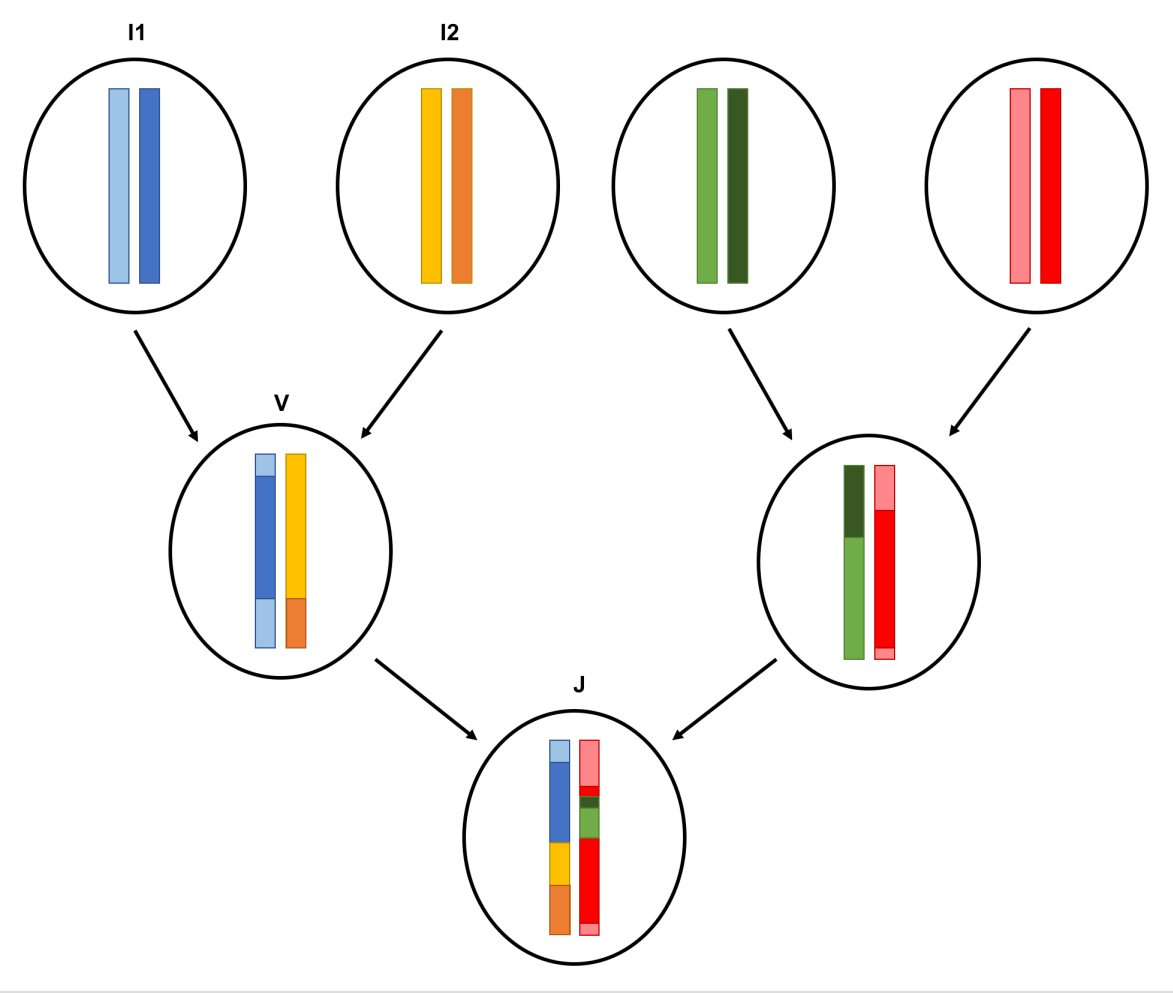
GWAS 7

Matti Pirinen
University of Helsinki
Week 4-5

GENOMIC RECOMBINATION



Sini Kerminen; www.helsinki.fi/~mjxpirin/stamp/

- Offspring inherits two genomes in continuous segments of the parent's two genomes
- Segments from recent ancestors are longer than from more distant ancestors
- This process determines correlations in GWAS data
  - Concept of genetic relatedness (last week's topic)
    - Sample-by-sample correlation
  - GWAS results at nearby SNPs are correlated (today's topic)
    - SNP-by-SNP correlation

HAPLOTYPES = HAPLOID GENOTYPES

1. True haplotypes

A	**T**	**C**

G	**C**	**A**

1. Individual has inherited a chromosome with alleles A-T-C from one parent and a chromosome with alleles G-C-A at the same SNPs from the other parent.

These are the two **haplotypes** of the individual at these 3 SNPs.

2. Observed (diploid) genotypes

A/G **C/T** **A/C**

2. Genotype data does not carry haplotype information for heterozygous loci:
We do not know whether **A** at SNP1 is coming from the same parent as **C** or as **T** at SNP2.

3. Possible haplotypes

A	C	A
A	C	C
A	T	A
A	T	C
G	C	A
G	C	C
G	T	A
G	T	C

3. **Haplotype phasing =** determining which are the two haplotypes behind the observed diploid genotypes

