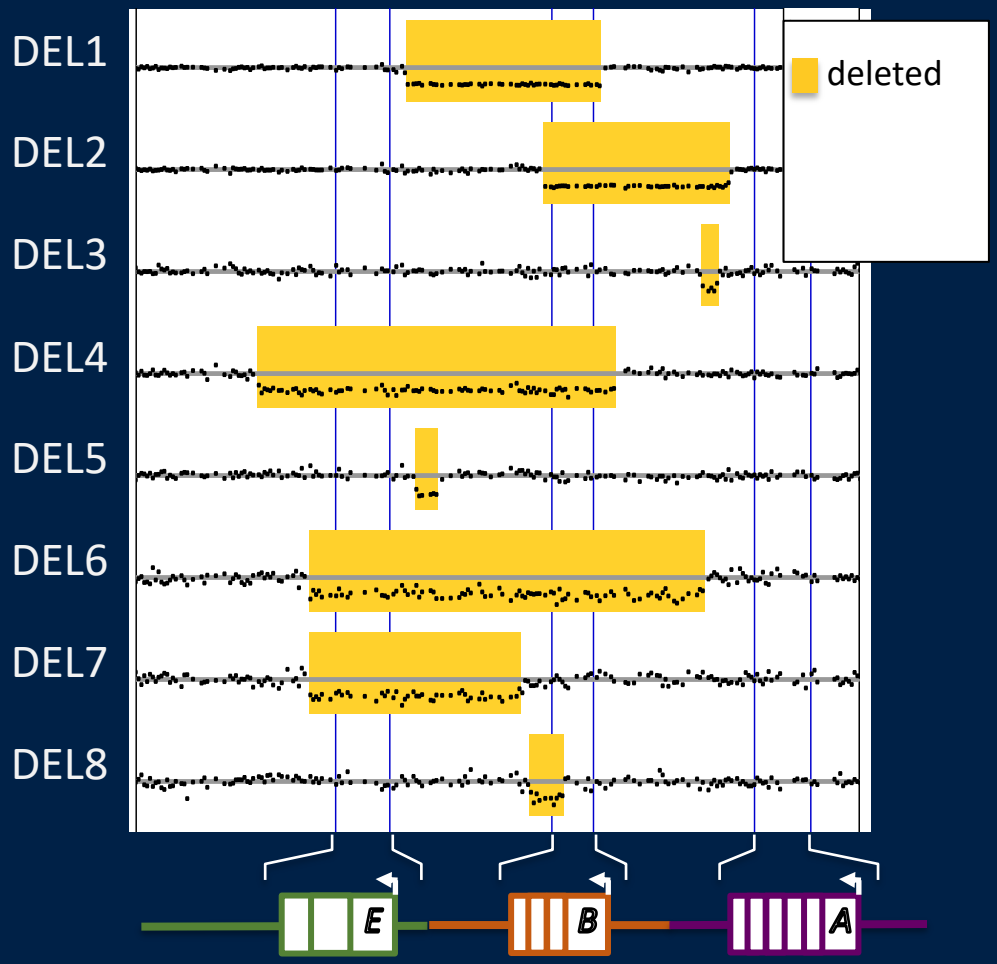
(this sample has a deletion in this region)

…our method infers the copy number

The region turned out to have *a lot* of structural variation

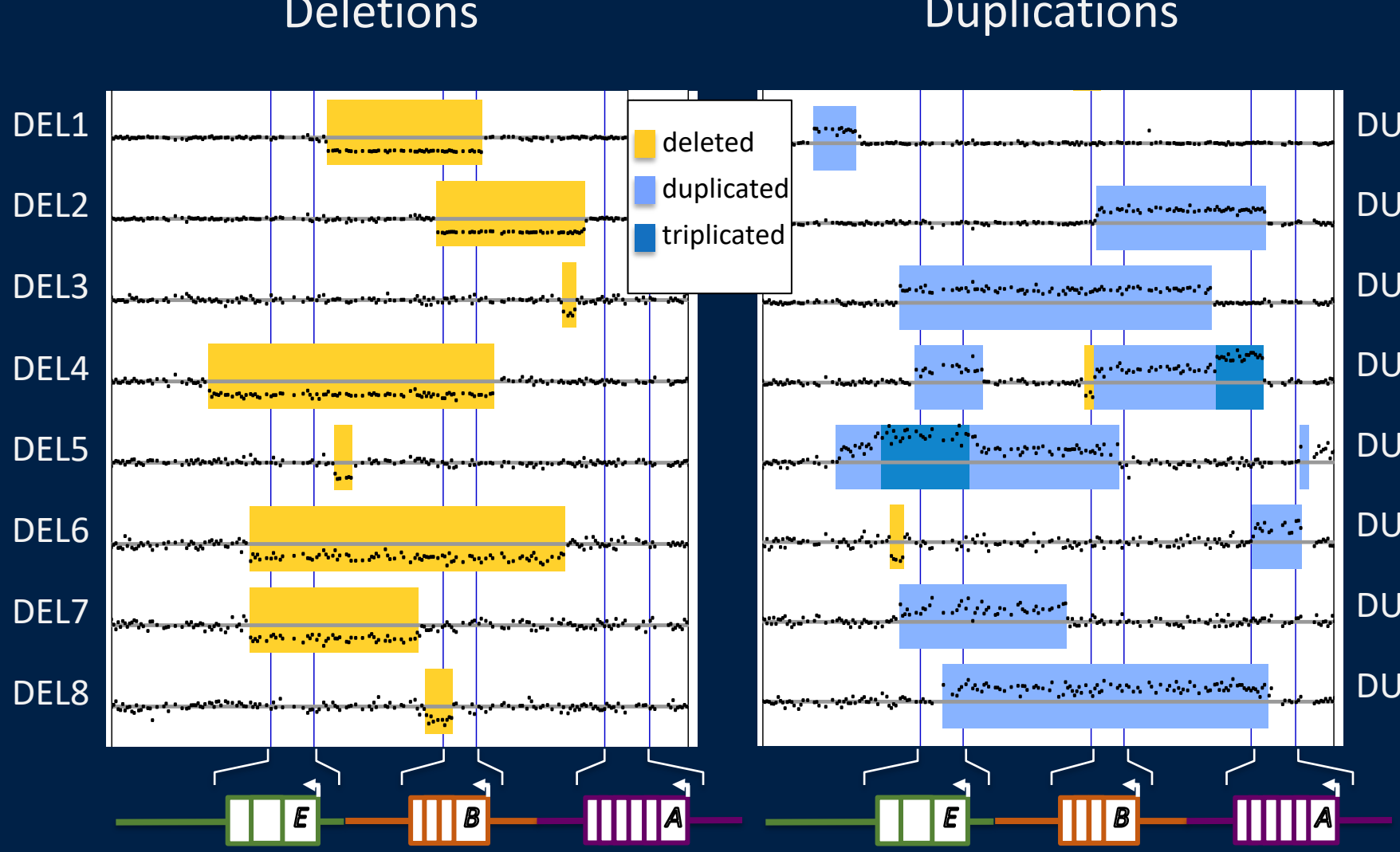
Deletions

Duplications



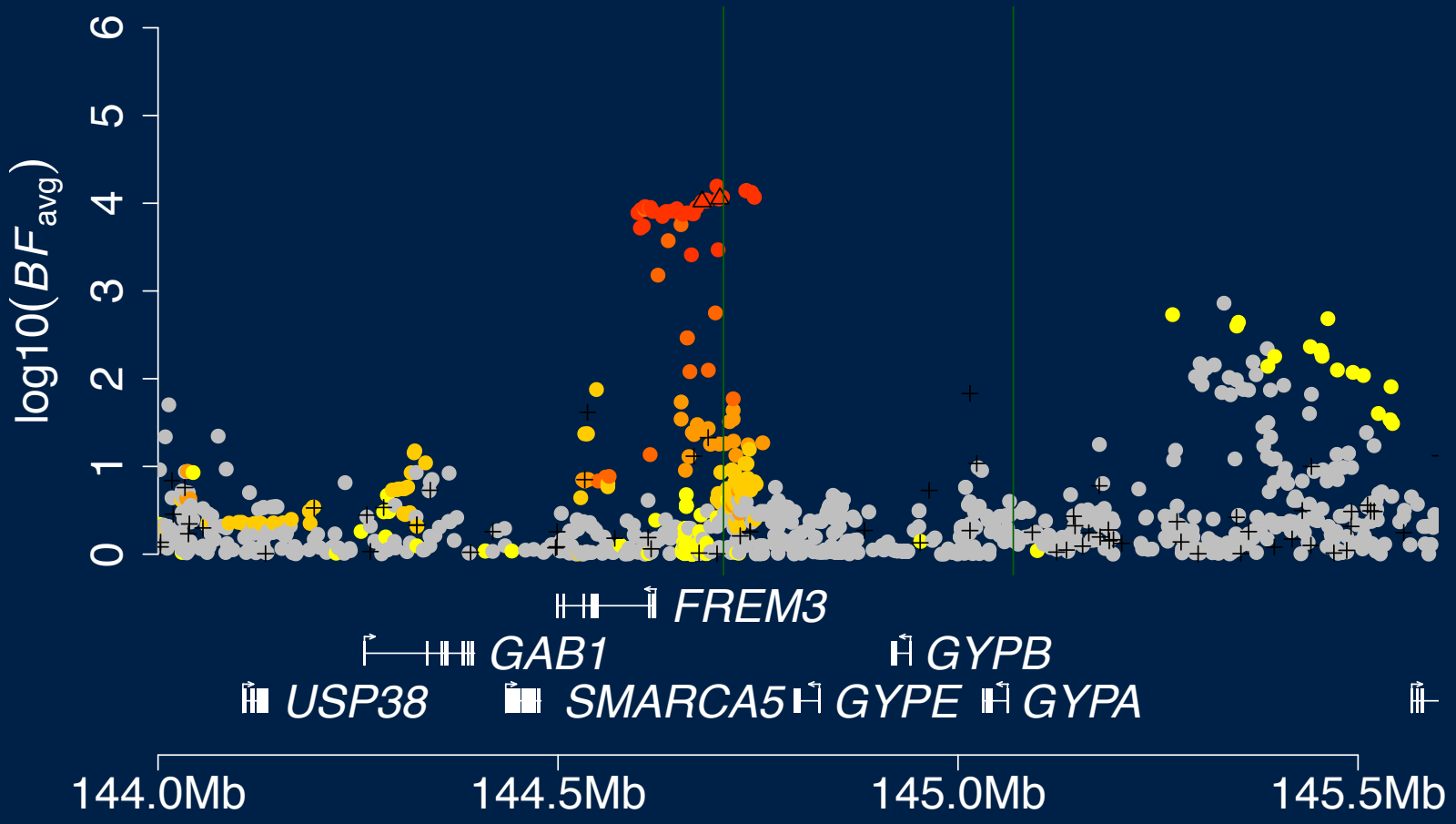
14% of Africans carry a CNV affecting these genes

