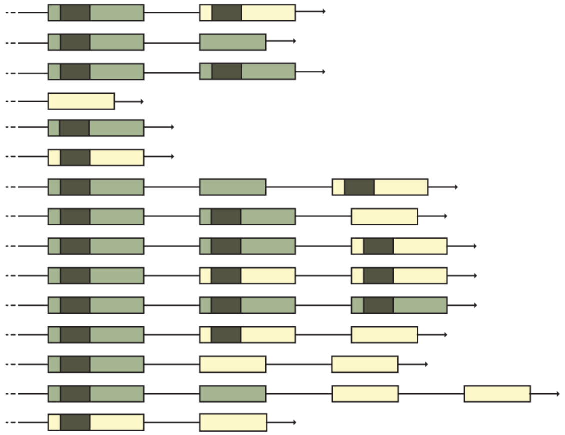
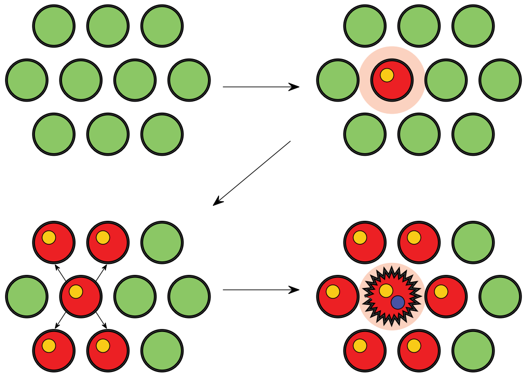
Structural and multi-allelic variation



Steve McCarroll

Harvard Medical School

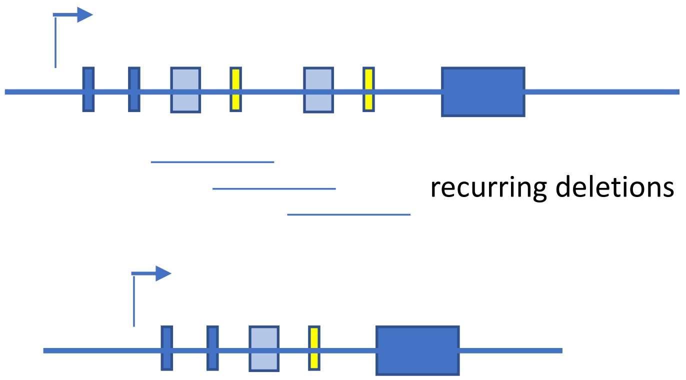
Broad Institute



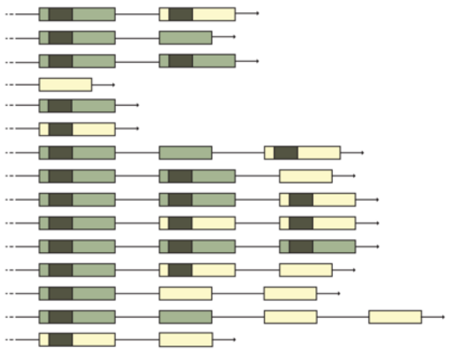
Structural variation

- Large CNV mutations – fairly easy to detect, imp. for some cases, usu. *de novo* (not large component of heritability)
- Common, diallelic structural alleles –
  - WGS data + new analysis methods have led to far better data resources
  - Today part of VCFs etc. from 1000 Genomes Project
  - Routinely imputed into GWASs and meta-analyses
- Structurally unstable loci
  - have rearranged multiple times among human ancestors
  - many structurally and functionally distinct alleles
  - more challenging to analyze

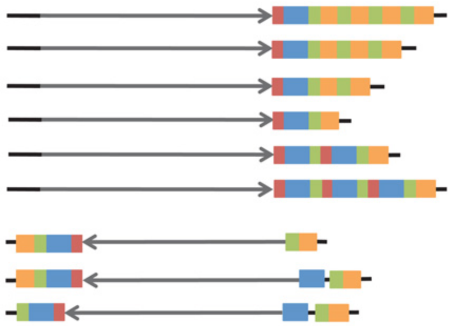
Loci with recurring structural mutations and many functionally distinct alleles



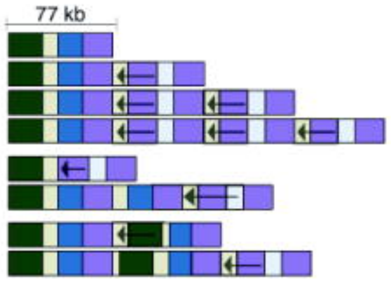
Haptoglobin (*HP*)
Boettger, et al., ...*Nat Genet* 2016