EXAMPLE: 2500BPS REGION FROM CHR 1

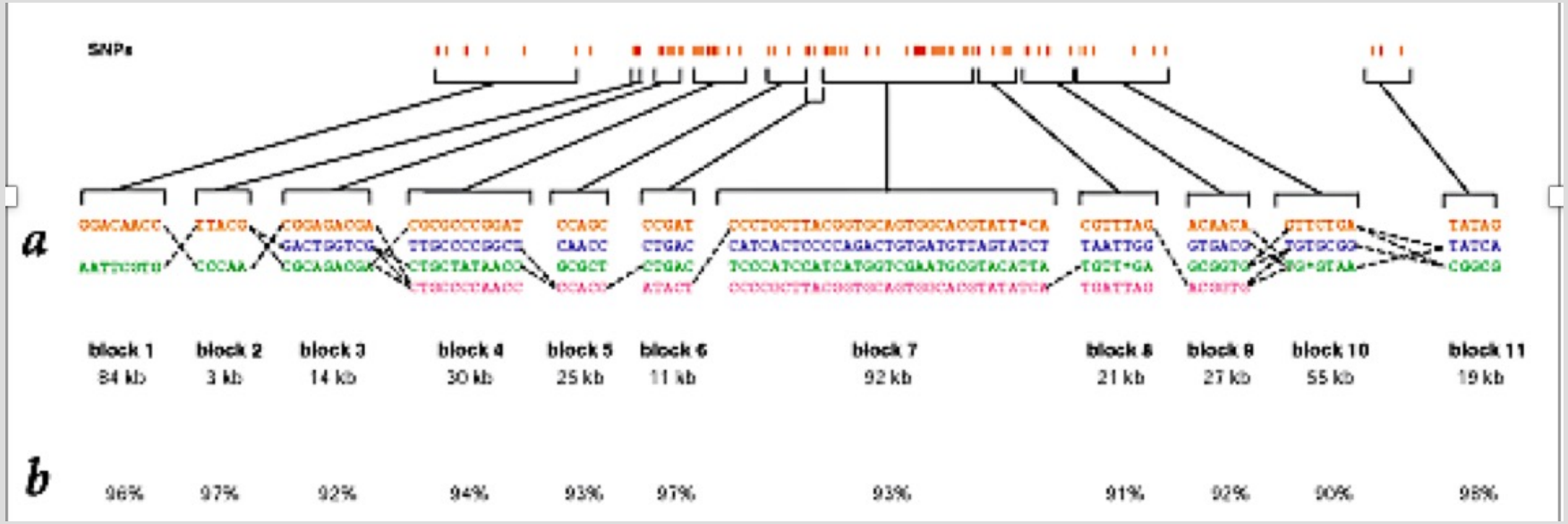
Southern Han Chinese

Finns

Luhya in Kenya

From: LDHap
https://ldlink.nci.nih.gov/?tab=ldhap

HAPLOTYPE BLOCK STRUCTURE



Haplotype is the sequence of alleles on the same chromosome, or, more generally, sequence of alleles inherited from the same parent.

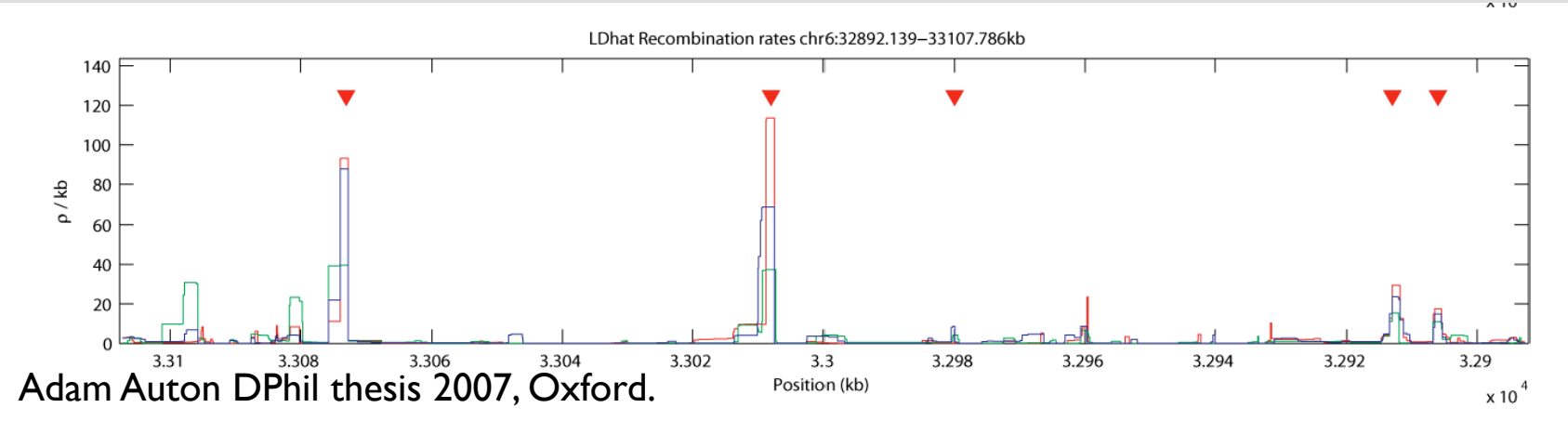
Daly et al. 2001. Nat Gen.

SNP variation in the population is organized as **haplotype blocks,**
where recombination seems to happen mainly between the blocks but little within any block.

Consequences for GWAS:

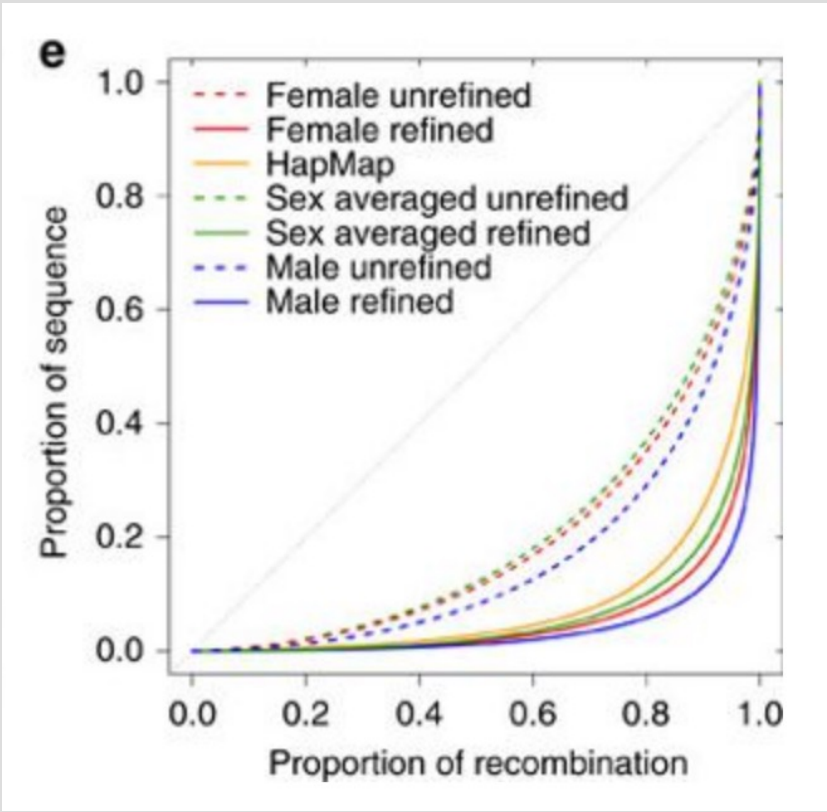
1. A relatively small number of tag-SNPs chosen from the blocks can capture most of (common) variation
2. It may be difficult to know which variant within a block is the causal variant as they are highly correlated

RECOMBINATION HOTSPOTS



Adam Auton DPhil thesis 2007, Oxford.

Recombination rate as function of physical position on MHC region shows 5 hotspots (triangles). Different colors are rates in different continental populations.



Bherer. et al. 2017 *Nature Communications* **8**, 14994

Most recombination events occur in a small proportion genome sequence.