The region turned out to have *a lot* of structural variation



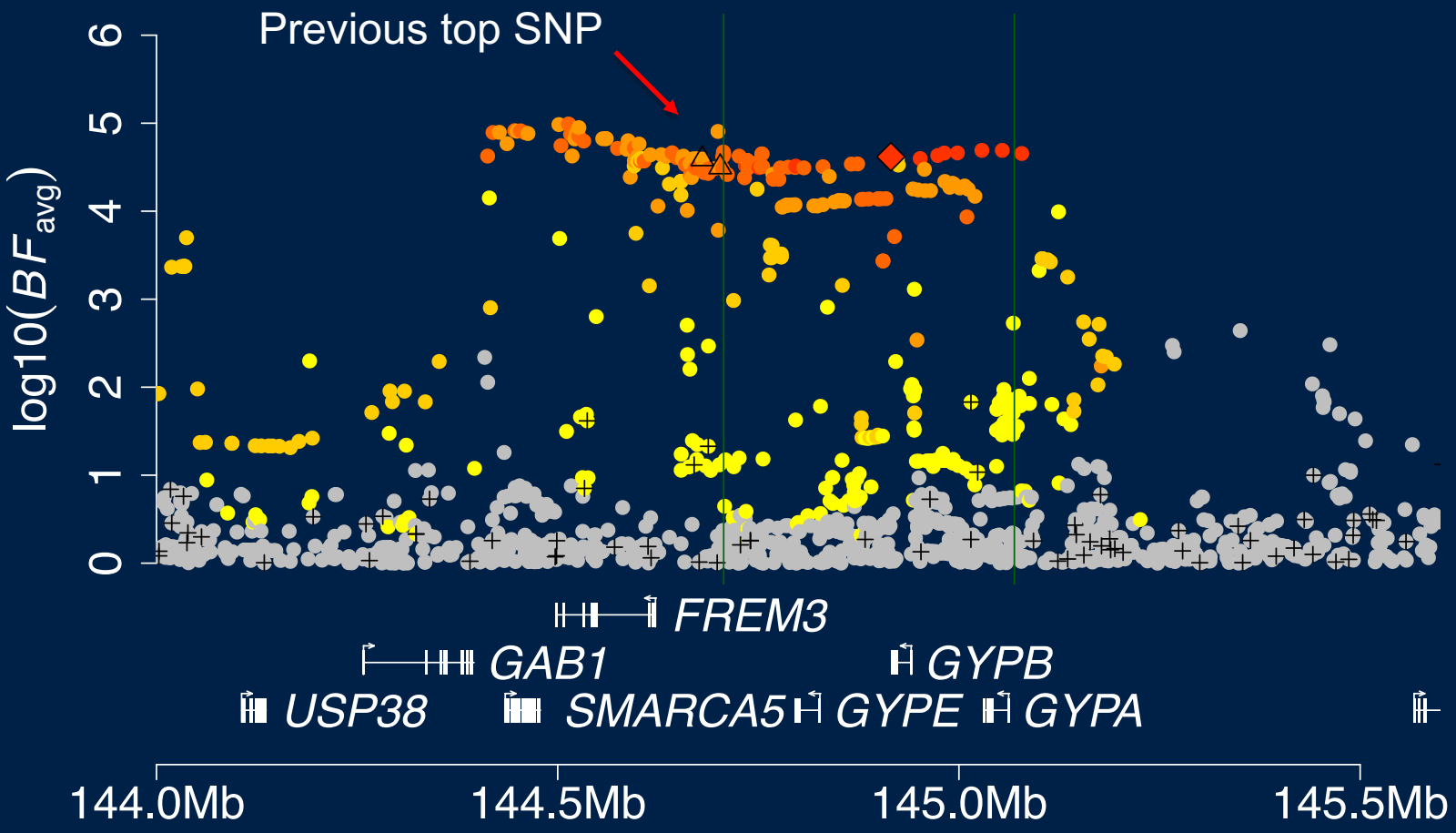
14% of Africans carry a CNV affecting these genes

Before fine-mapping



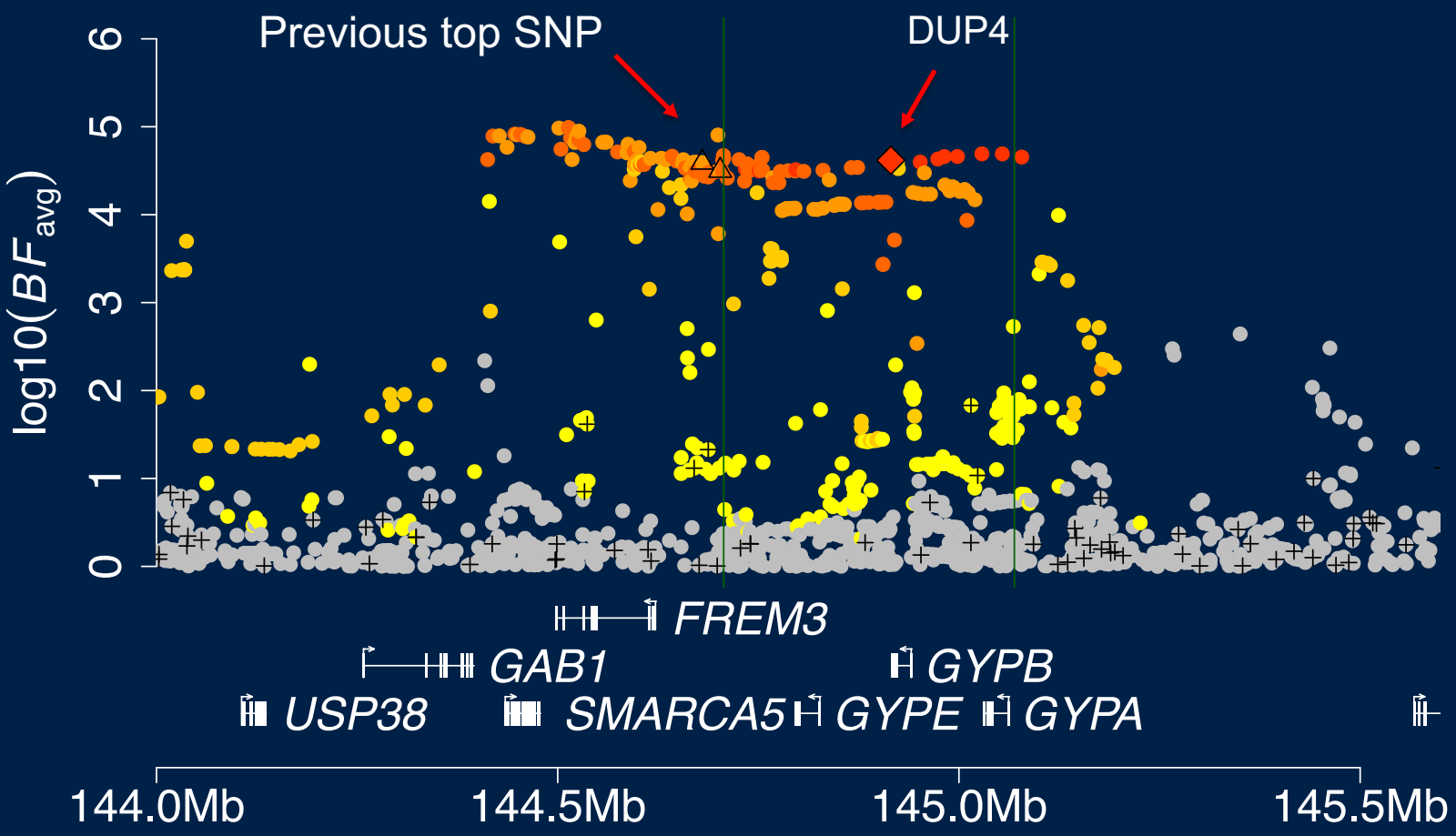
Original GWAS result

After fine-mapping

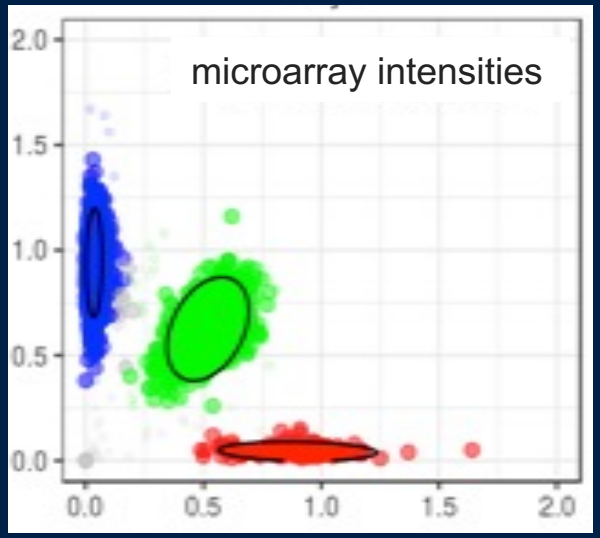


Result after incorporating genetic variation discovered in sequenced samples.

After fine-mapping



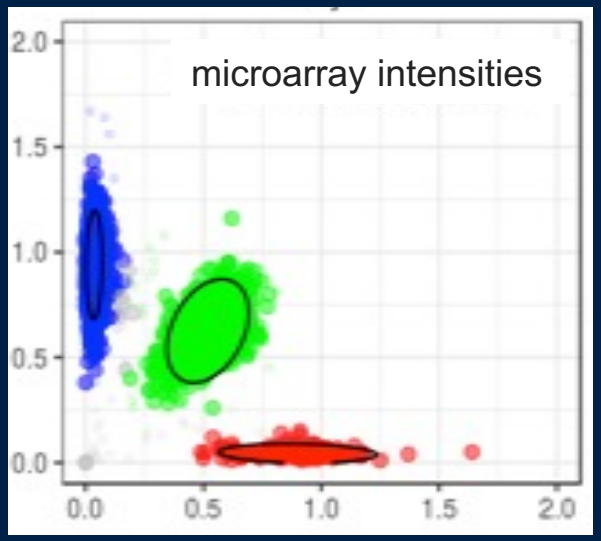
Confirming structural variants using cluster plots



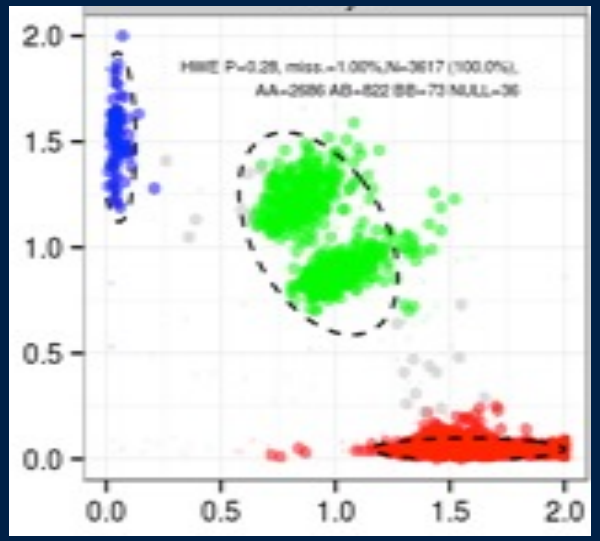
This is how a microarray cluster plot should look: 3 clusters for AA / AB / BB genotypes

Confirming structural variants using cluster plots

Actually this signal was evident in our cluster plots



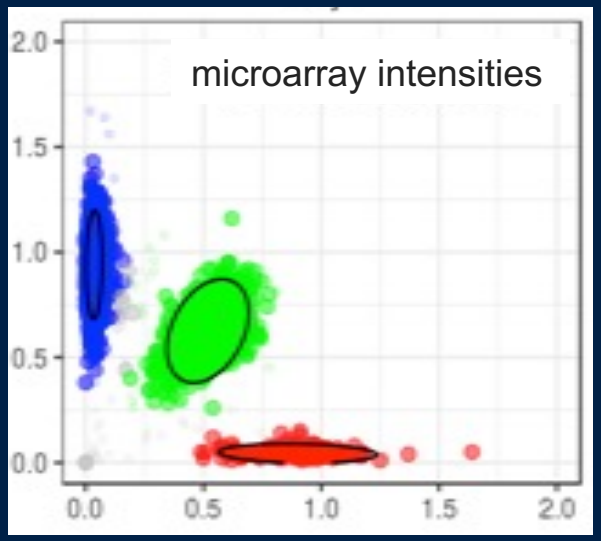
This is how a microarray cluster plot should look: 3 clusters for AA / AB / BB genotypes