Complement component 4 (*C4*)
Sekar, et al., *Nature* 2016

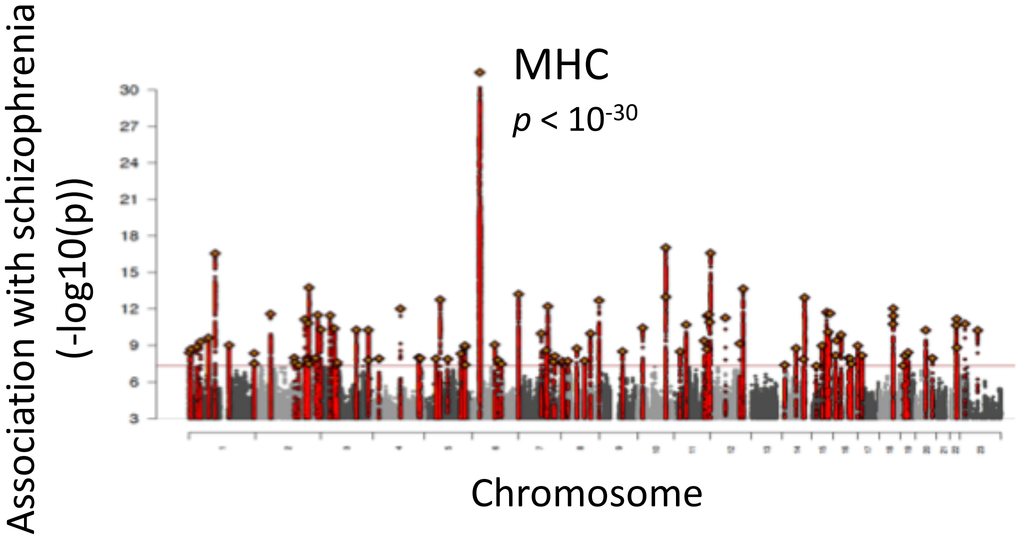


17q21.1 / *MAPT*
Boettger, et al., *Nat Genet* 2012



AMY1 / AMY2
Usher, *et al., Nat Genet* 2015

Strongest association in schizophrenia is to the MHC locus

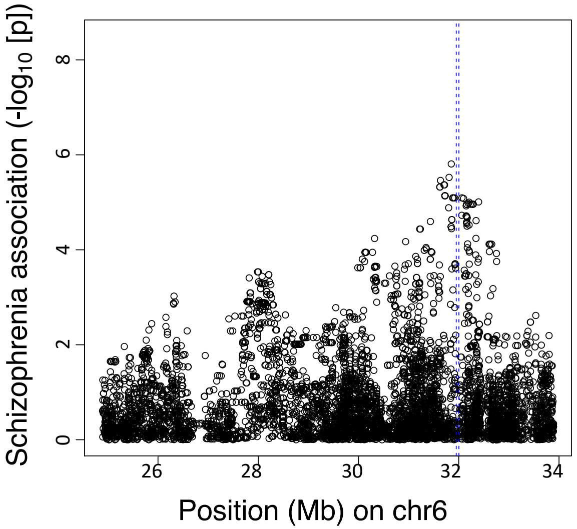


MHC

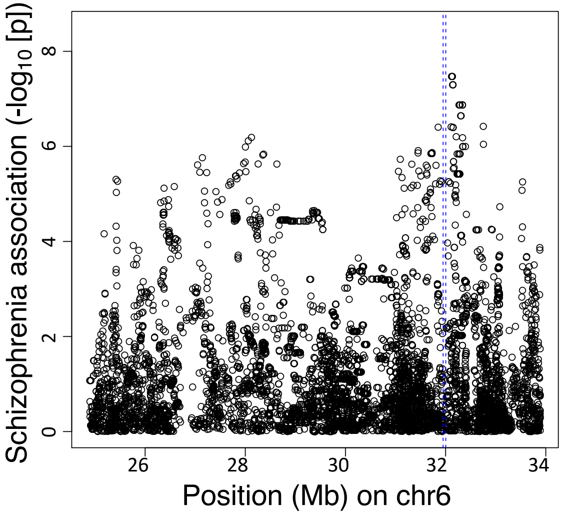
$p < 10^{-30}$

Several cohorts share a curious association peak

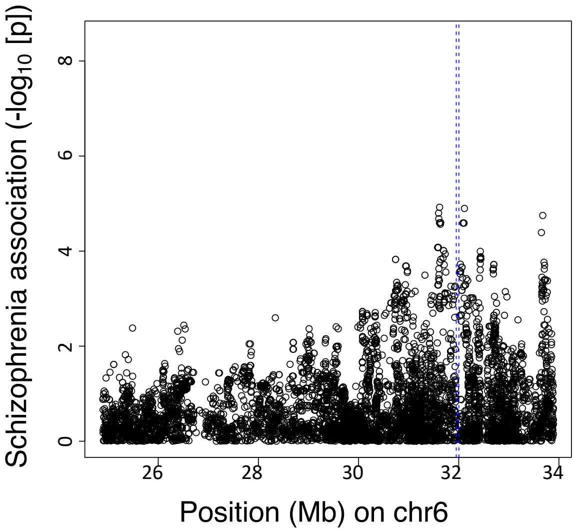
Sweden



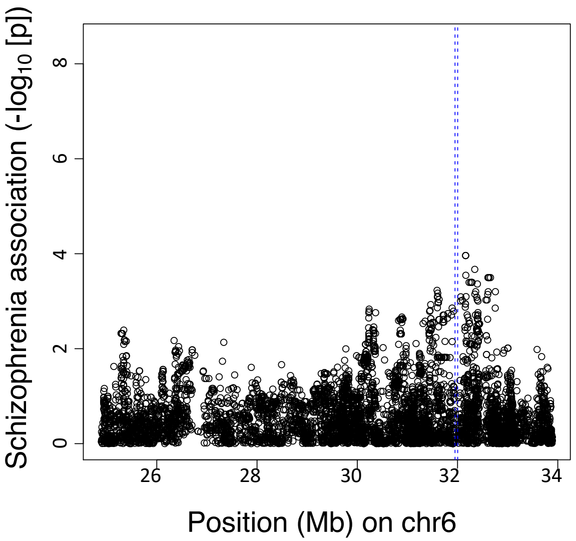
England



Ireland

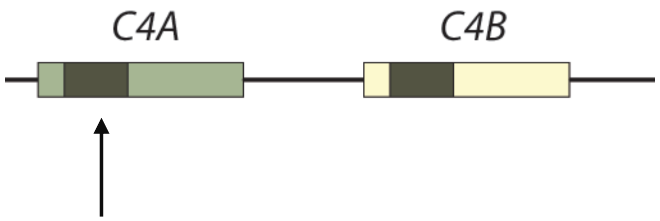


Scotland



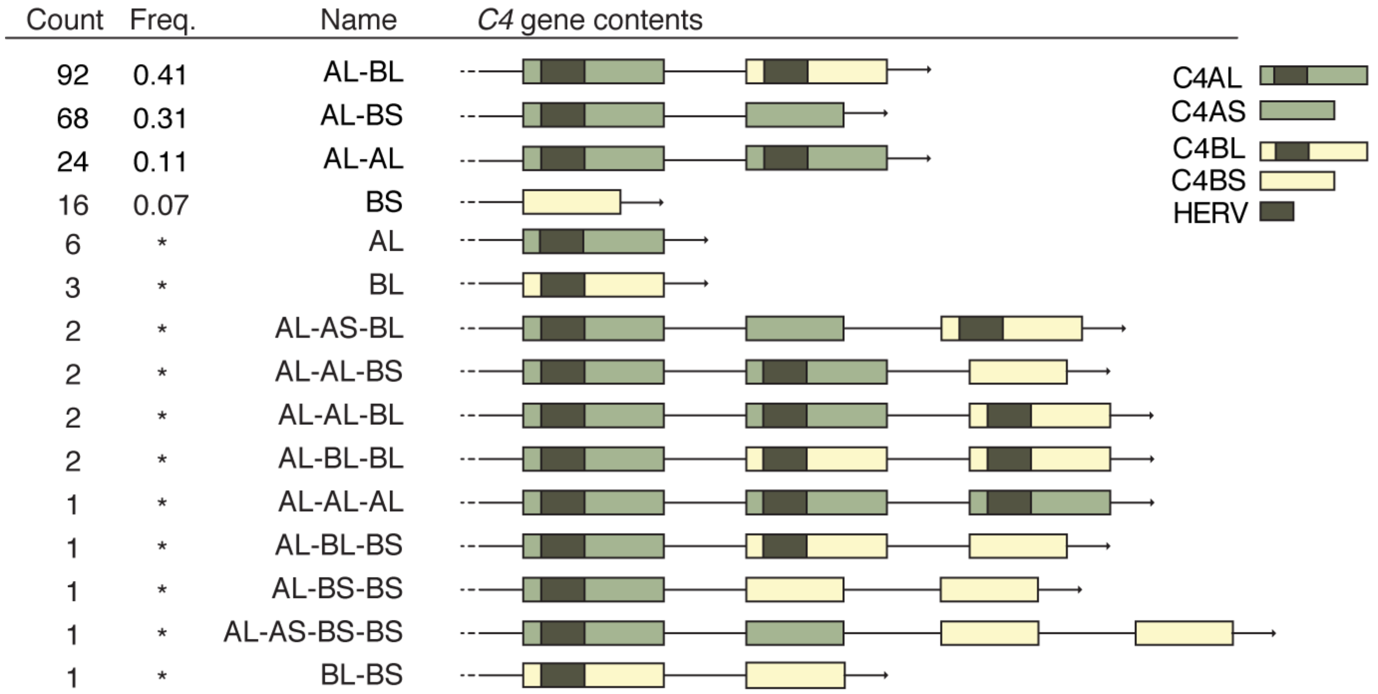
Doesn’t correspond to the linkage disequilibrium **around any known variant**