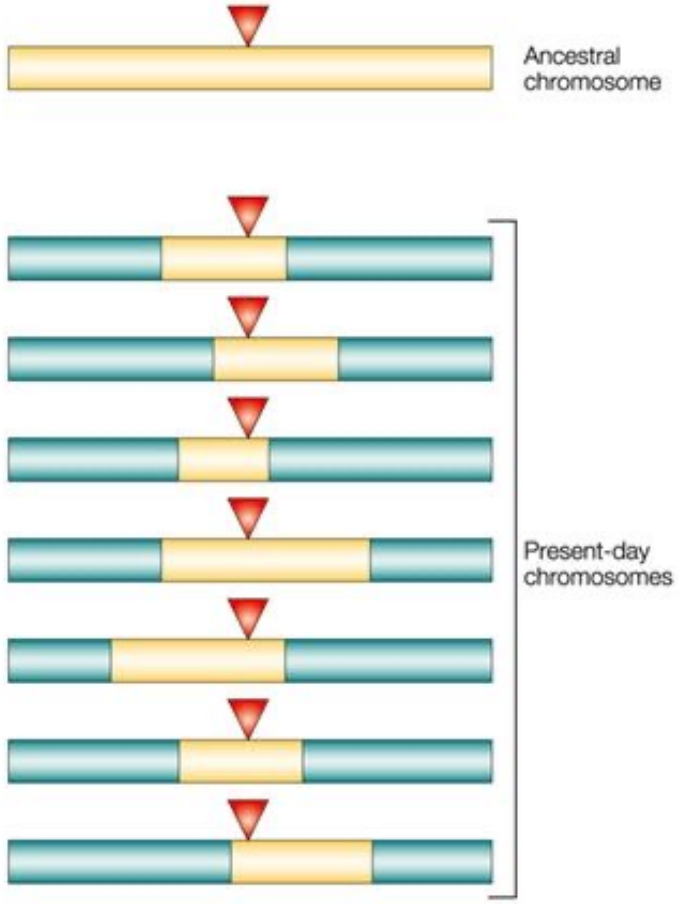
PHASING LEVERAGES SHARING OF HAPLOTYPE BLOCKS

Unphased genotypes	Possible phasing A		Possible phasing B		Possible phasing C		Possible phasing D	
A/C	A	C	A	C	A	C	A	C
G/T	G	T	G	T	T	G	T	G
A/T	A	T	T	A	A	T	T	A
Population haplotype frequency	55%	0%	15%	5%	2%	3%	0%	20%
Population frequency of unordered haplotype pair	0%		2 × (15% × 5%) = 1.5%		2 × (2% × 3%) = 0.12%		0%	
Posterior probability of unordered haplotype pair	0%		1.5% / (1.5% + 0.12%) = 93%		0.12% / (1.5% + 0.12%) = 7%		0%	

Browning & Browning. (2011) Nature Reviews Genetics **volume 12**: 703–714 Nature Reviews | Genetics

Consider one individual with a heterozygous genotype at each of three SNPs in a region. There are four possible haplotype configurations that are consistent with the genotype data (possible phasing patterns A–D). Suppose that haplotype frequencies are available from other individuals in the population at these sites (provided below each phasing pattern). The population frequency of a haplotype pair is obtained using the Hardy–Weinberg principle (independence of the two haplotypes within an individual); the factor of two in the frequency of the haplotype pairs accounts for both possible assignments of maternal and paternal origin to the two haplotypes. The posterior probabilities of the phased data are obtained from the population frequencies of the possible haplotype pairs. In this example, the posterior probability of phasing B (93%) is much greater than that of phasing C (7%).

EROSION OF HAPLOTYPE BACKGROUND OF A VARIANT

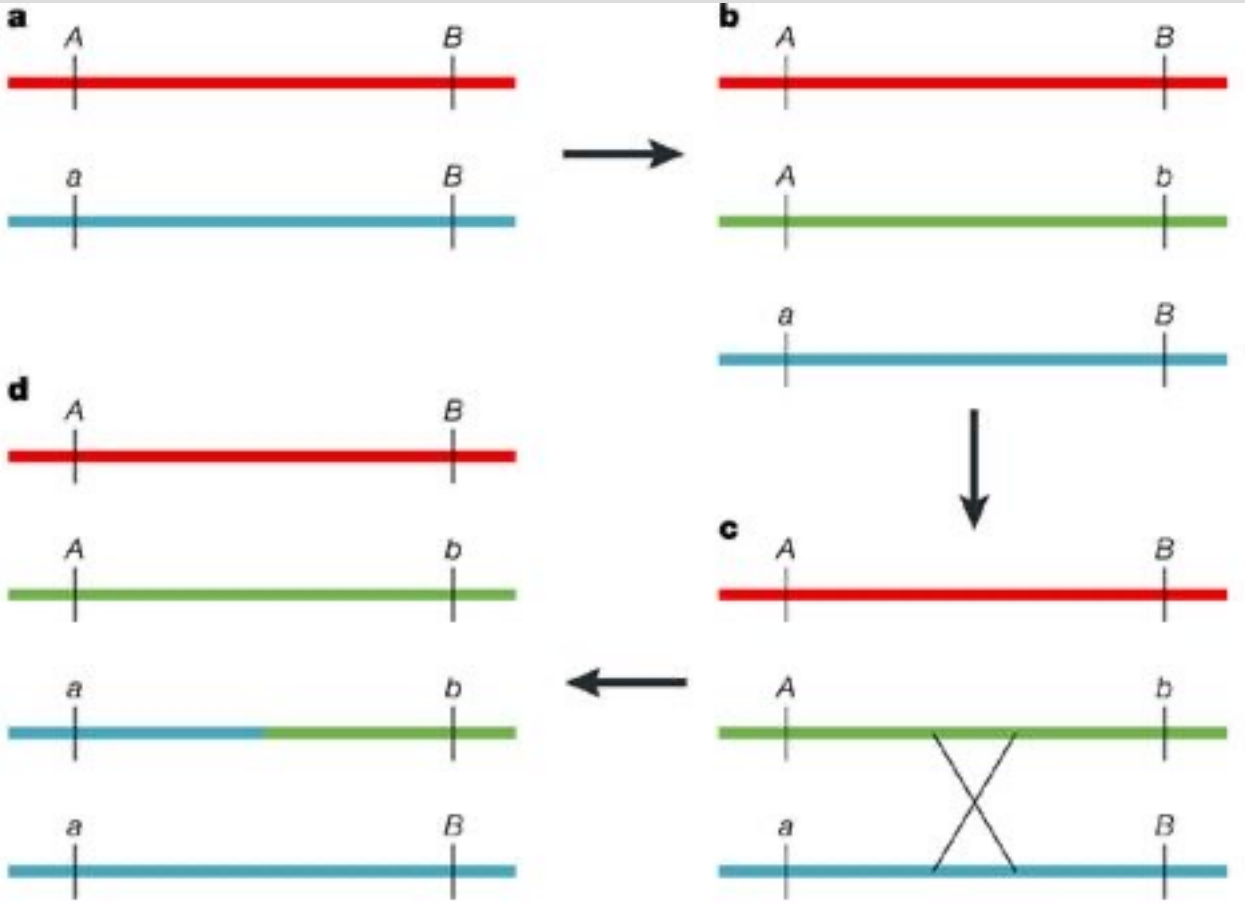


Ardlie et al. Nat Rev Gen 2002

The mutation is indicated by a red triangle. Chromosomal stretches derived from the common ancestor of all mutant chromosomes are shown in yellow, and new stretches introduced by recombination are shown in blue. Markers that are physically close (that is, in the yellow regions of present-day chromosomes) tend to remain associated with the ancestral mutation even as recombination limits the extent of the region of association over time.