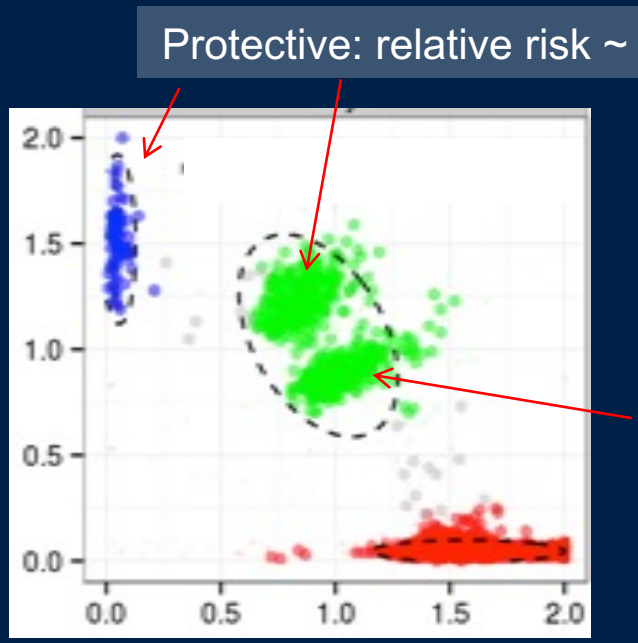
What we saw in this region

Confirming structural variants using cluster plots

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

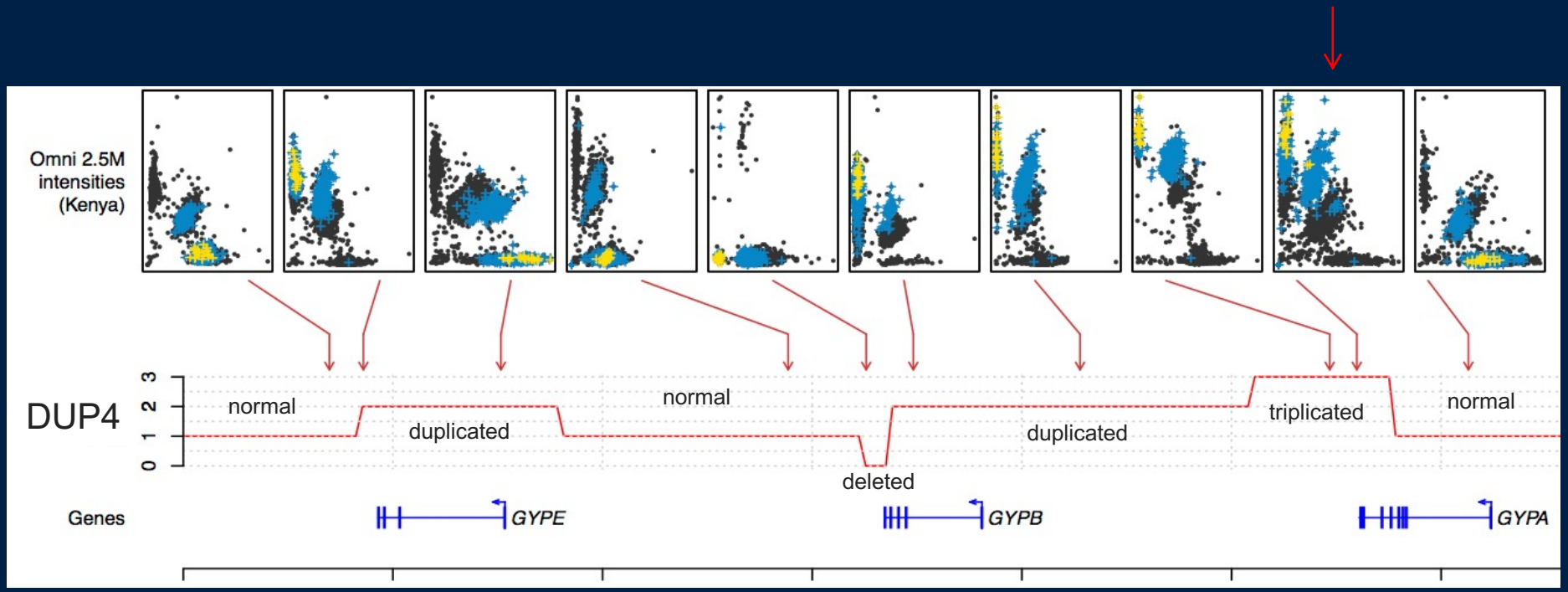


This is how a microarray cluster plot should look: 3 clusters for AA / AB / BB genotypes



What we saw in this region

Confirming structural variants using cluster plots

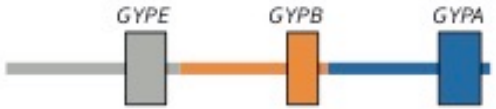


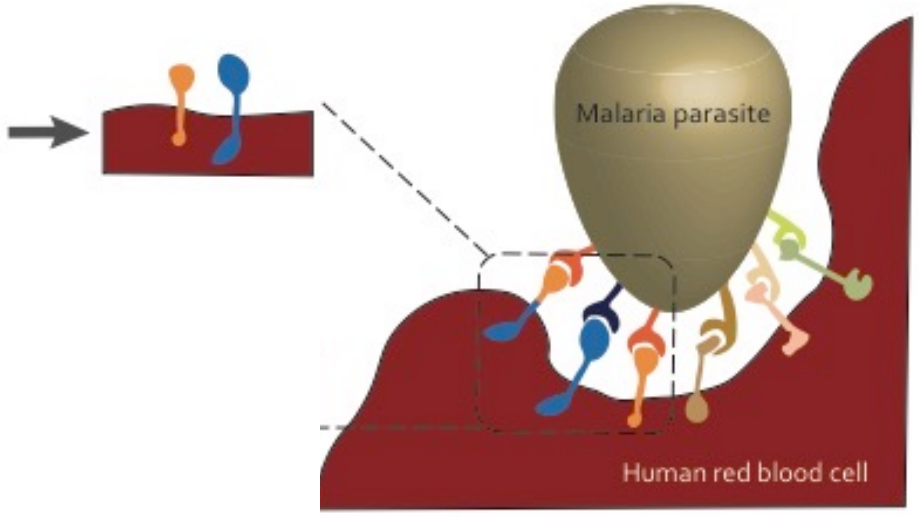
We were able to use cluster plots to confirm individuals in our GWAS really do carry the complicated structural variant “DUP4”.

DUP4 is pretty complicated – what could it be?

What is DUP4?

"Normal" haplotype:

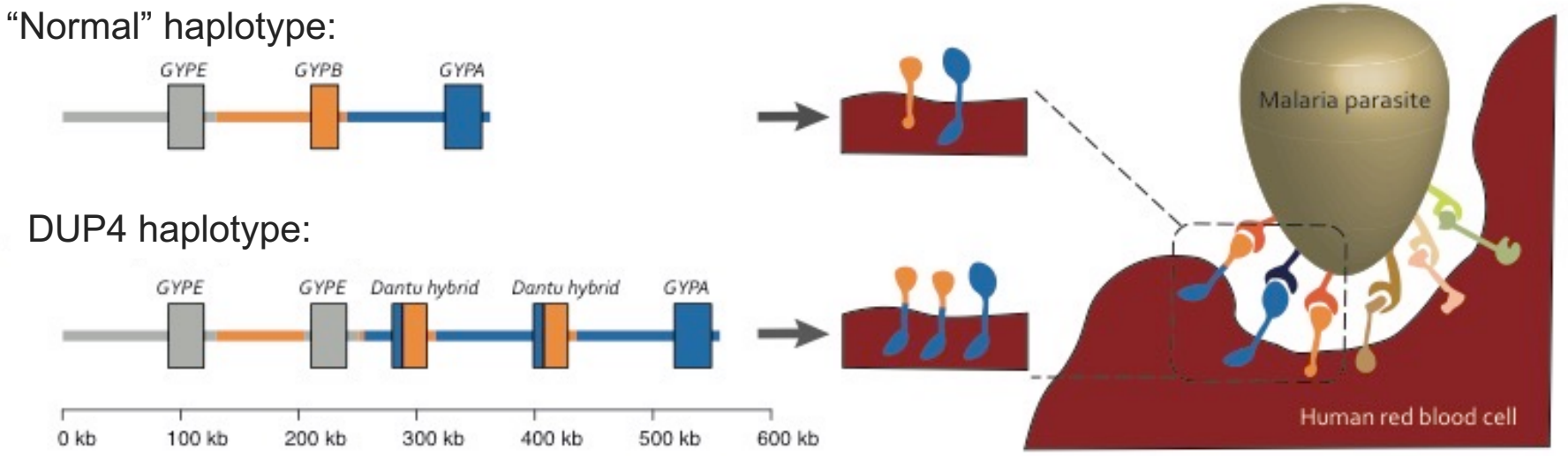




Leffler et al, "*Resistance to malaria through structural variation of red blood cell invasion receptors*", Science (2017)

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

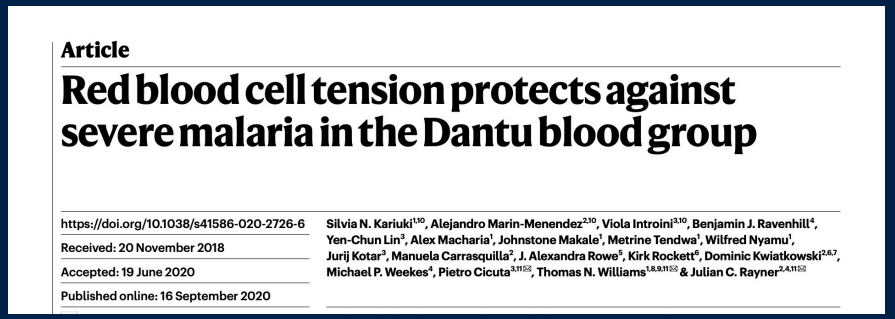
What is DUP4?