Complement component 4 (*C4*) genes



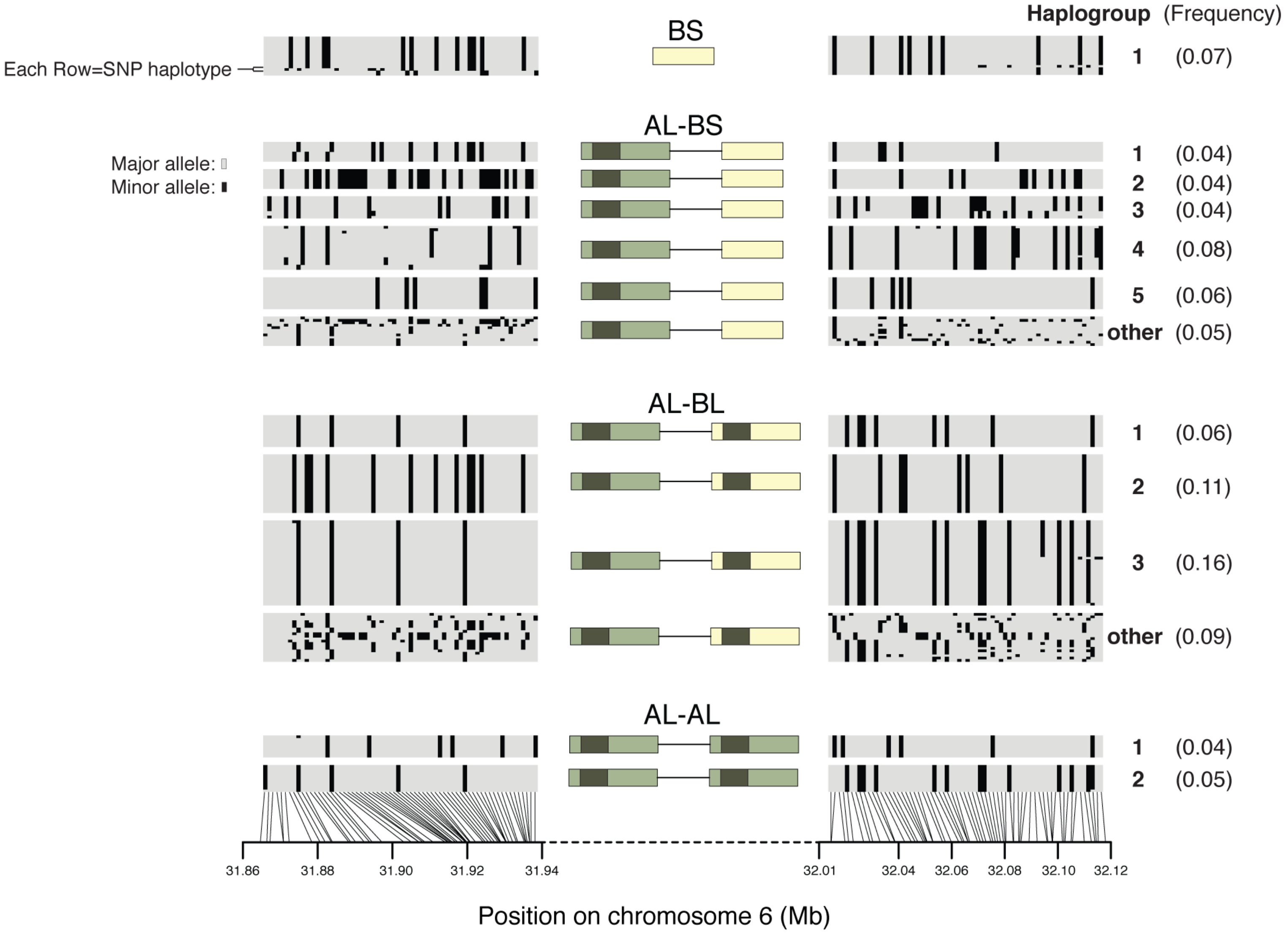
Ancient retroviral insertion
brain-specific enhancer

Paralogous genes
encoded proteins opsonize material for elimination,
bind to different sites in tissues



Count	Freq.	Name	*C4* gene contents
92	0.41	AL-BL	
68	0.31	AL-BS	
24	0.11	AL-AL	
16	0.07	BS	
6	*	AL	
3	*	BL	
2	*	AL-AS-BL	
2	*	AL-AL-BS	
2	*	AL-AL-BL	
2	*	AL-BL-BL	
1	*	AL-AL-AL	
1	*	AL-BL-BS	
1	*	AL-BS-BS	
1	*	AL-AS-BS-BS	
1	*	BL-BS	

C4 structures form haplotypes with SNPs

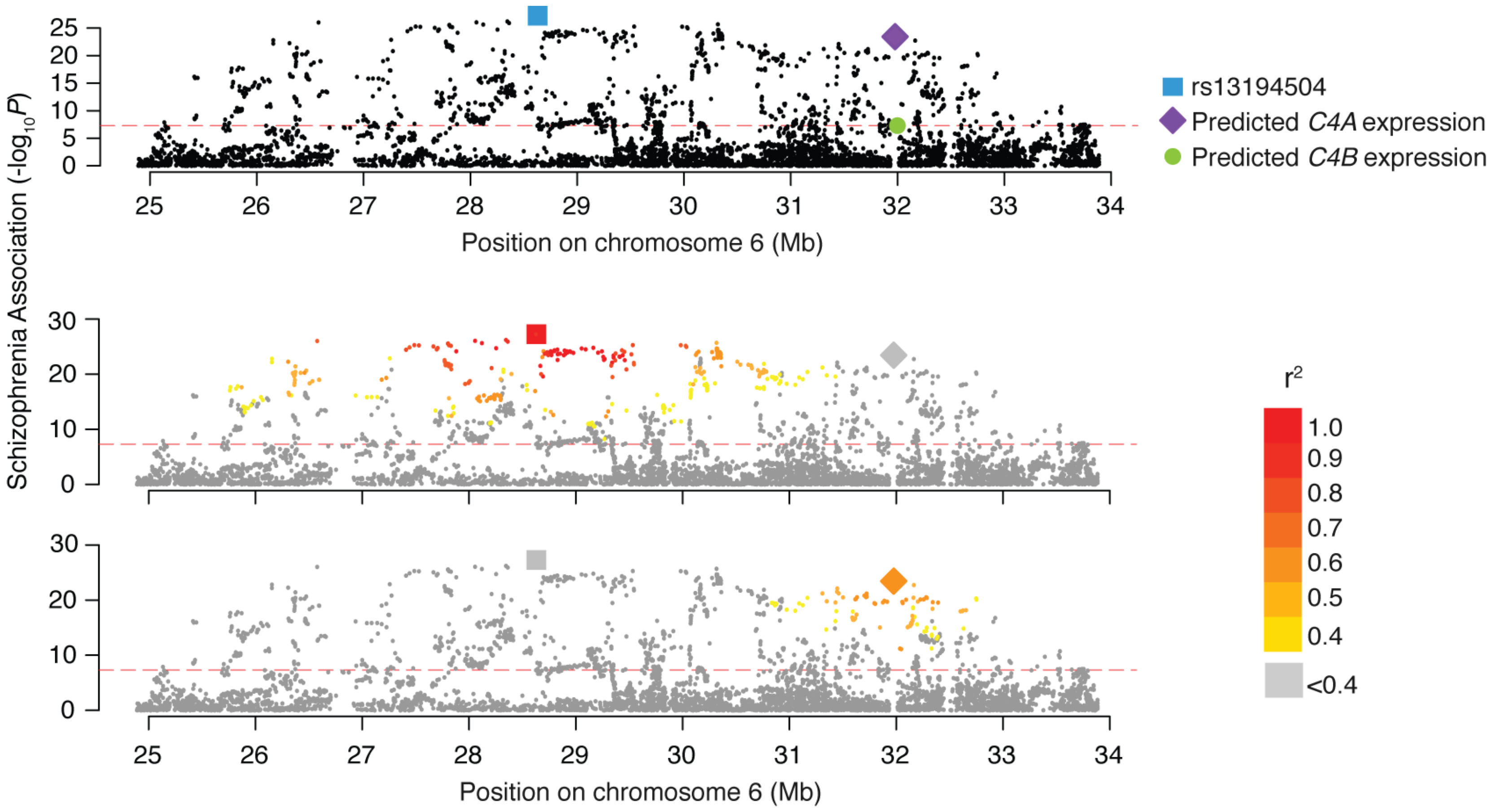


1. If C4 affected phenotype, could generate unconventional patterns of association across SNPs.

2. Might be possible to analyze C4 structures by imputation from existing SNP data.

Sekar, *et al., Nature* 2016

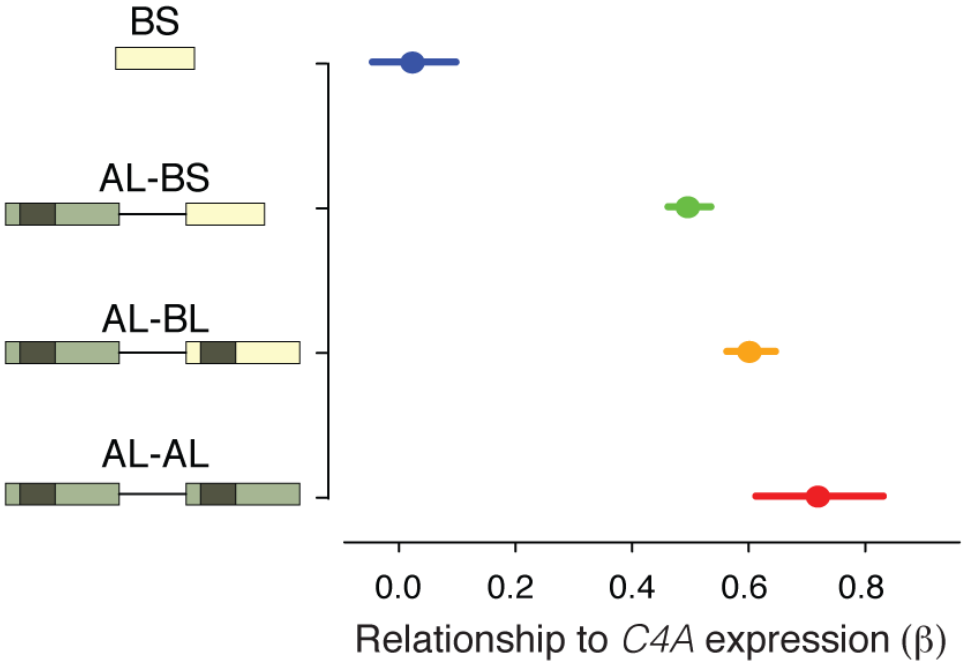
Imputing C4 alleles reveals two association peaks in the MHC – one at *C4*