Exactly shared segment around the variant is shrinking over time among the carriers of the variant.

LINKAGE DISEQUILIBRIUM



Ardlie et al. Nat Rev Gen 2002

a | At the outset, there is a polymorphic locus with alleles *A* and *a*. **b** | When a mutation occurs at a nearby locus, changing an allele *B* to *b*, this occurs on a single chromosome bearing either allele *A* or *a* at the first locus (*A* in this example). So, early in the lifetime of the mutation, only three out of the four possible haplotypes will be observed in the population. The *b* allele will always be found on a chromosome with the *A* allele at the adjacent locus. **c** | The association between alleles at the two loci will gradually be disrupted by recombination. **d** | This will result in the creation of the fourth possible haplotype and an eventual decline in LD among the markers in the population as the recombinant chromosome (*a*, *b*) increases in frequency.

LDPAIR

CEU
(Central Europe)

rs7837688 chr8:128539360	rs4242382 chr8:128517573 A	rs4242382 chr8:128517573 G		
G	0	180	180	(0.909)
T	16	2	18	(0.091)
	16 (0.081)	182 (0.919)	198	

Haplotypes

G_G: 180 (0.909)
T_A: 16 (0.081)
T_G: 2 (0.01)
G_A: 0 (0.0)

Statistics

D': 1.0
R^2: 0.8791
Chi-sq: 174.0659
p-value: <0.0001

rs7837688(G) allele is correlated with rs4242382(G) allele
rs7837688(T) allele is correlated with rs4242382(A) allele

LWK
(Kenya)

rs7837688 chr8:128539360	rs4242382 chr8:128517573 A	rs4242382 chr8:128517573 G		
G	40	139	179	(0.904)
T	5	14	19	(0.096)
	45 (0.227)	153 (0.773)	198	

Haplotypes

G_G: 139 (0.702)
G_A: 40 (0.202)
T_G: 14 (0.071)
T_A: 5 (0.025)

Statistics

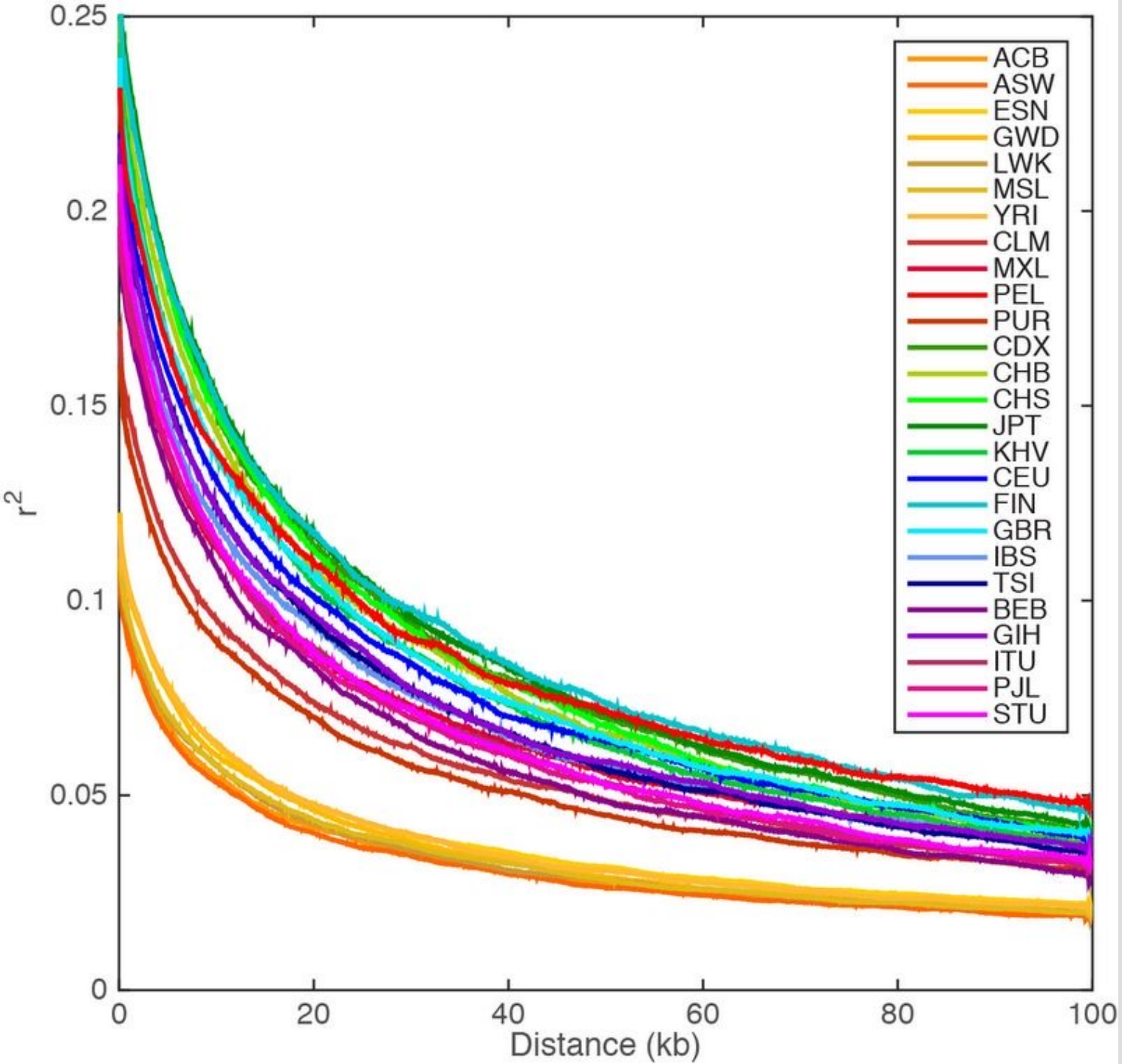
D': 0.0464
R^2: 0.0008
Chi-sq: 0.1541
p-value: 0.6946

rs7837688 and rs4242382 are in linkage equilibrium

D' is a normalized version of D that has maximum of 1.