Leffler et al, "*Resistance to malaria through structural variation of red blood cell invasion receptors*", Science (2017)

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:

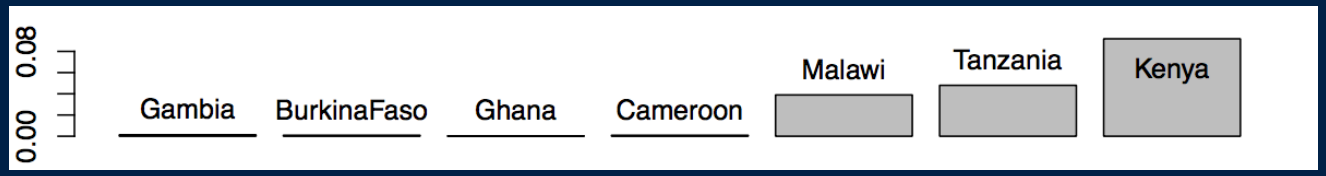
1 in 44,112	Londoners*
0 in 1,000	Germans†
1 in 320	African Americans†
0 in 2870	Gambians‡

...but found at high frequency in east Africa

The Dantu blood group has been found in:

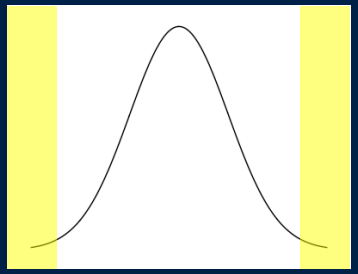
1 in 44,112	Londoners*
0 in 1,000	Germans†
1 in 320	African Americans†
0 in 2870	Gambians‡
1 in 12	Malawians‡
1 in 6	Kenyans (from the Kilifi region)‡

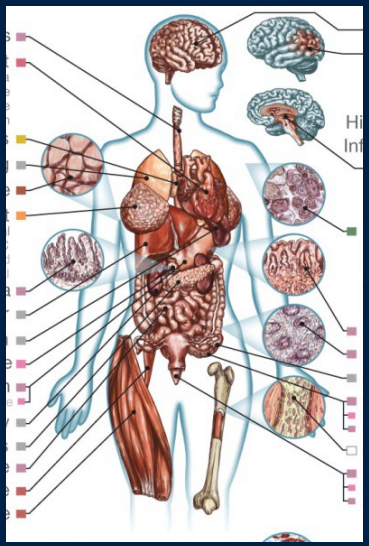
Allele frequency:



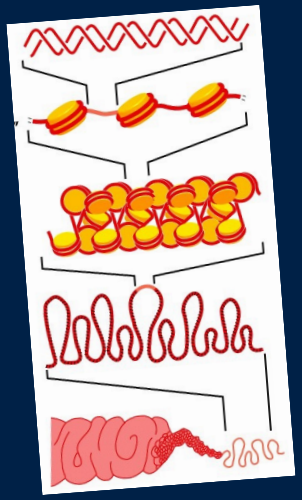
West Africa ⟷ East Africa

The circle of genetic causation

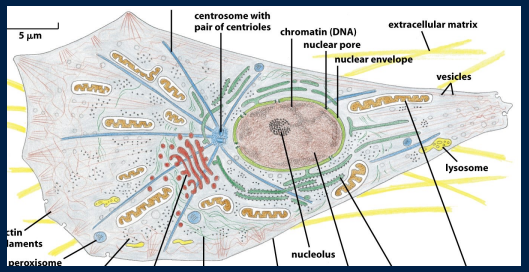




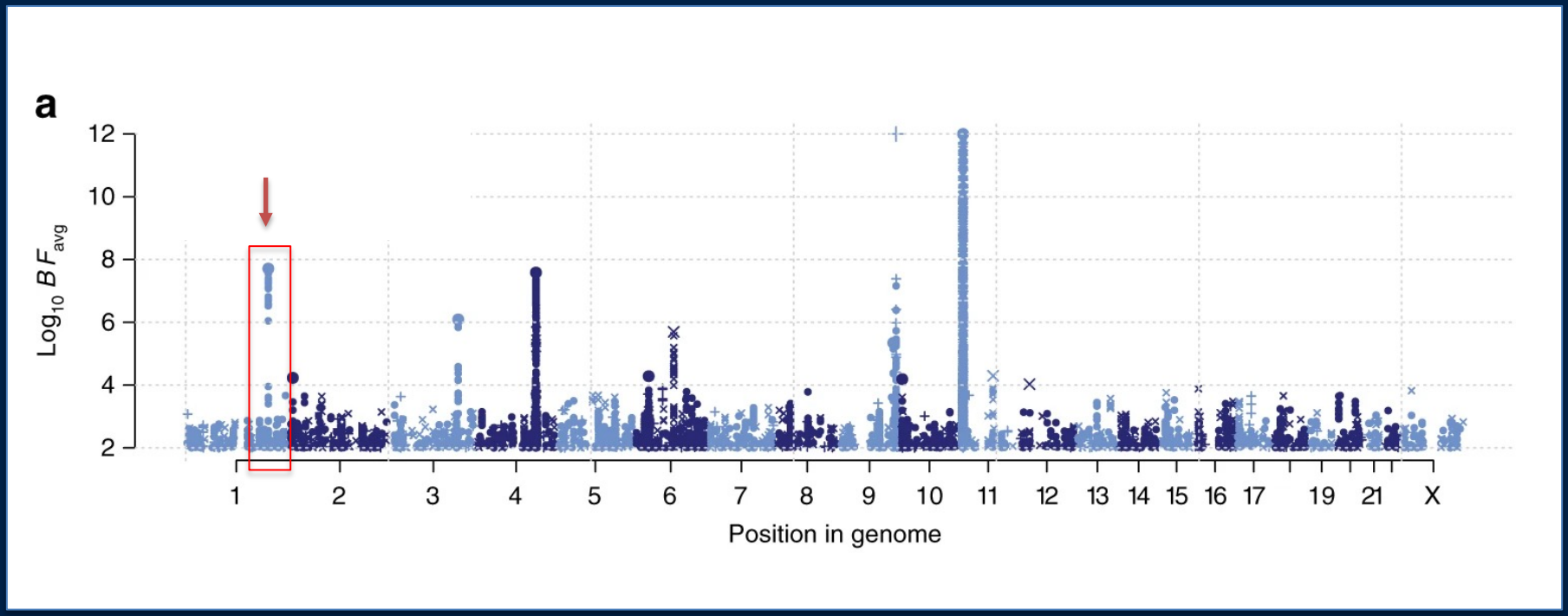
...that combine to make individuals...



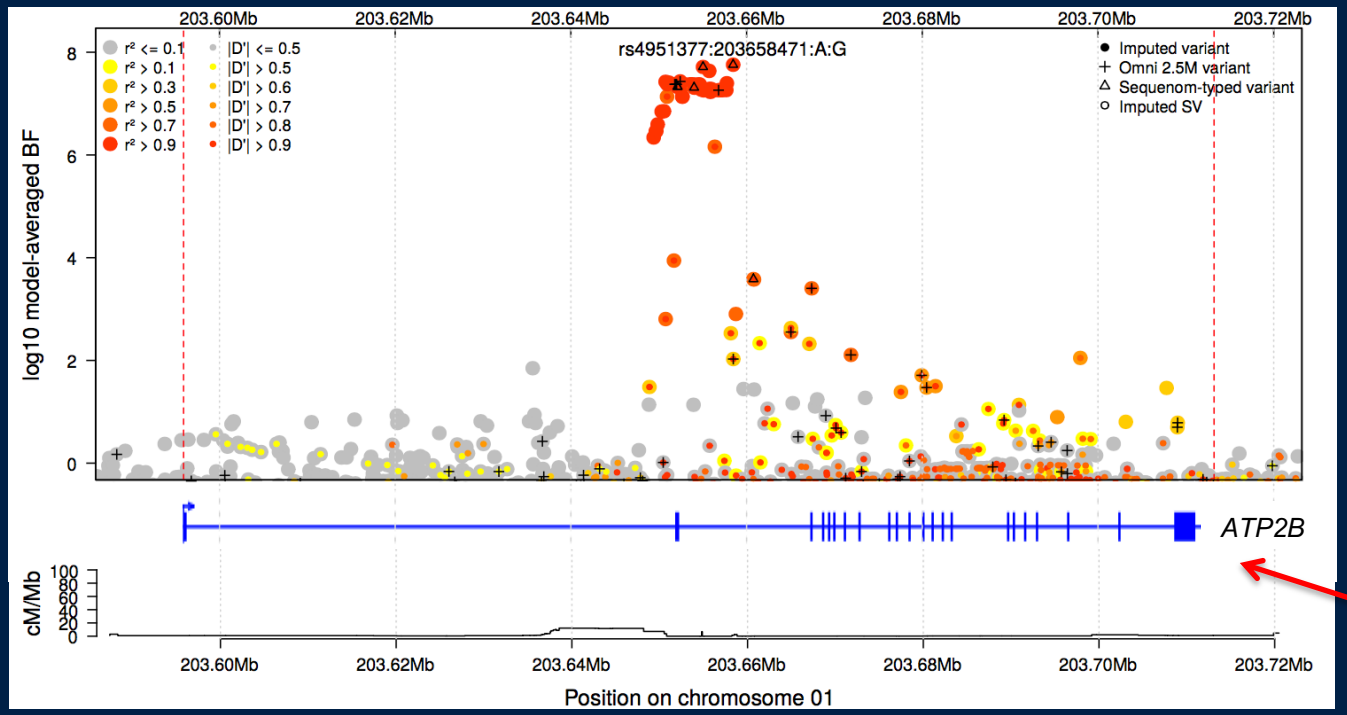
Fine-mapping example 2:
expression complexity