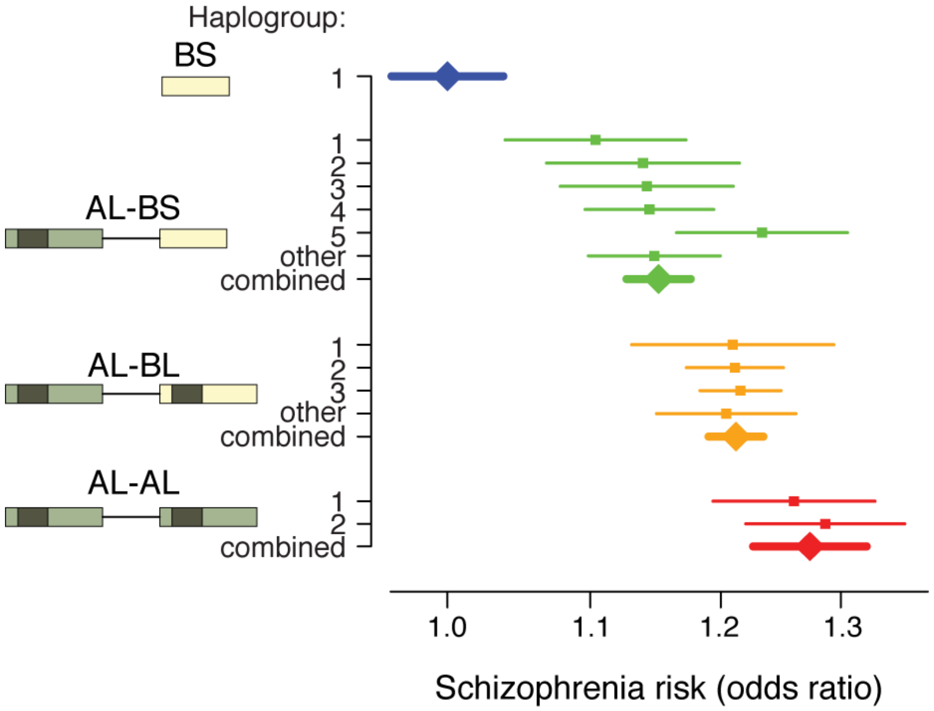
Sekar, *et al.*, *Nature* 2016

Associations to specific *C4* alleles

The more **C4A RNA expression** an allele generates...

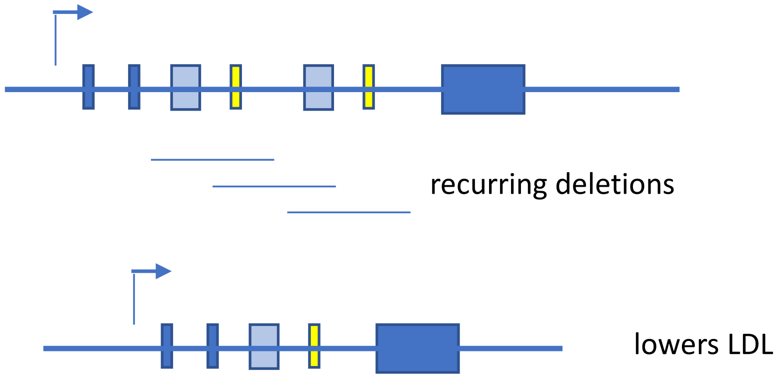


... the more **schizophrenia risk** it confers

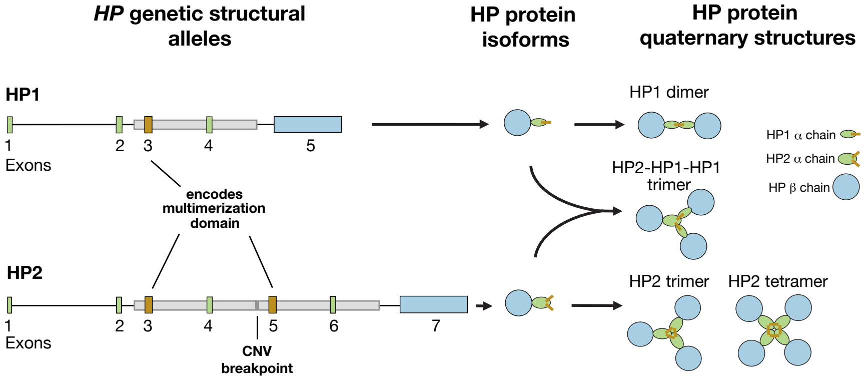


Sekar, *et al.*, *Nature* 2016

Recurring exon deletions in haptoglobin (*HP*) and blood cholesterol



Haptoglobin (*HP*)



Deletions (in fact, reversions of an ancient duplication) of two exons that encode a multimerization domain

Recur in every generation

- A few common alleles (due to old mutations)
- Also many rare alleles (due to new mutations)

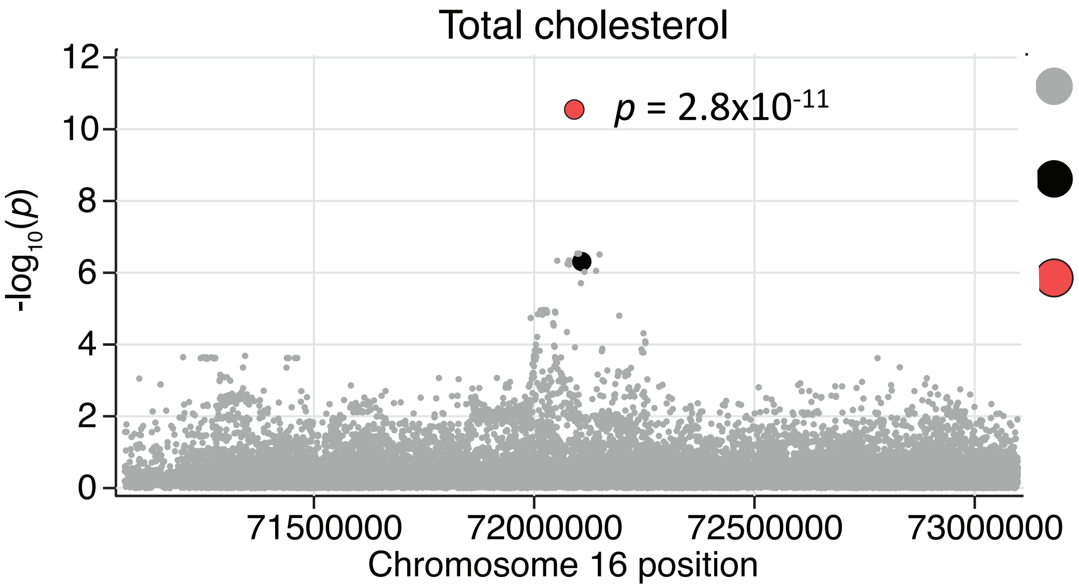
Cause HP to circulate in the blood as a dimer (rather than a trimer)
The dimer is a more-efficient antioxidant for ApoE

Reduce blood cholesterol by 2.1 mg/dl
Explain GWAS associations near this locus

Act together with a nearby SNP (1.5 mg/dl effect) that regulates *HP* expression level