From LDpair
https://ldlink.nci.nih.gov/

DECAY OF LD IN HUMANS



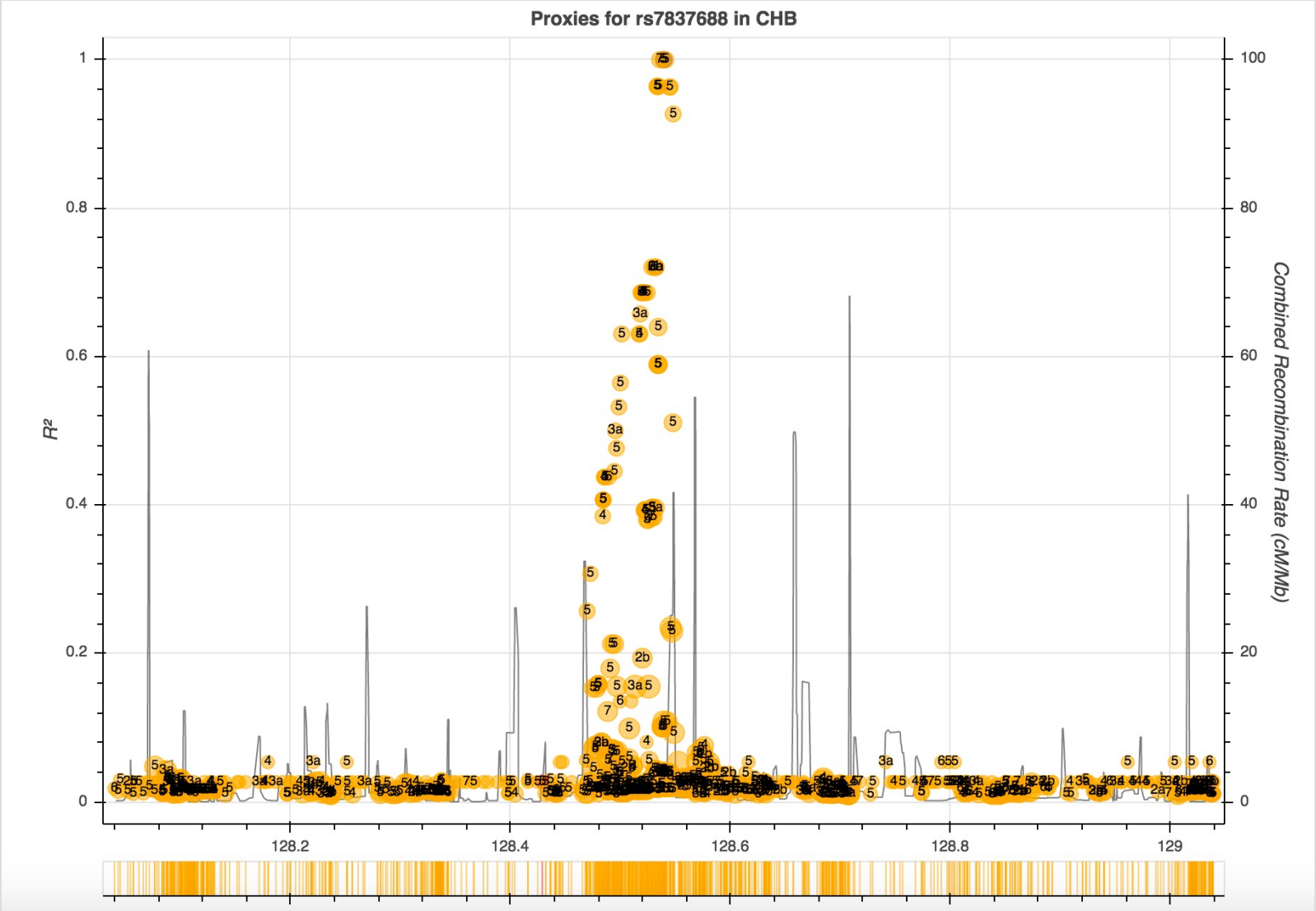
Linkage disequilibrium was calculated around 10,000 randomly selected polymorphic sites in each population, having first thinned each population down to the same sample size (61 individuals). The plotted line represents a 5 kb moving average.

Finns have the longest span of LD together with Asian populations followed by European populations.

African populations have clearly shorter span of LD.

1000 Genomes. Nature 2015.

LD FRIENDS



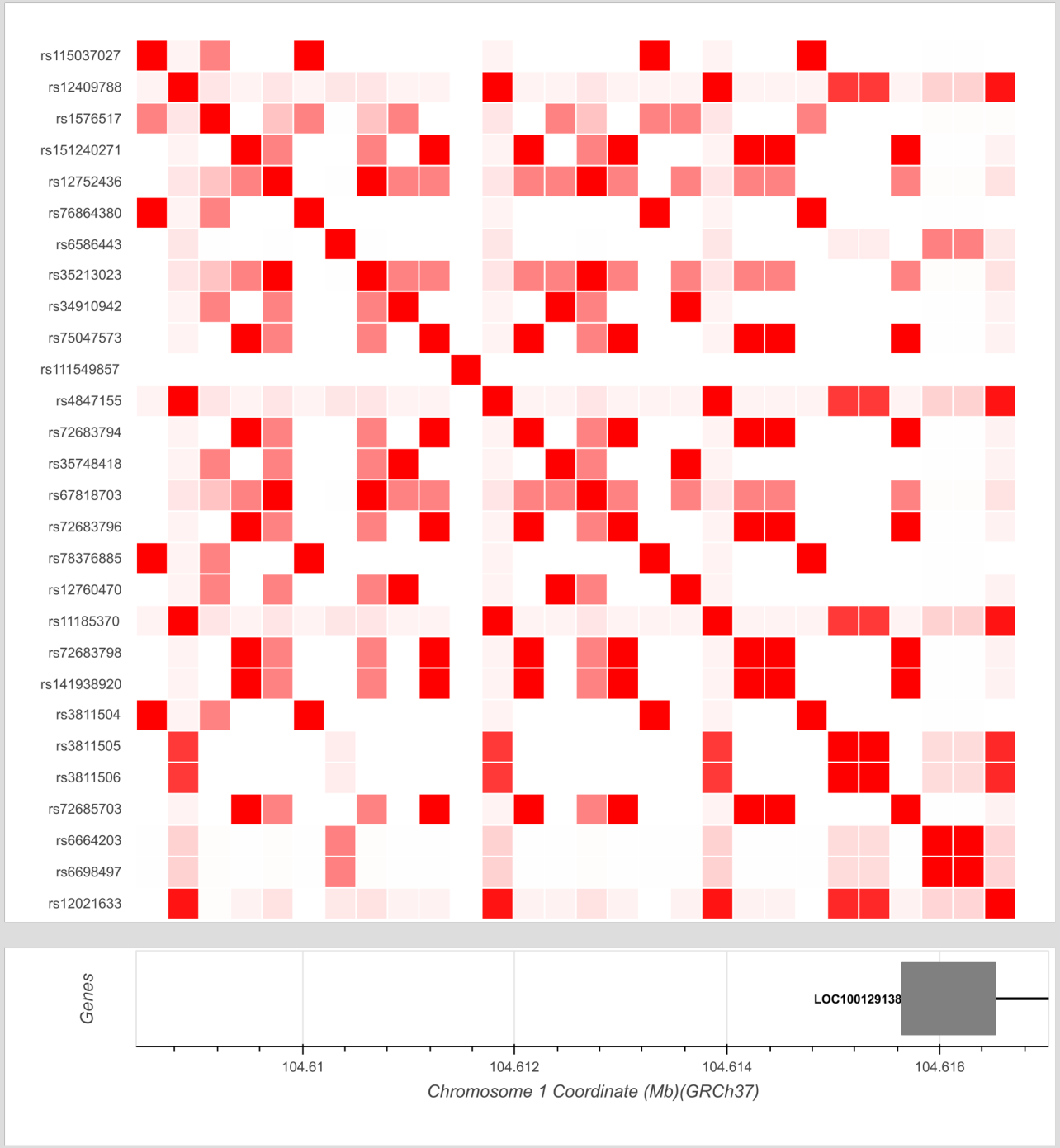
LDproxy gives all other variants that are highly correlated with the target SNP.