Natural resistance is driven by red blood cell variation



Association near 2nd exon of *ATP2B4*

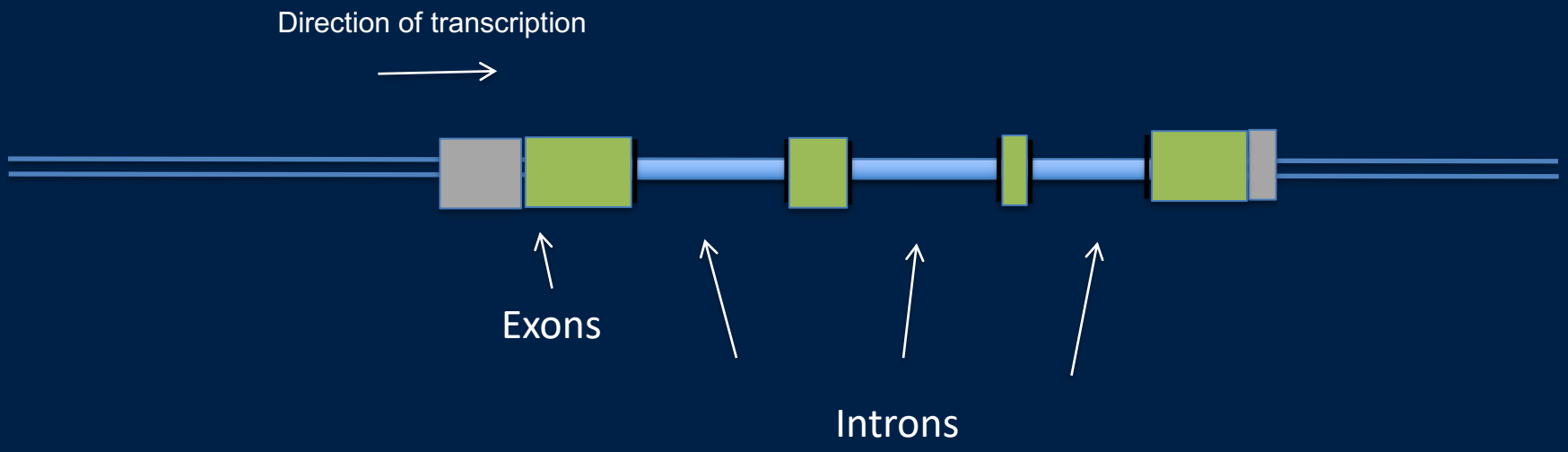


"Canonical" gene model for ATP2B4

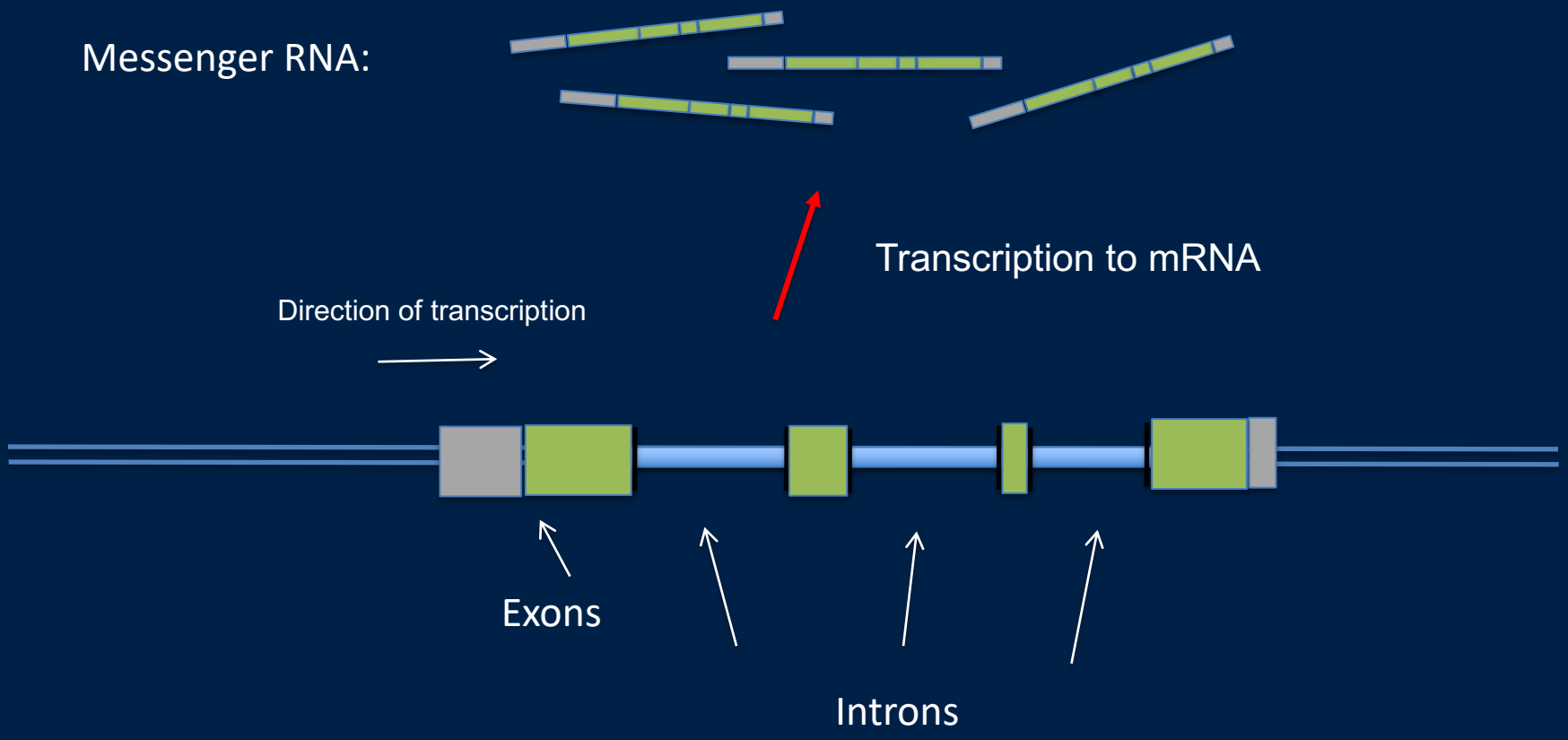
ATP2B4 = a red cell "calcium pump"

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



Cartoon of a gene

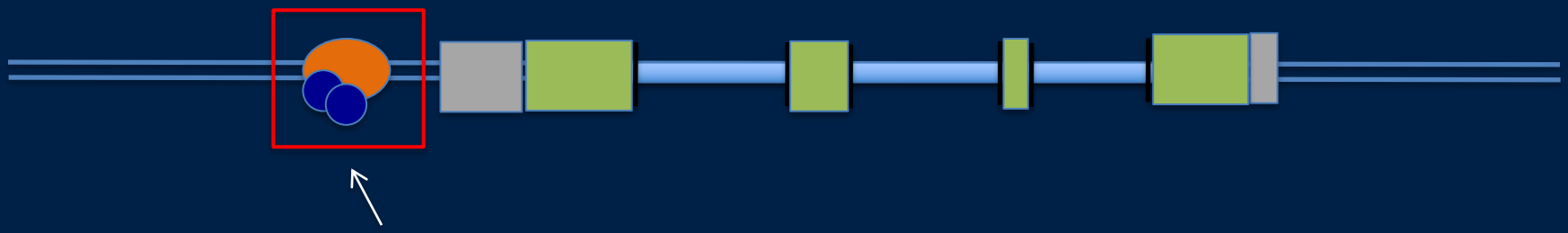


Cartoon of a gene

Messenger RNA:

Transcription to mRNA

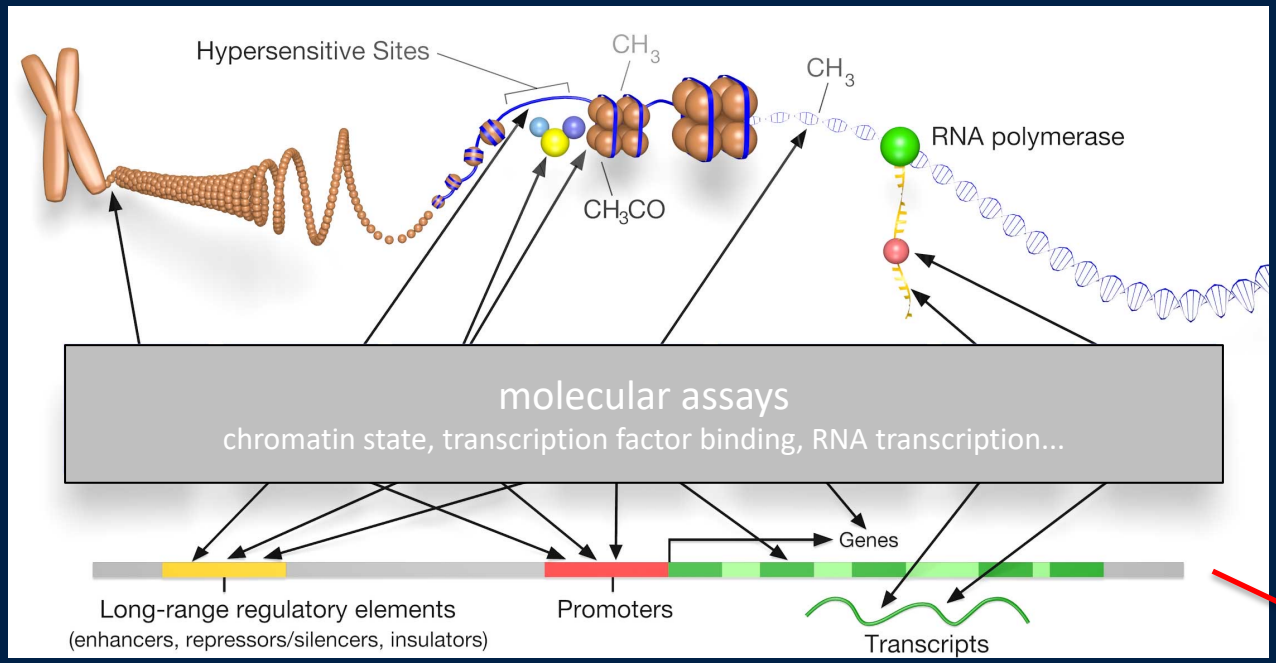
Direction of transcription



The promoter region.

In order for this to take place, the DNA upstream of the gene must be accessible and helpers known as *transcription factors* must be able to bind.

Two ways to look at transcription



Can look at chromatin state

RNA expression

ENCODE

ROADMAP epigenomics PROJECT

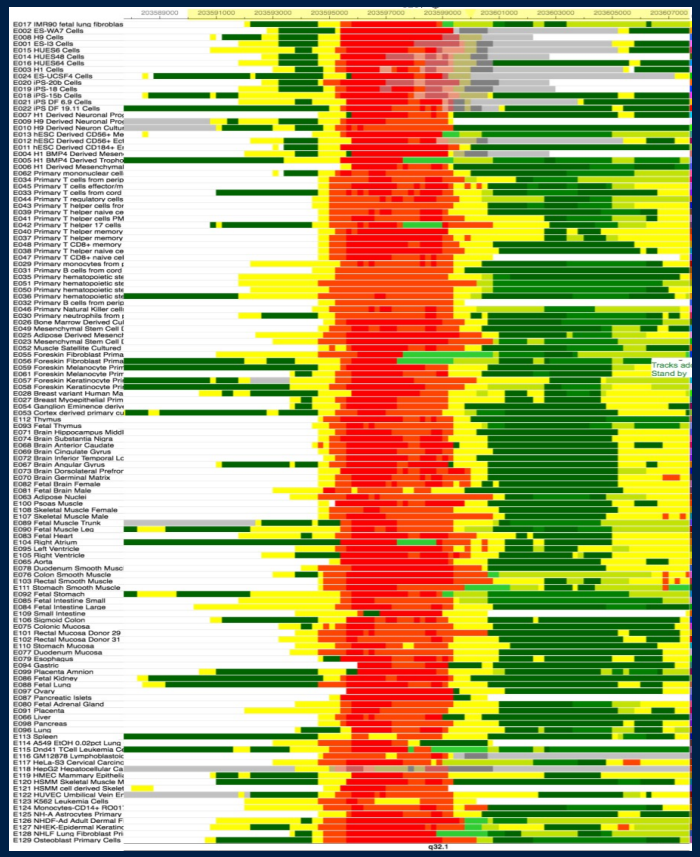
GTEx Portal

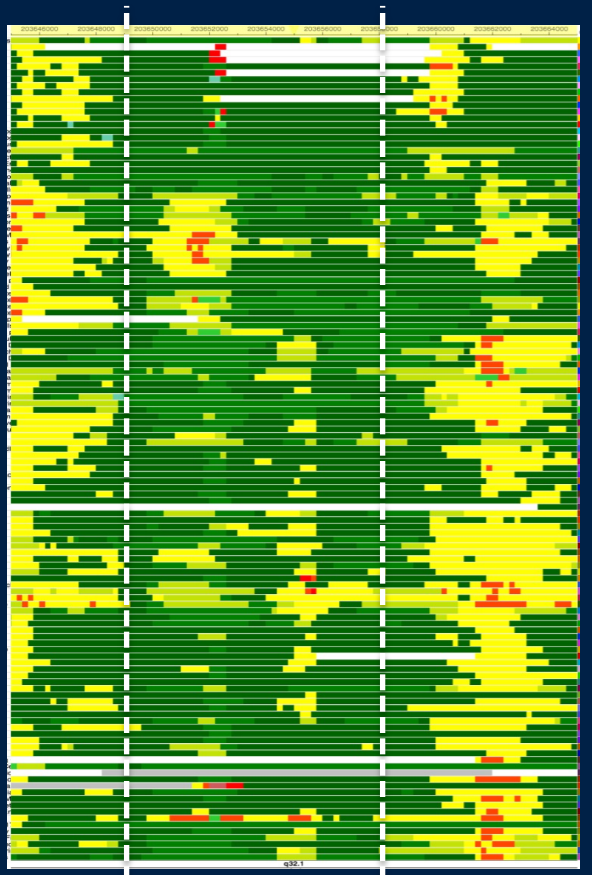
ATP2B4 is widely expressed...