Boettger, et al., ...*Nat Genet* 2016

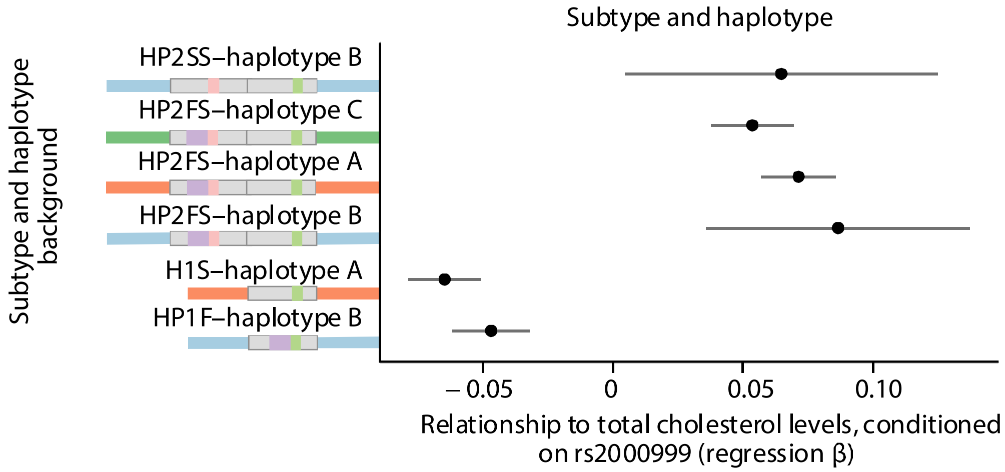
An imputed HP2/HP1 predictor associates much more strongly than any SNP near *HP* does



● 16q22 SNPs

● GWAS index SNP (rs2000999)

● HP2 vs. HP1



Boettger, et al., …*Nat Genet* 2016

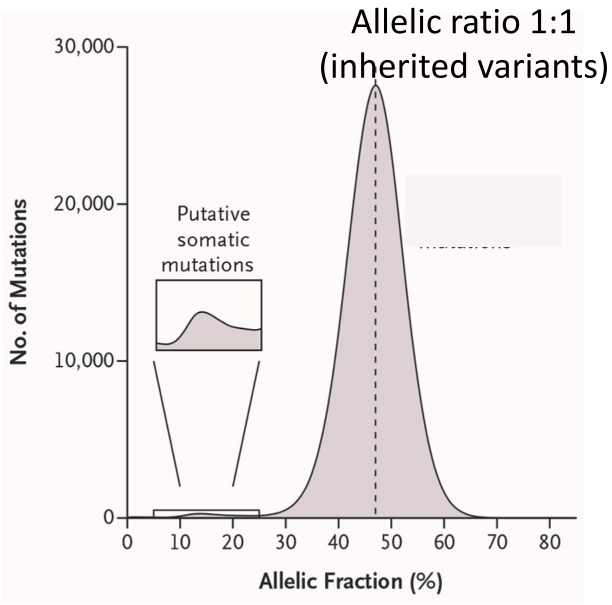
How much “missing heritability” is explained?

- **Across the genome**, very small contribution, because only 0-1 loci with complex variation (that we know of) implicated in any given disease (smaller-scale VNTRs could contribute also though)
- At **individual loci**,
    - locus explained 2-4 times more variance than the “lead SNP” did (may be true at many other loci also)
        - individual SNPs only partially correlated with the full spectrum of allelic influences at the locus
    - value not from ∂h^2 but from *series of alleles* with *interpretable effects*

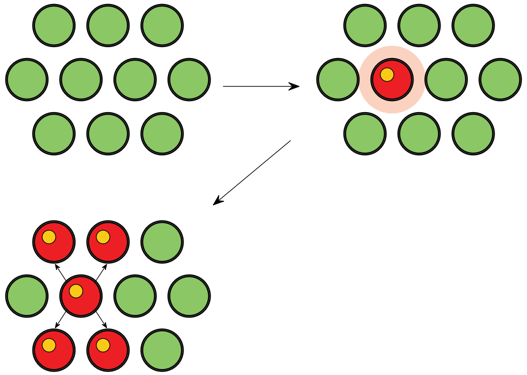
Acquired mutations, clonal expansions, and missing heritability

Really?

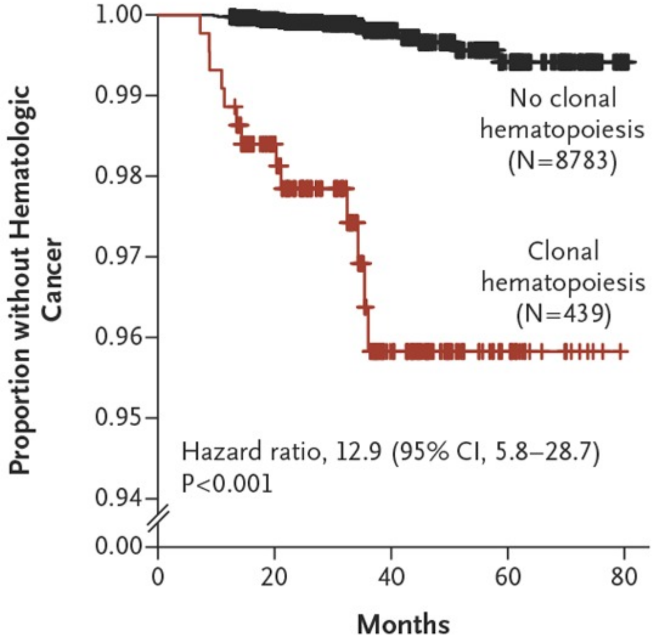
Common clonal expansions of blood cells with mutations



Somatic mutations
concentrated in **blood-cancer** genes
(*TET2, AXSL1, DNMT3A*)



Model:
acquired mutation
+ clonal expansion

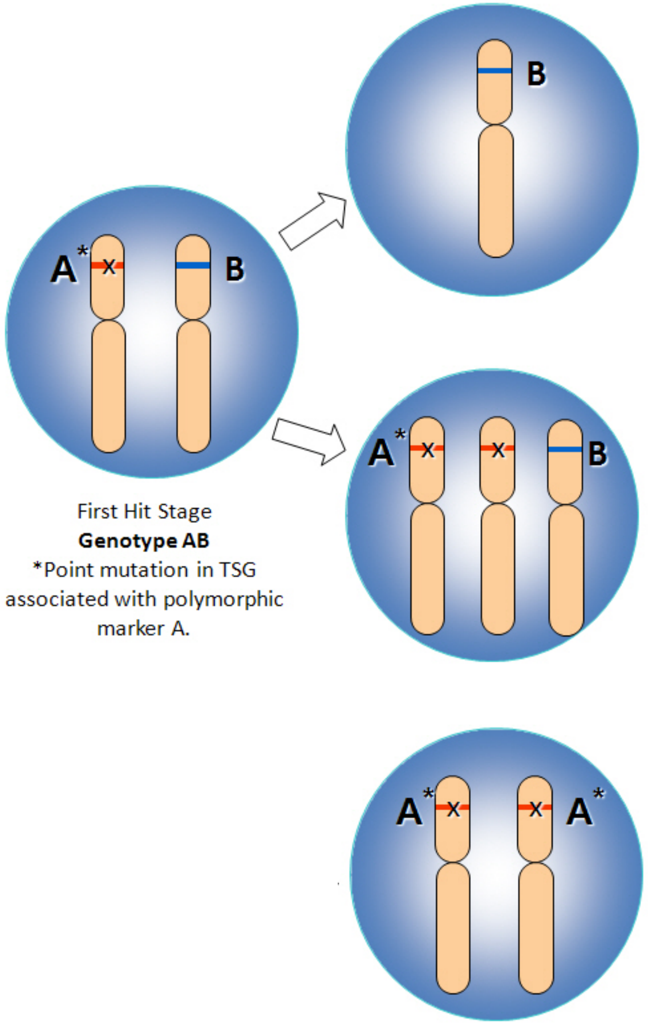


12x increased risk for later blood cancer

Genovese, ..., McCarroll. *NEJM* 2015
discovered independently by Jaiswal, ... Ebert. *NEJM* 2015