Note how r^2 decays after recombination hotspots on either side of the variant.

LDproxy reported 3 variants in perfect LD ($r^2=1$) with this SNP.

We call SNPs in high LD with each other as ”LD-friends”. Definition of ”high” can vary.

https://ldlink.nci.nih.gov/

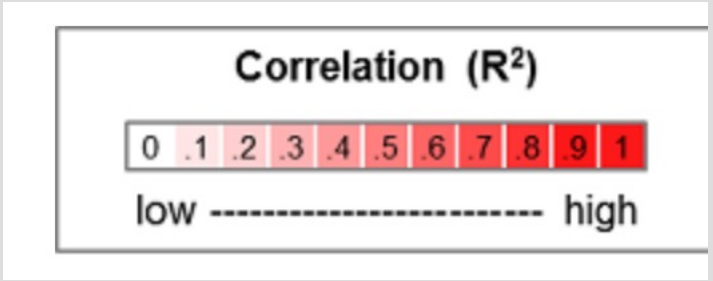


LDMATRIX

https://ldlink.nci.nih.gov/?tab=ldmatrix

Makes LD matrix for any 1000G population
For given set of variants.

Example:
CEU population, R^2 values,
a region from chr 1.



rs115037027
rs12409788
rs1576517
rs151240271
rs12752436
rs76864380
rs6586443
rs35213023
rs34910942
rs75047573
rs111549857
rs4847155
rs72683794
rs35748418
rs67818703
rs72683796
rs78376885
rs12760470
rs11185370
rs72683798
rs141938920
rs3811504
rs3811505
rs3811506
rs72685703
rs6664203
rs6698497
rs12021633