1st exon

2nd exon

Chromatin states in 130 cell types

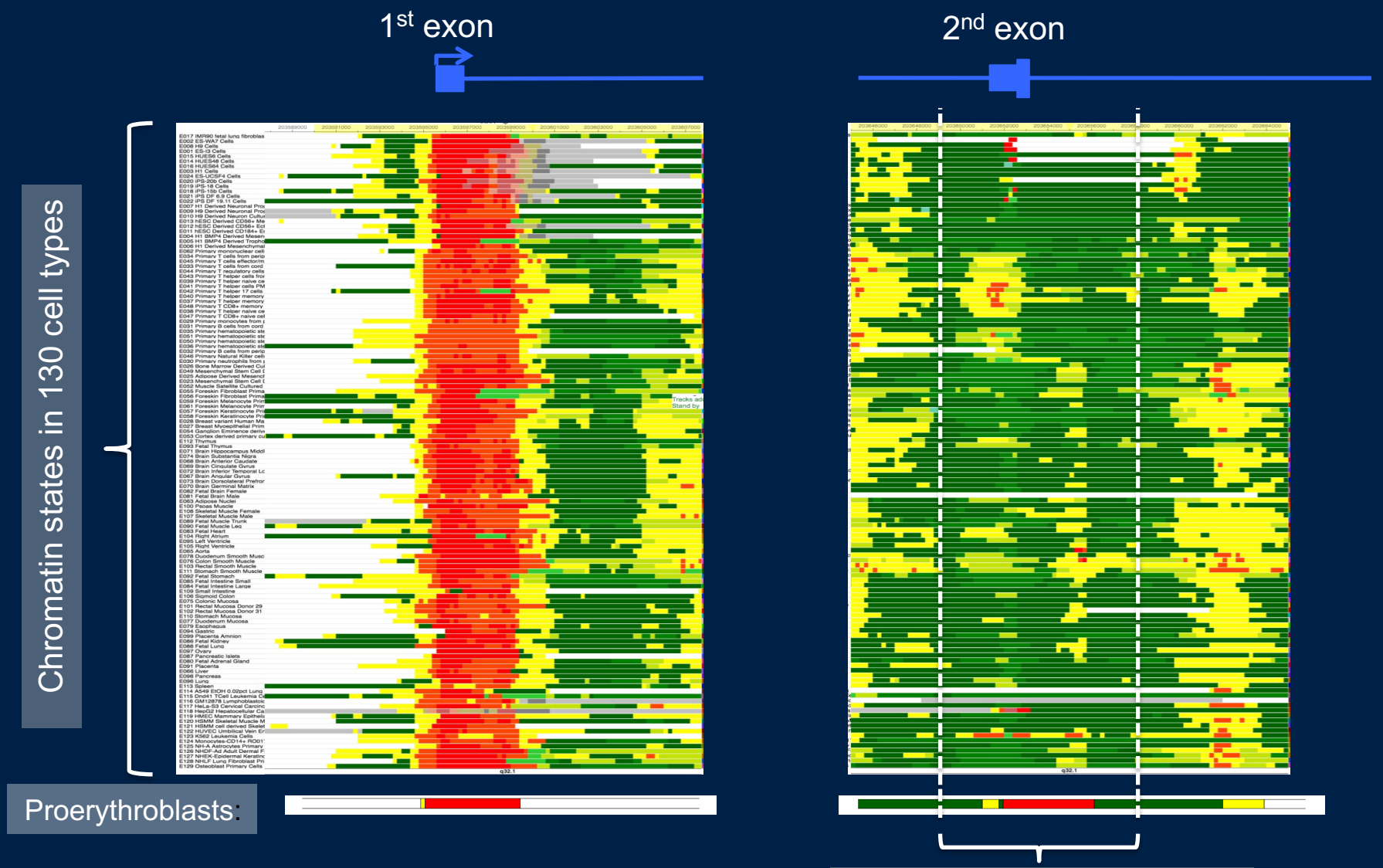




Data from ENCODE / Roadmap

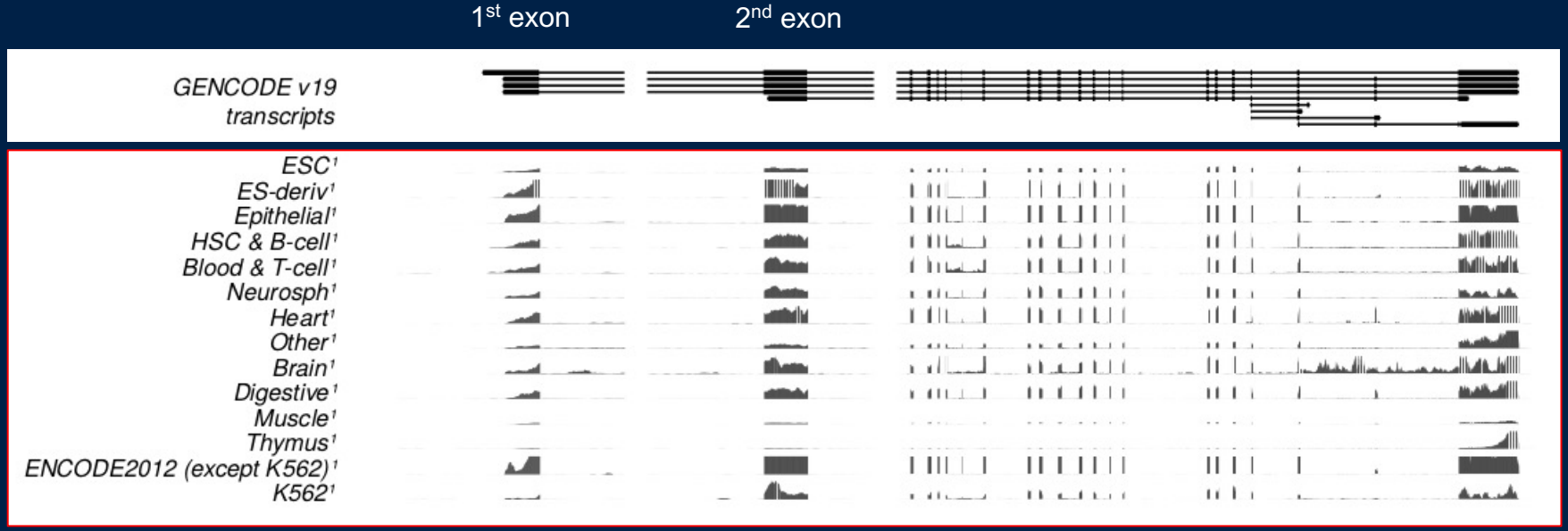
Malaria-associated region

...but shows chromatin differences in RBCs



ATP2B4 is widely expressed…

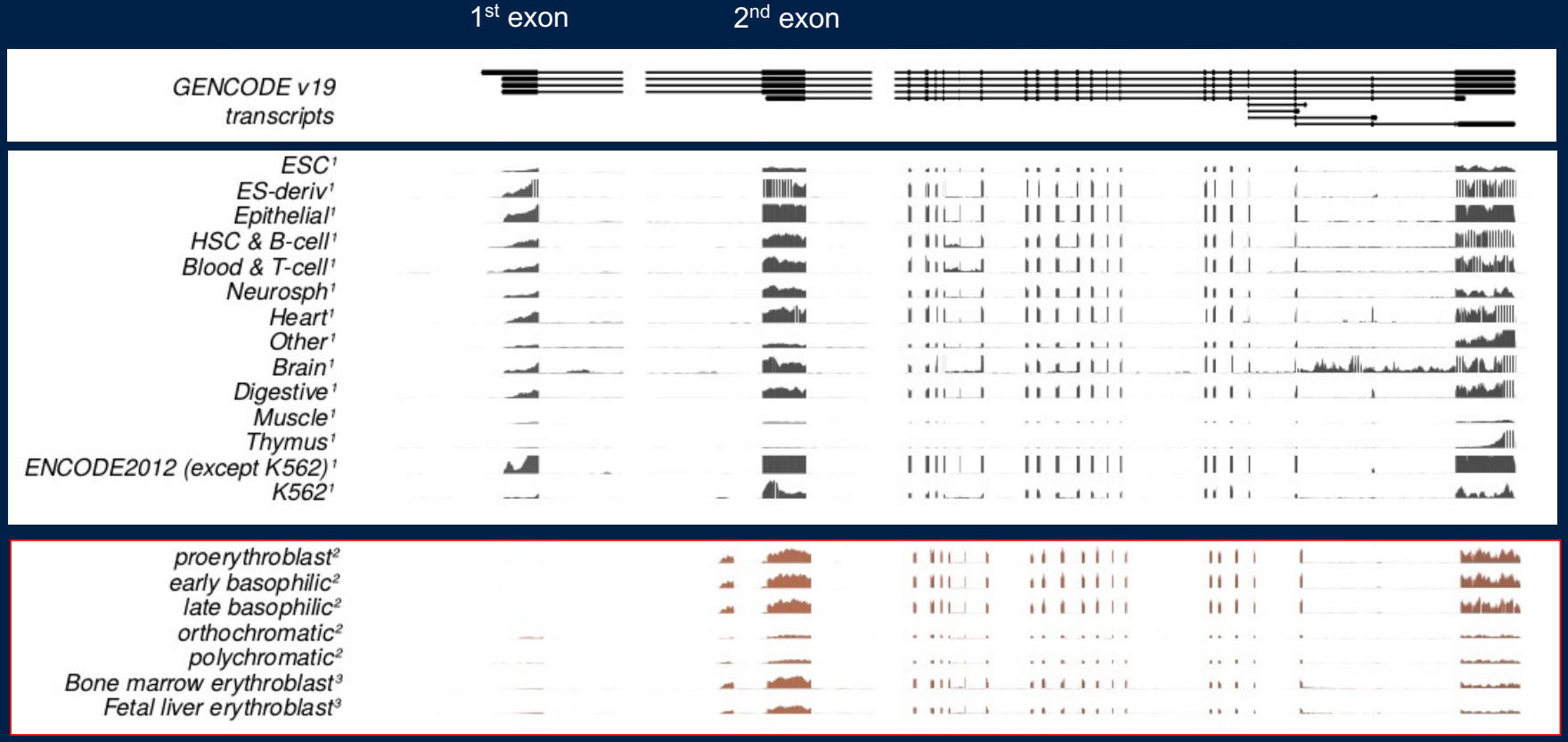
Measured RNA transcription (RNA-seq)



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

ATP2B4 has an erythroid-specific transcript

Measured RNA transcription (RNA-seq)