Also strongly affects cardiovascular disease risk
(Jaiswal, Kathiresan, Ebert, *et al.,* 2017)

A subset of acquired mutations affect entire segments or chromosomes


B
A*
x
B
First Hit Stage
Genotype AB
*Point mutation in TSG associated with polymorphic marker A.
A*
x
x
B
A*
x
x
A*

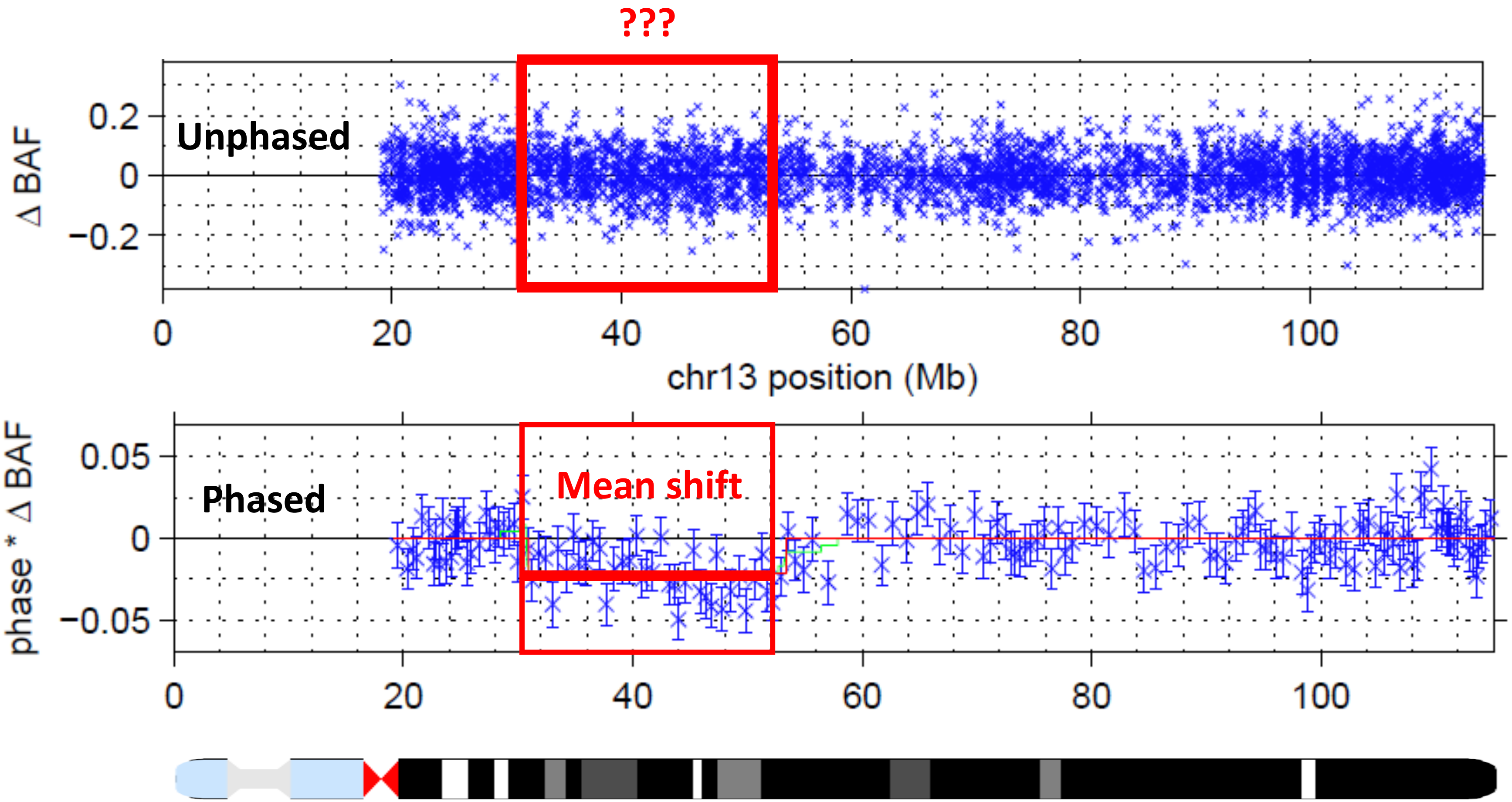
Loss / deletion

Gain / duplication

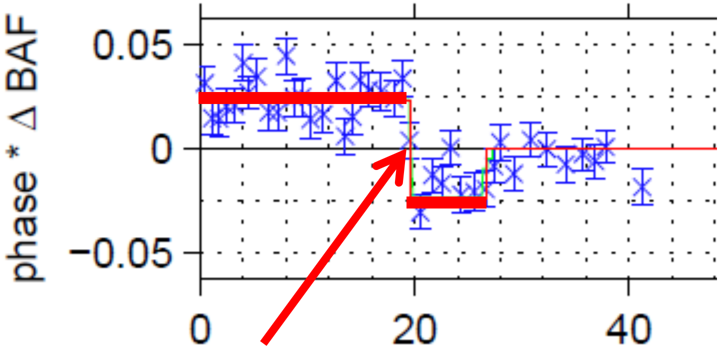
CNN-LOH
(copy-number-neutral
loss of heterozygosity)

All affect
allelic ratios
along a genomic segment

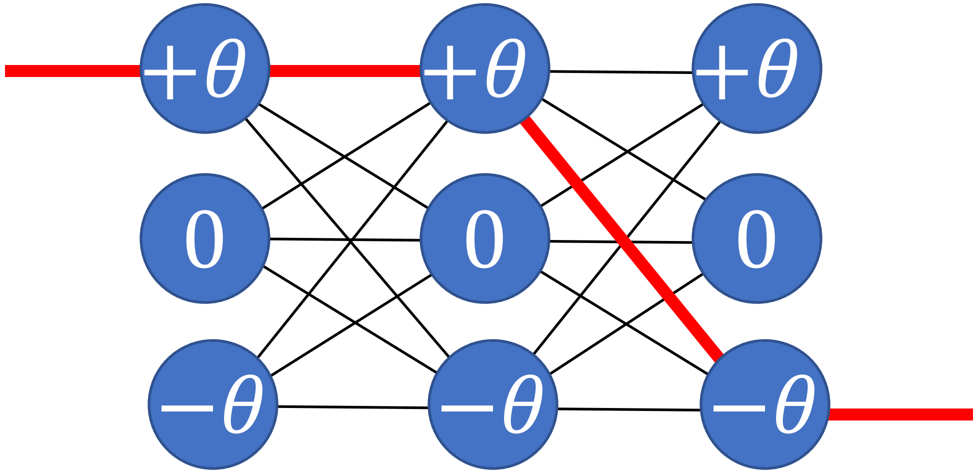
Knowing the chromosomal phase of heterozygous SNPs helps detect mosaic segmental mutations



Phasing without relatives is imperfect, but this can be addressed computationally



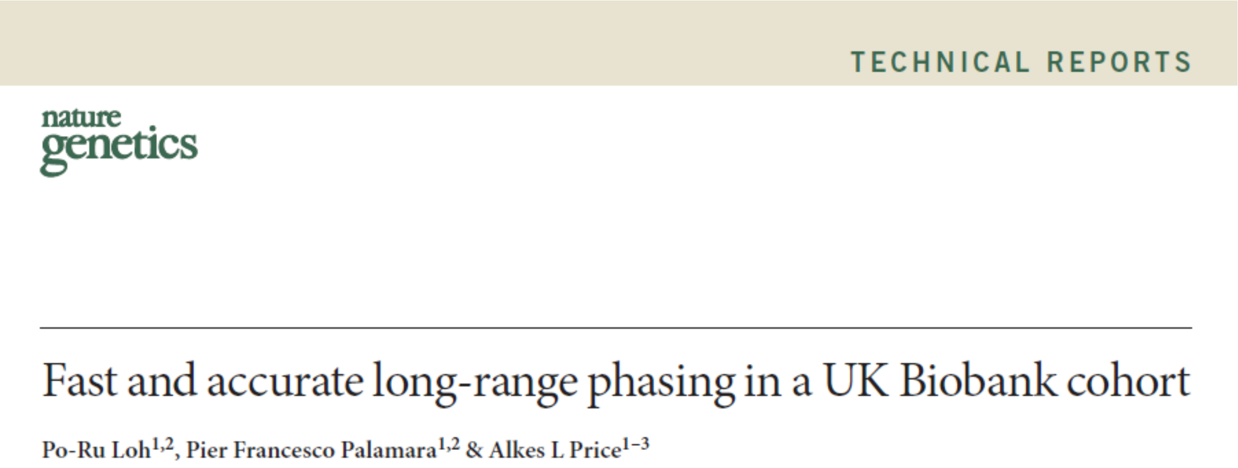
Phase switch error



- HMM:
  - 1 parameter: θ = |ΔBAF| in mosaic region
  - 3 states: E[phase*ΔBAF] = $+\theta, 0, -\theta$
- Detection procedure:
  - Compute LRT statistic for testing $\theta \neq 0$
  - Calibrate empirically using permutation