LD IN 4 GENES IN 2 FINNISH COHORTS

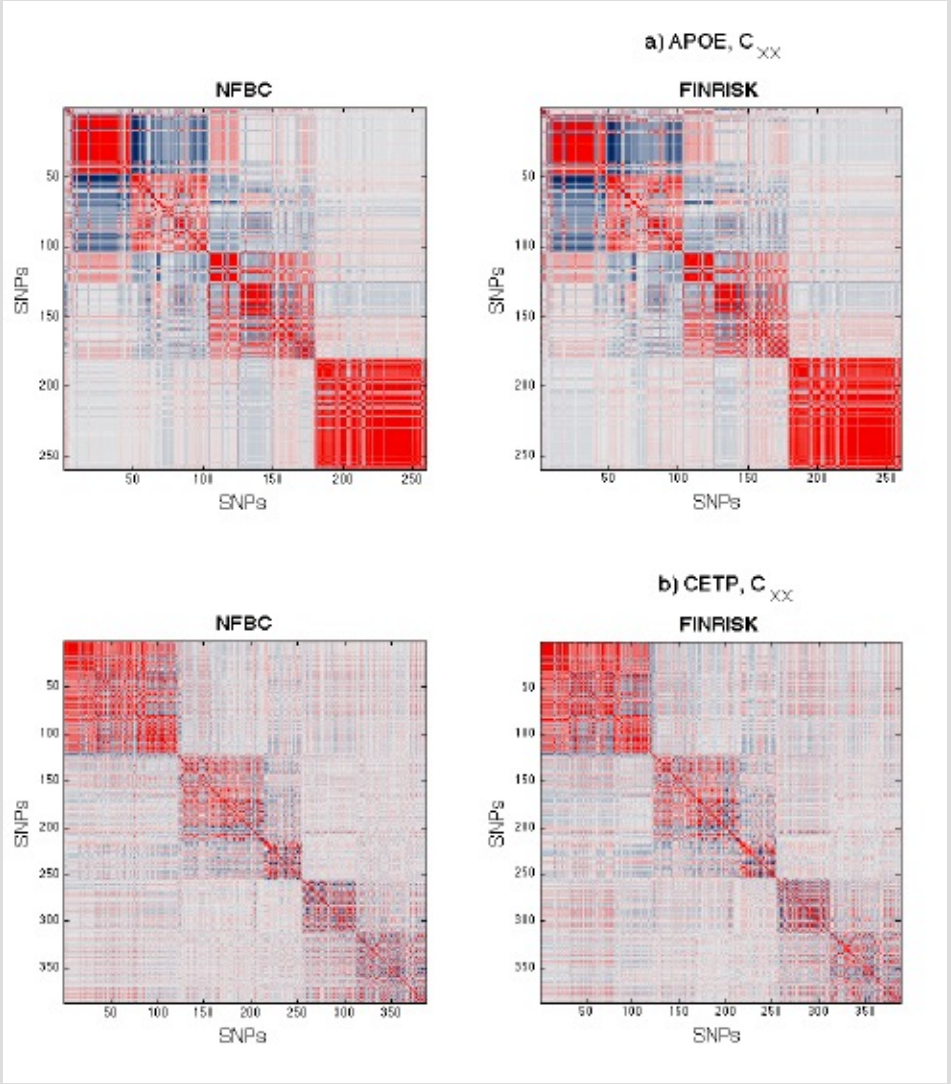
Genes:
APOE
CETP
GCKR
PCSK9

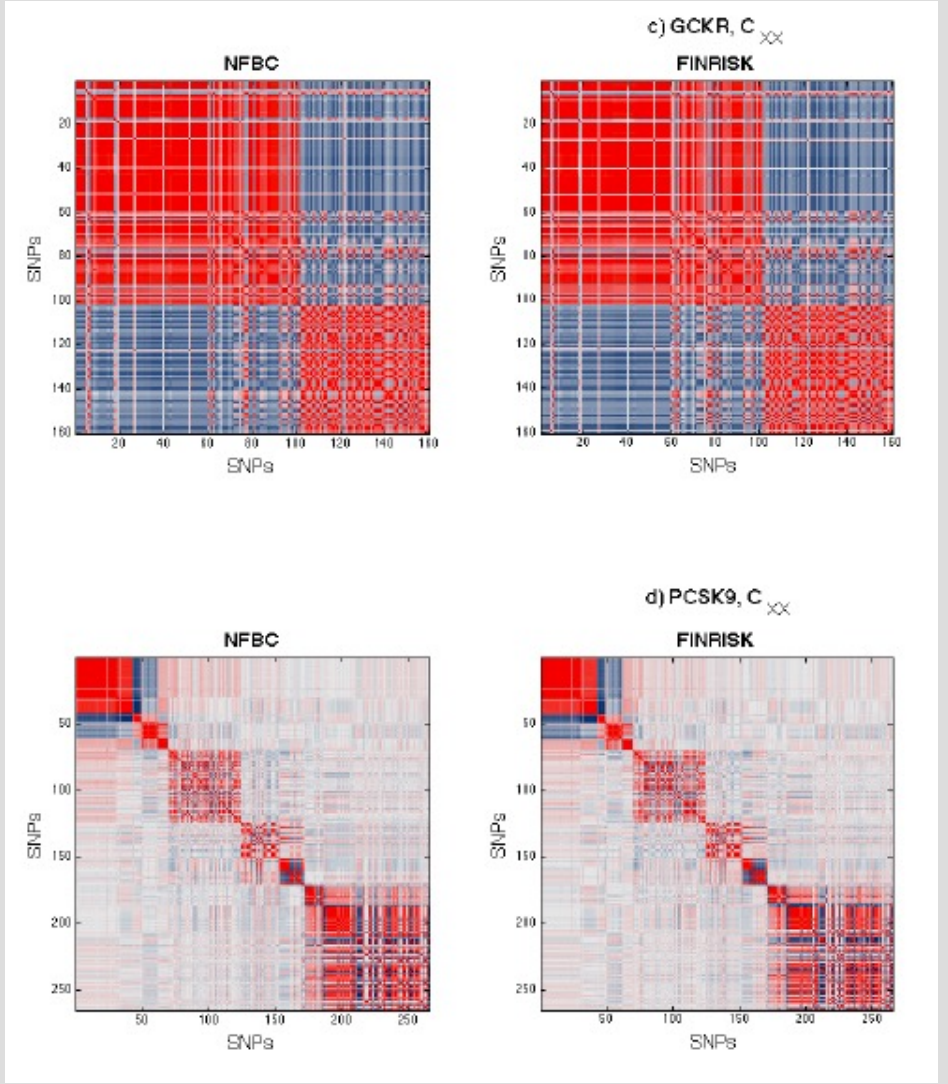
Cohorts:

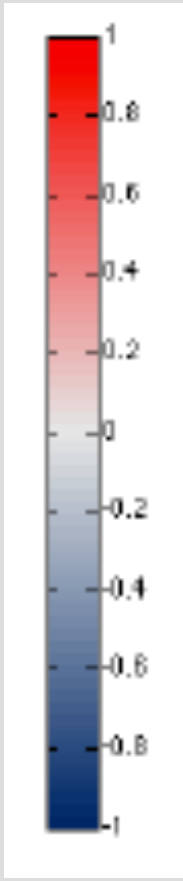
NFBC= Northern Finland Birth Cohort 1966 (n~ 4000)

FINRISK (n~3000)

Note how similar LD is across these two Finnish cohorts.