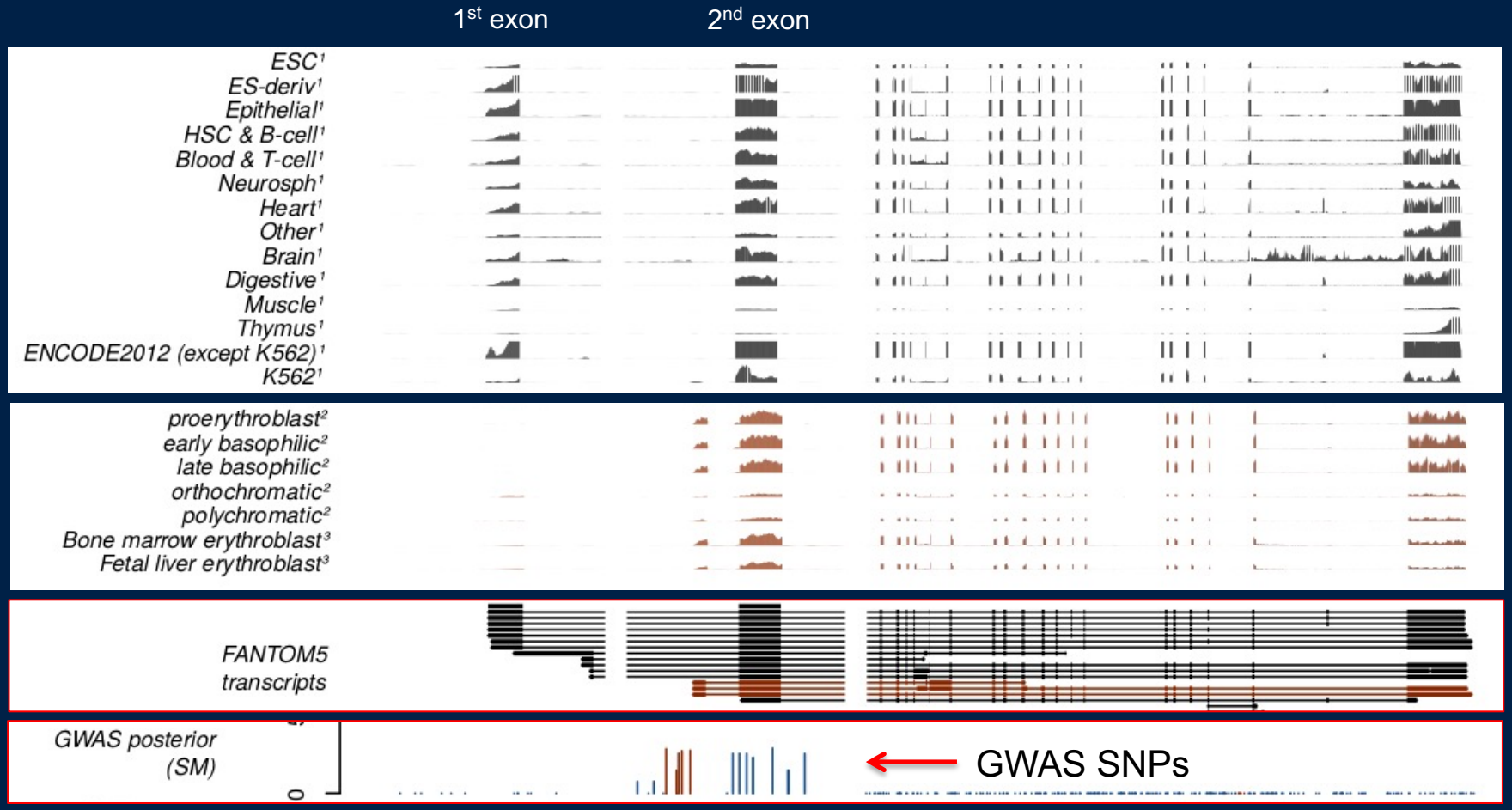


Erythroid cells show a different expression pattern.

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.

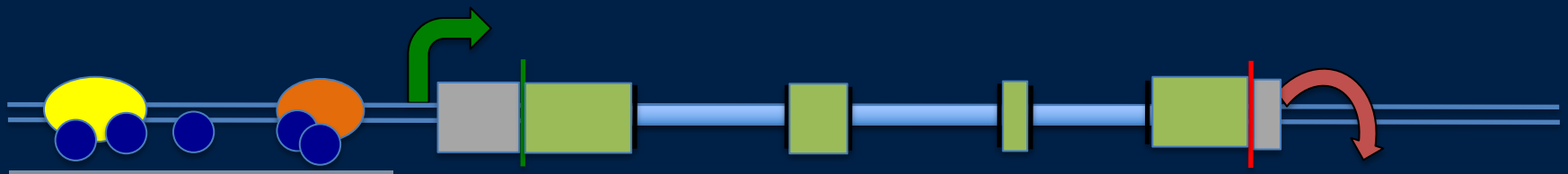
ATP2B4 has an erythroid-specific transcript

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?

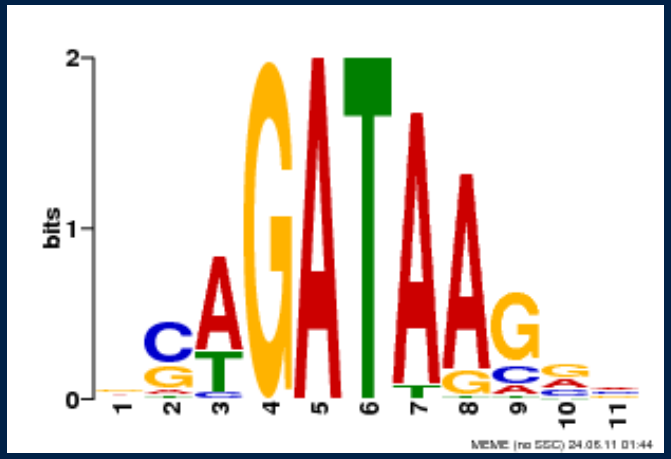
What is different about RBCs?



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.

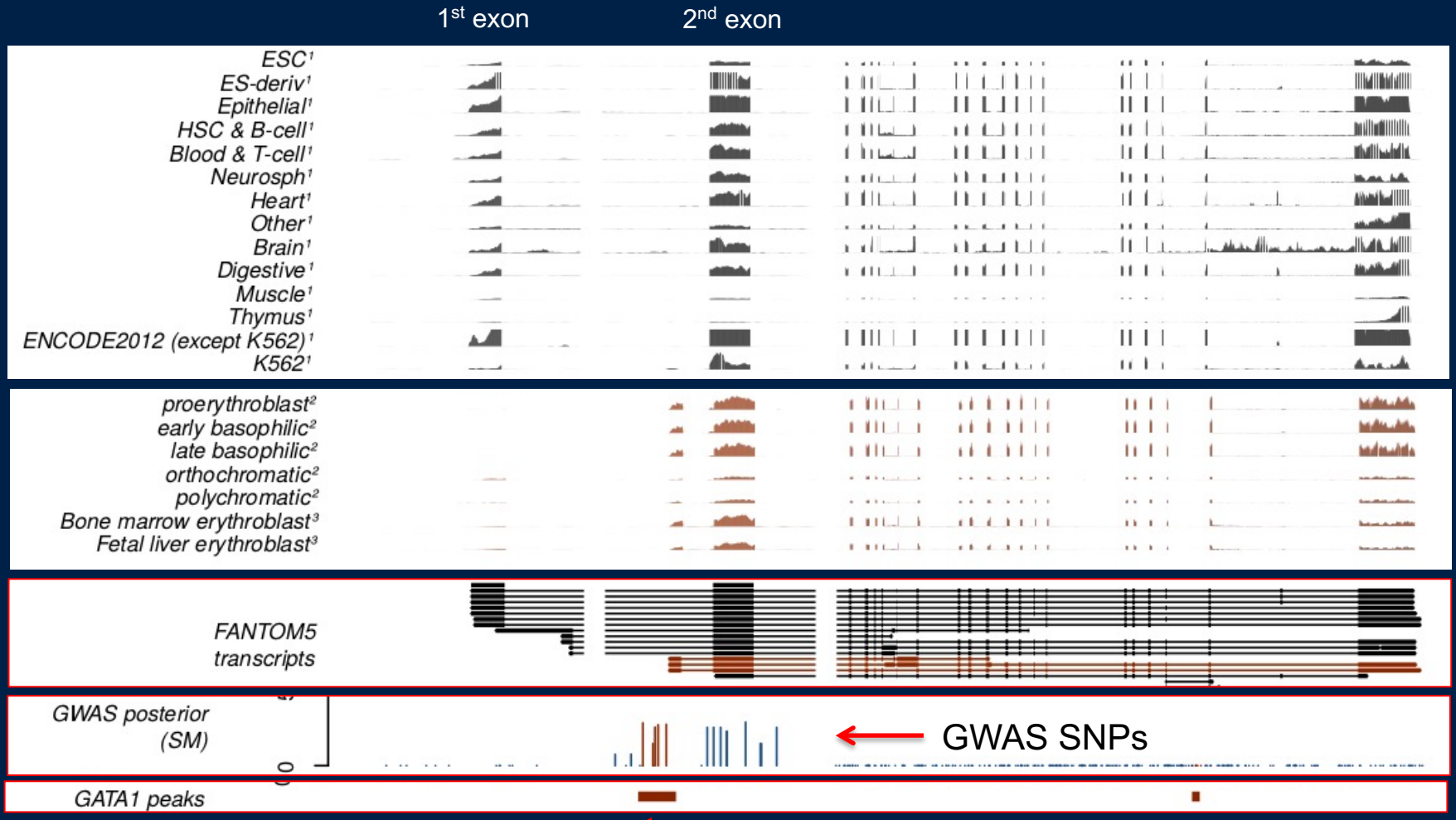
GATA1 is named after the DNA motif it recognises:



v1.factorbook.org

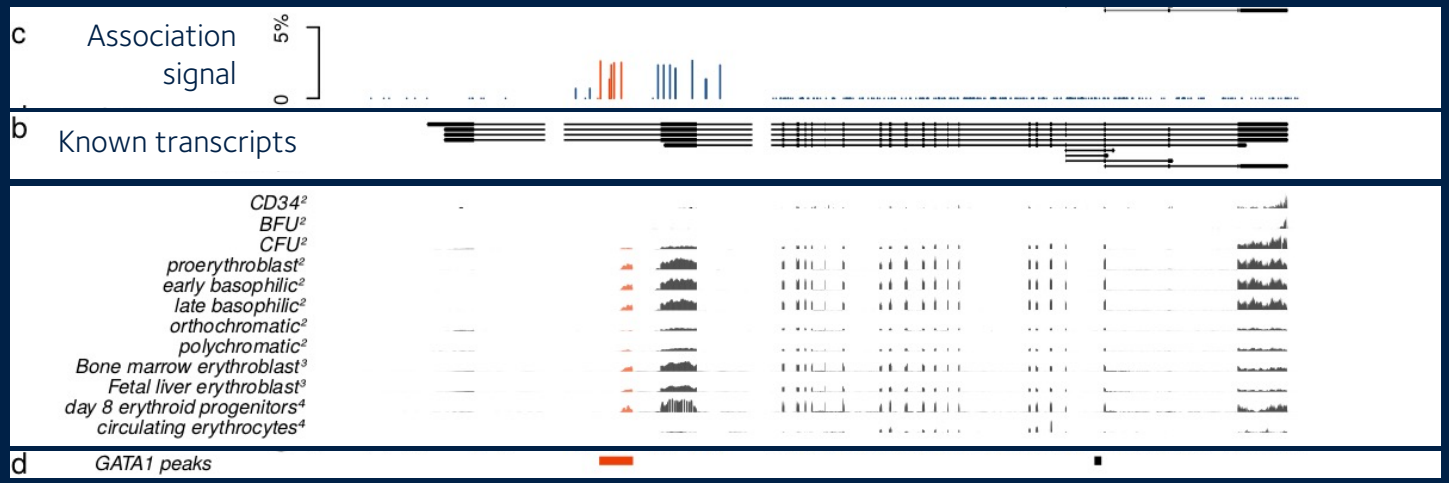
GATA1 binds just upstream of 2nd exon

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.