Giulio Genovese

Recent innovations allow population-scale phasing

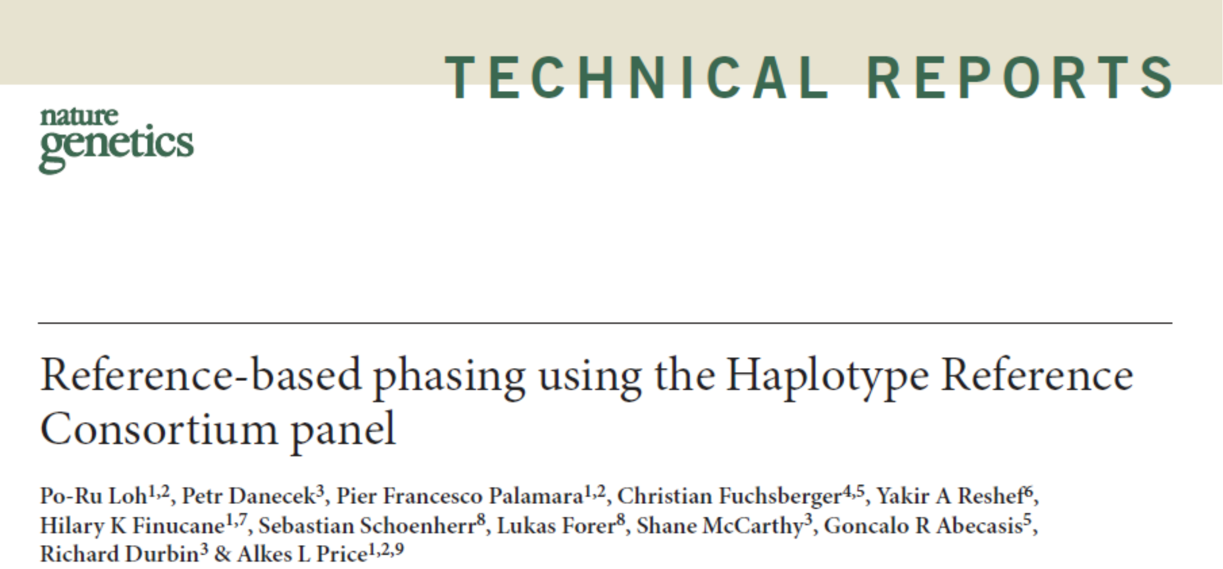


TECHNICAL REPORTS

nature genetics

Fast and accurate long-range phasing in a UK Biobank cohort

Po-Ru Loh[1,2], Pier Francesco Palamara[1,2] & Alkes L Price[1–3]



TECHNICAL REPORTS

nature genetics

Reference-based phasing using the Haplotype Reference Consortium panel

Po-Ru Loh[1,2], Petr Danecek[3], Pier Francesco Palamara[1,2], Christian Fuchsberger[4,5], Yakir A Reshef[6], Hilary K Finucane[1,7], Sebastian Schoenherr[8], Lukas Forer[8], Shane McCarthy[3], Goncalo R Abecasis[5], Richard Durbin[3] & Alkes L Price[1,2,9]

Po-Ru Loh

Alkes Price

Finding clones in the vast UK Biobank cohort

SNP data from 150k people

We identified >8,000 mosaic segmental mutations (at allelic fractions 1% and up)

Mosaic mutations cluster in genomic hotspots

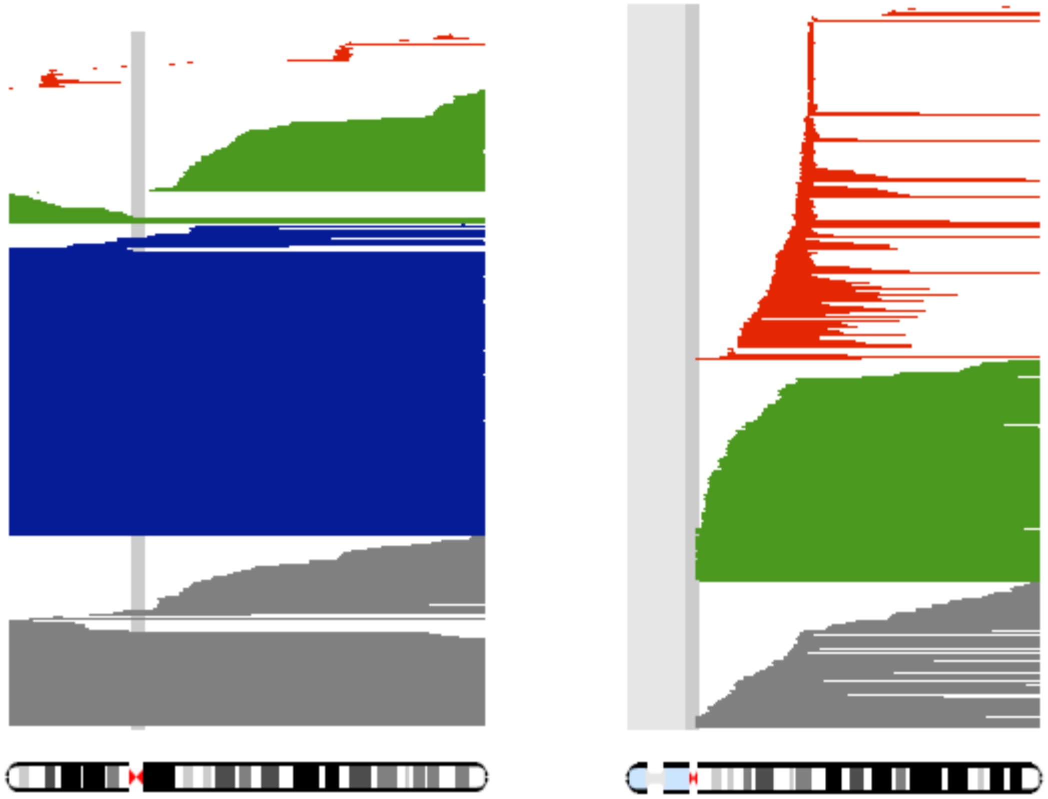
Pileups of detected events (indicated with horizontal lines):

— Loss

— CNN–LOH

— Gain

— Unknown



Acquired trisomy 12

Acquired 13q deletion

Chromosome 12

Chromosome 13

cf. Solimini et al., 2012 *Science*,
Jacobs et al. 2012 *Nat Genet,* Laurie et al. 2012 *Nat Genet*
Machiela et al. 2015 *AJHG,* Vattathil & Scheet 2016 *AJHG*