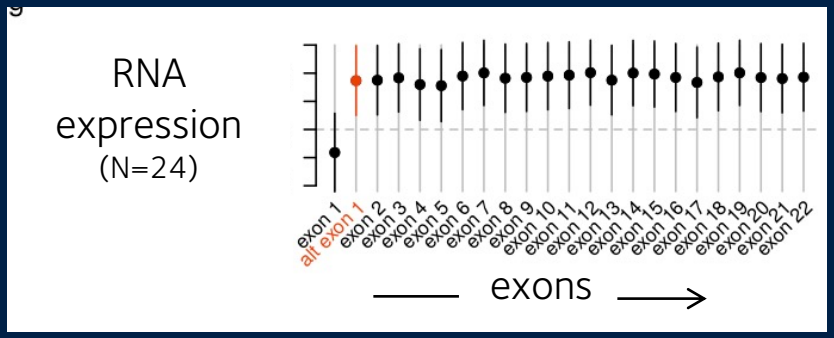
One of the malaria-associated SNPs disrupts the GATA site



rs10715451

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

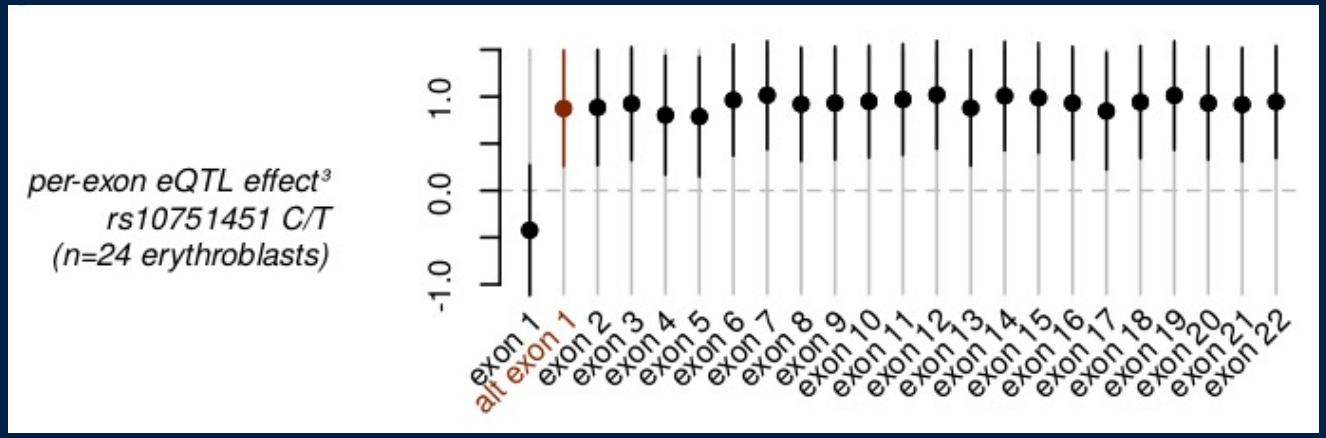
...GGAGCG**A**TAAGATA... (malaria risk allele)



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

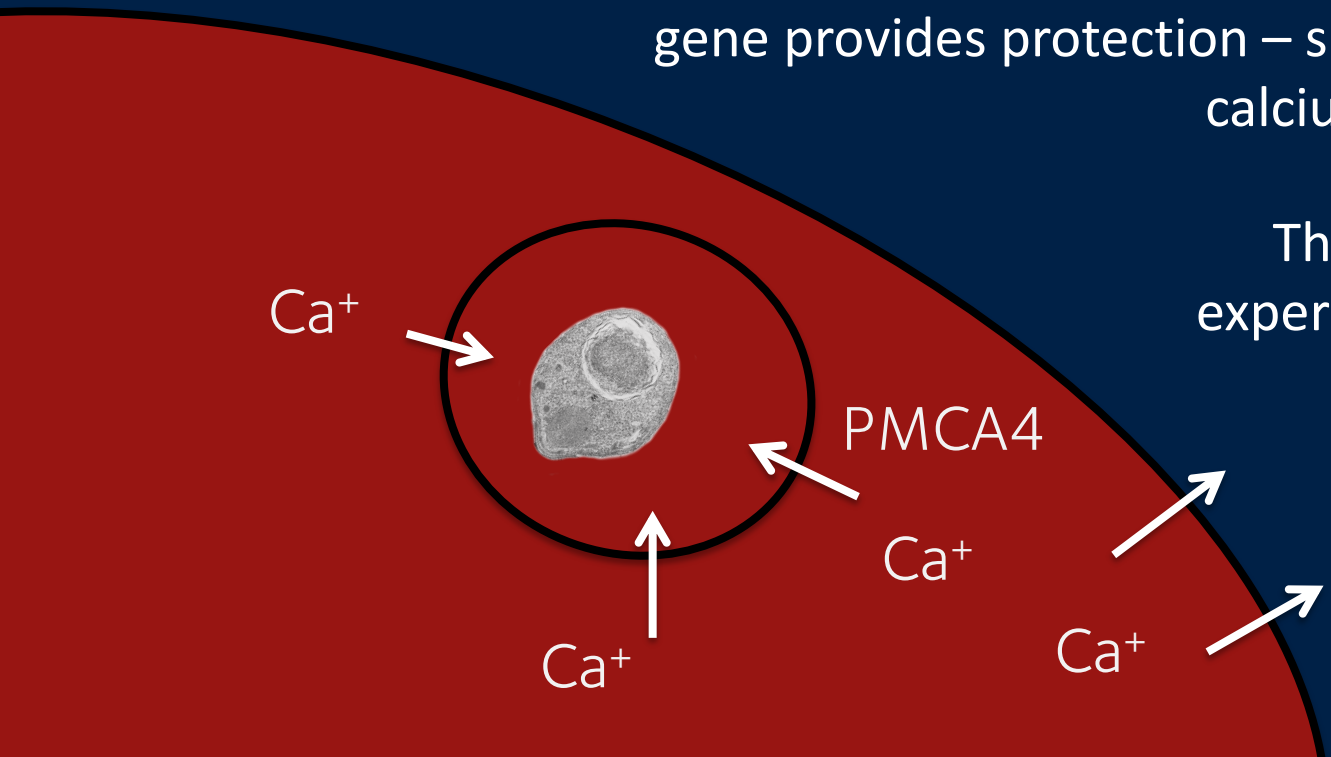
Functional hypothesis

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.

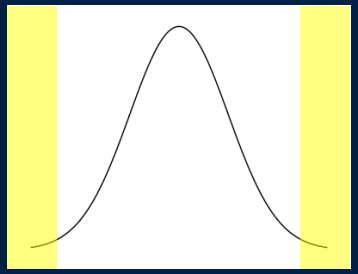
This is a hypothesis - not experimentally tested (yet)!



Zambo et al, Cell Calcium 2017

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

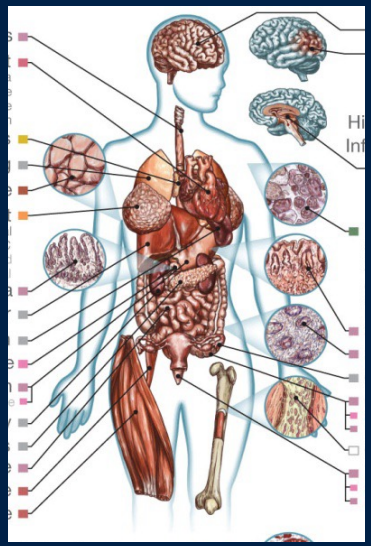
Clinical phenotype measurements

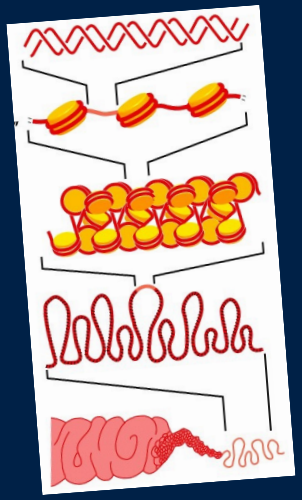
...that gets physically packaged up into chromosomes...

There is complex biology at all stages

Any complication that can happen, does happen!

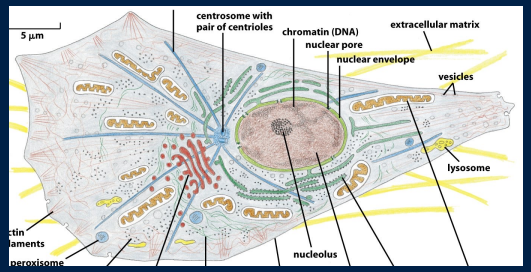
Biomarker measurements





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...

Biology from GWAS

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!

Fine-mapping success stories

Fetal haemoglobin modifiers in sickle cell disease. Gene editing is now possible!

CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia

H. Frangoul, D. Altshuler, M.D. Cappellini, Y.-S. Chen, J. Domm, B.K. Eustace, J. Foell, J. de la Fuente, S. Grupp, R. Handgretinger, T.W. Ho, A. Kattamis, A. Kernytsky, J. Lekstrom-Himes, A.M. Li, F. Locatelli, M.Y. Mapara, M. de Montalembert, D. Rondelli, A. Sharma, S. Sheth, S. Soni, M.H. Steinberg, D. Wall, A. Yen, and S. Corbacioglu

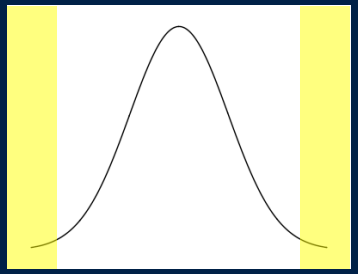
Transfusion-dependent β-thalassemia (TDT) and sickle cell disease (SCD) are severe monogenic diseases with severe and potentially life-threatening manifestations. BCL11A is a transcription factor that represses γ-globin expression and fetal hemoglobin in erythroid cells. We performed electroporation of CD34+ hematopoietic stem and progenitor cells obtained from healthy donors, with CRISPR-Cas9 targeting the *BCL11A* erythroid-specific enhancer. Approximately 80% of the alleles at this locus were modified, with no evidence of off-target editing. After undergoing myeloablation, two patients — one with TDT and the other with SCD — received autologous CD34+ cells edited with CRISPR-Cas9 targeting the same *BCL11A* enhancer. More than a year later, both patients had high levels of allelic editing in bone marrow and blood, increases in fetal hemoglobin that were distributed pancellularly, transfusion independence, and (in the patient with SCD) elimination of vaso-occlusive episodes. (Funded by CRISPR Therapeutics and Vertex Pharmaceuticals; ClinicalTrials.gov numbers, NCT03655678 for CLIMB THAL-111 and NCT03745287 for CLIMB SCD-121.)

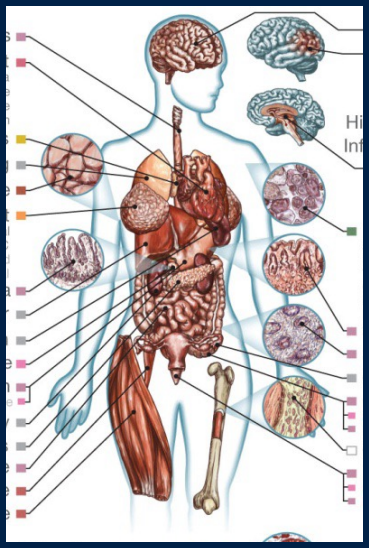
Lecture plan

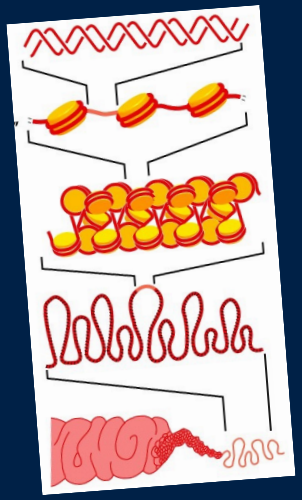
- Recap from last lecture – GWAS and the common variant / common trait hypothesis
- How polygenic are traits anyway?
- The challenge of fine-mapping
- Summary

Conclusions and summary

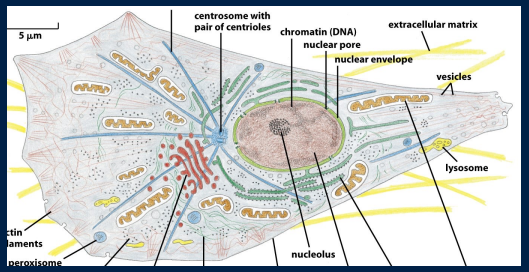
- Most human traits are *highly heritable*
- For 'complex' traits, the effects are made up of many genetic variants often with modest effects.
- Traits vary in genetic architecture - sometimes up to tens of thousands of polymorphisms are involved!
- Fine-mapping is generally hard, but sometimes possible
- A major frontier is to understand the biology and translate these findings into clinically useful insights and predictions.







Thanks!


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