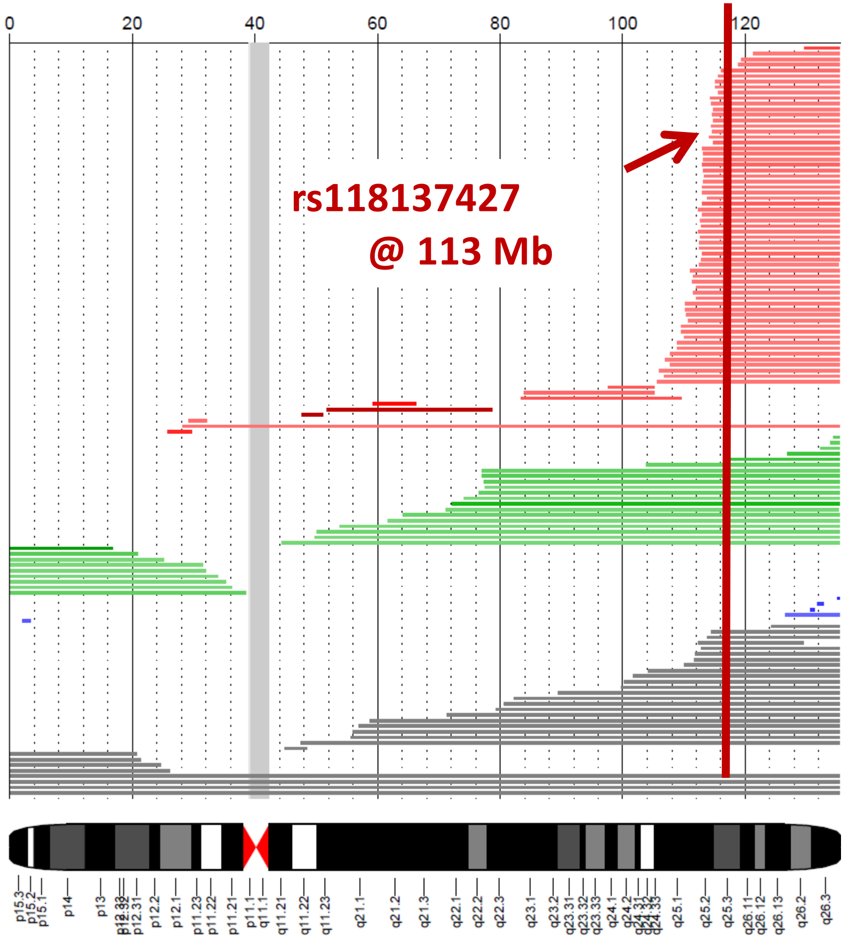
Can **inherited variation** shape our risk for developing clones with **specific somatic mutations** during our lifetime?

We treated recurring mutations (at the same locus) as a phenotype, then did a GWAS on each such phenotype (for each locus).

Power limited by modest number of “cases” (10s-100s for most loci)
Still, found cis- associations (and causal alleles) at many loci

- Causal variants are low-frequency (0.05 – 0.5%)
  but large odds ratios (19-700)

Example locus: chr10q deletions



rs118137427 @ 113 Mb

Observed in 60 people

All 60 have the minor allele of rs118137427 (MAF 5%) ($p < 10^{-41}$)

The haplotype with the minor allele is always the one that is deleted

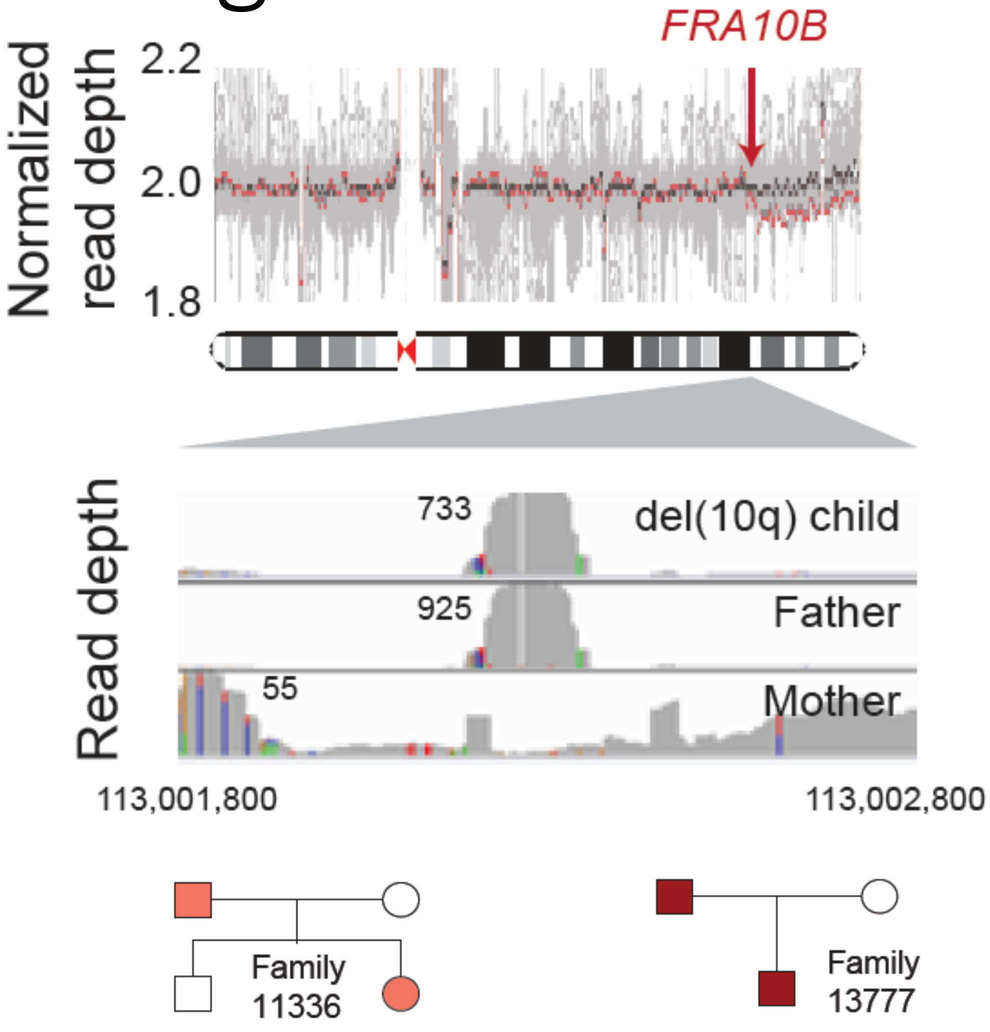
Loh, Genovese, Handsaker et al., submitted

Mechanism: An inherited allele makes a “fragile site” much more fragile

A genomic “fragile site” (VNTR) close to the deletion breakpoint

In WGS data, we see that people with acquired 10q deletions appear to have an **expanded FRA10B** site.

The mutation phenotype segregates in families, together with the repeat expansion



Loh, Genovese, Handsaker et al., submitted

A lesson from clonal hematopoiesis

- Dichotomy between *inheritance* (heritable, firm, predictable) and *acquired mutations* (capricious, random) is not as firm as we had thought

Lessons from multi-allelic variation and clonal hematopoiesis

- Interesting sources of unexplained heritability, but what was essential in studying them were **SNP data from vast numbers of people** and **new ways to think about old data types**
    - Imputation and IBD analyses become every more enabling and powerful as data sets expand; likely to allow many new kinds of analyses

Legacies of “missing heritability” mania

- Reminder that there is much to be learned
- Antidote to smugness
- Encourage exploration of new ideas

Legacies of “missing heritability” mania

- Reminder that there is much to be learned
- Antidote to smugness
- Encourage exploration of new ideas

• Too-easy excuse to abandon patient, consistent application of any one form of genetic analysis (SNP arrays ... WES ...) and lurch to applying new, glamorous expensive technology at small *n*

> Although we are finding sources of missing heritability,
> what enabled the above studies was large, widely available SNP data sets
> because • it is available **for so many people**
> • imputation and IBD become so much more powerful with sample size

Acknowledgements

Heritable influences on clonal expansions

Po-Ru Loh

Giulio Genovese

Bob Handsaker

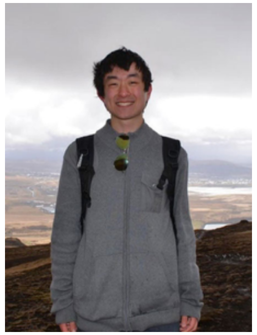
Alkes Price

Complex and multi-allelic genome variation

Aswin
Sekar

Katy
Tooley

Linda Boettger



Nolan
Kamitaki

Support: NGHRI (2 R01 HG006855)