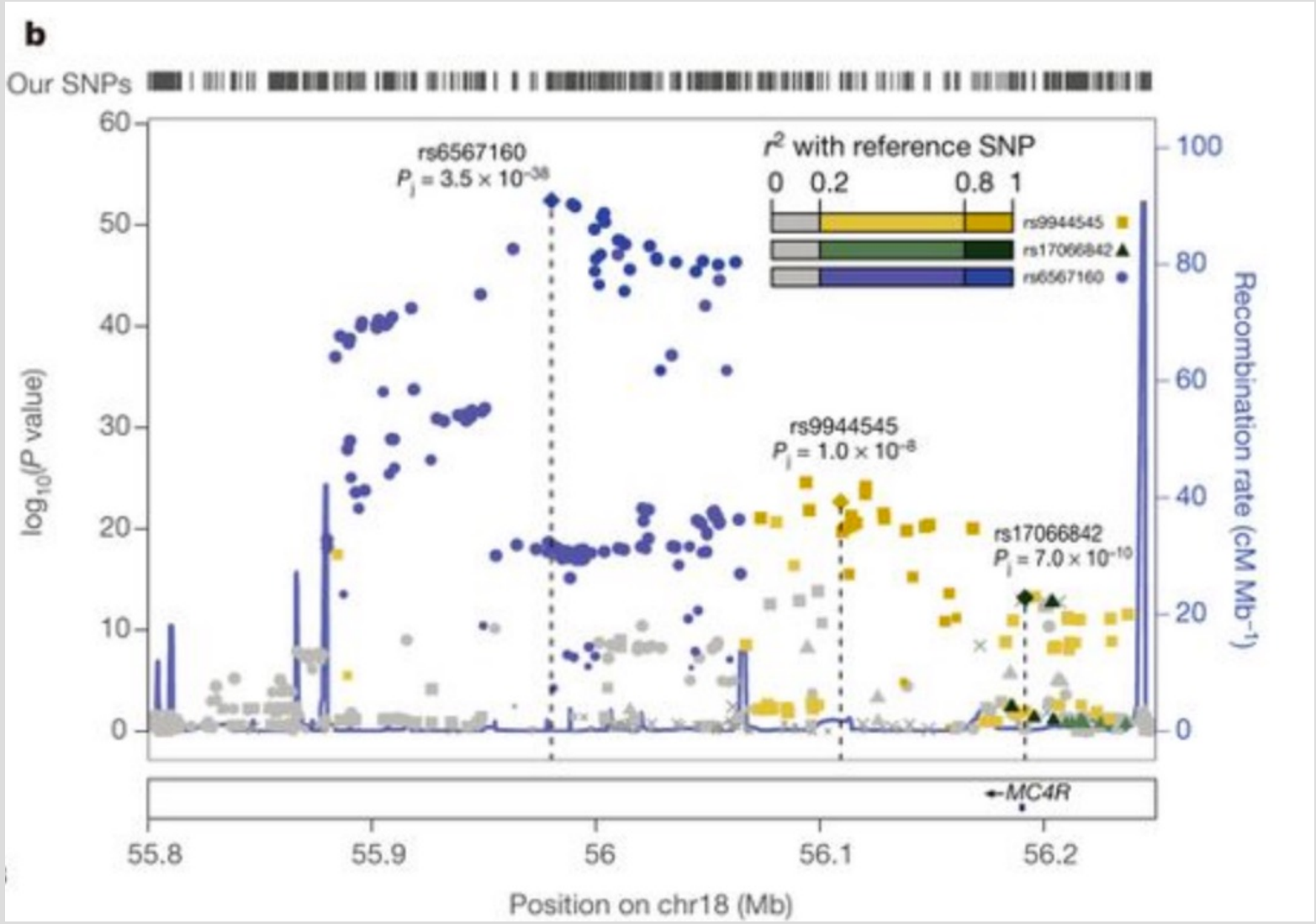






Cichonska et al. 2016 Bioinformatics

MC4R - BMI REGION (LOCKE ET AL. 2015 NATURE)

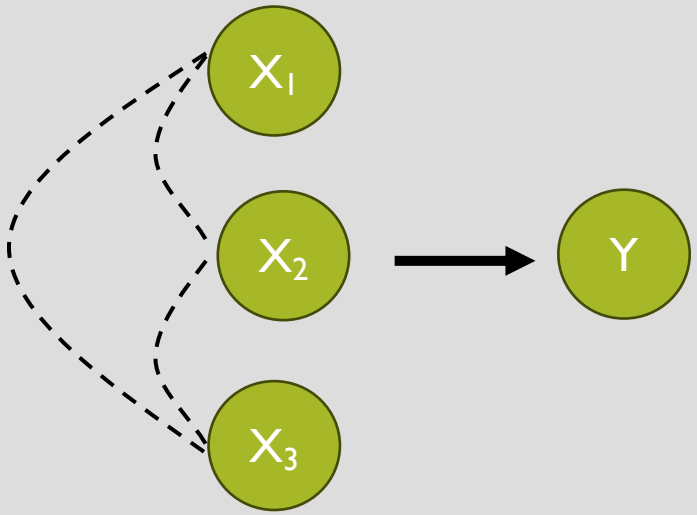


SNPs having similar P-values are in high LD with each other. Their genotype data look almost the same and therefore the regression model gives them essentially the same results.

It is not clear based on these marginal P-values alone which (if any of these) variants are the causal ones.

Three lead SNPs have been chosen and LD w.r.t them is shown for their LD-friends

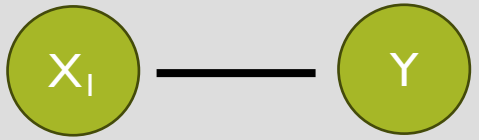
True model:
SNP 2 has a causal effect on Y ($\lambda_2 \neq 0$).
SNPs 1 and 3 are correlated with SNP 2 but do not have causal effects on Y.
($\lambda_1 = \lambda_3 = 0$)



We could estimate the causal effects by a joint model:

$$Y = \mu + X_1\lambda_1 + X_2\lambda_2 + X_3\lambda_3 + \varepsilon$$

In marginal models, that test only one SNP at a time, SNPs 1 and 3 are associated with Y
($\beta_1 \neq 0, \beta_3 \neq 0$).
Thus $\beta_1 \neq \lambda_1$ and $\beta_3 \neq \lambda_3$.
However, $\beta_2 = \lambda_2$ in this example.



Marginal model for SNP1

$$Y = \mu + X_1\beta_1 + \varepsilon$$

estimates marginal effect β_1
(not the causal effect λ_1).

MARGINAL EFFECT AT A NON-CAUSAL SNP

Consider SNPs N and C of which C is causal and its allele 1 has effect size λ
What is the marginal effect β at SNP N due to its LD with SNP C ?

Haplotype	0 – 0	1 – 0	0 – 1	1 - 1
Frequency	f_{00}	f_{10}	f_{01}	f_{11}
Effect	0	0	λ	λ

$$\beta = \mathrm{E}(Y \mid N = 1) - \mathrm{E}(Y \mid N = 0)$$
$$= \frac{f_{10}}{f_{10}+f_{11}} \times 0 + \frac{f_{11}}{f_{10}+f_{11}} \times \lambda - \frac{f_{00}}{f_{00}+f_{01}} \times 0 - \frac{f_{01}}{f_{00}+f_{01}} \times \lambda$$
$$= \left(\frac{f_{11}}{f_{10}+f_{11}} - \frac{f_{01}}{f_{00}+f_{01}}\right) \times \lambda = \left(\frac{f_{11}}{f_A} - \frac{f_{01}}{1-f_A}\right) \times \lambda$$
$$= \left(\frac{f_{11}(1-f_N) - f_{01}f_N}{f_N(1-f_N)}\right) \times \lambda = \frac{f_{11} - (f_{11}+f_{01})f_N}{f_N(1-f_N)} \times \lambda$$
$$= \frac{f_{11} - f_C f_N}{\sqrt{f_N(1-f_N)f_C(1-f_C)}} \sqrt{\frac{f_C(1-f_C)}{f_N(1-f_N)}} \times \lambda = r_{NC} \sqrt{\frac{f_C(1-f_C)}{f_N(1-f_N)}}\ \lambda$$

f_N = allele 1 frequency at SNP N
f_C = allele 1 frequency at SNP C
r_{NC} = correlation of allele 1 at N and C

Conclusion:
The marginal effect β at SNP N is shrunk towards 0 by correlation r_{NC} compared to SNP's C causal effect.
Also the allele frequencies affect the value.

MARGINAL EFFECT AT A NON-CAUSAL SNP

Marginal effect at SNP A is a linear combination of the causal effects of all variants in LD with A, where the weights are the correlations with A (after scaling the genotypes).

$$\beta_A^* = \lambda_A^* + r_{A1}\,\lambda_1^* + r_{A2}\lambda_2^* + r_{A3}\lambda_3^* + r_{A4}\lambda_4^* + \ldots$$

* denotes **scaled effect**: the allelic effect multiplied by $\sqrt{2f(1-f)}$, where f is the MAF of the SNP

MASKING EFFECT BY LD

		rs7687945 G	rs7687945 A		
rs356220	G	1 20.8%	1.16 (1.04-1.29) 41.9%	1.11	1.26 (1.16-1.37)
	A	1.31 (1.17-1.47) 27.9%	1.64 (1.41-1.90) 9.4%	1.39	
		1.18	1.25		
		1.07 (0.98-1.15)			

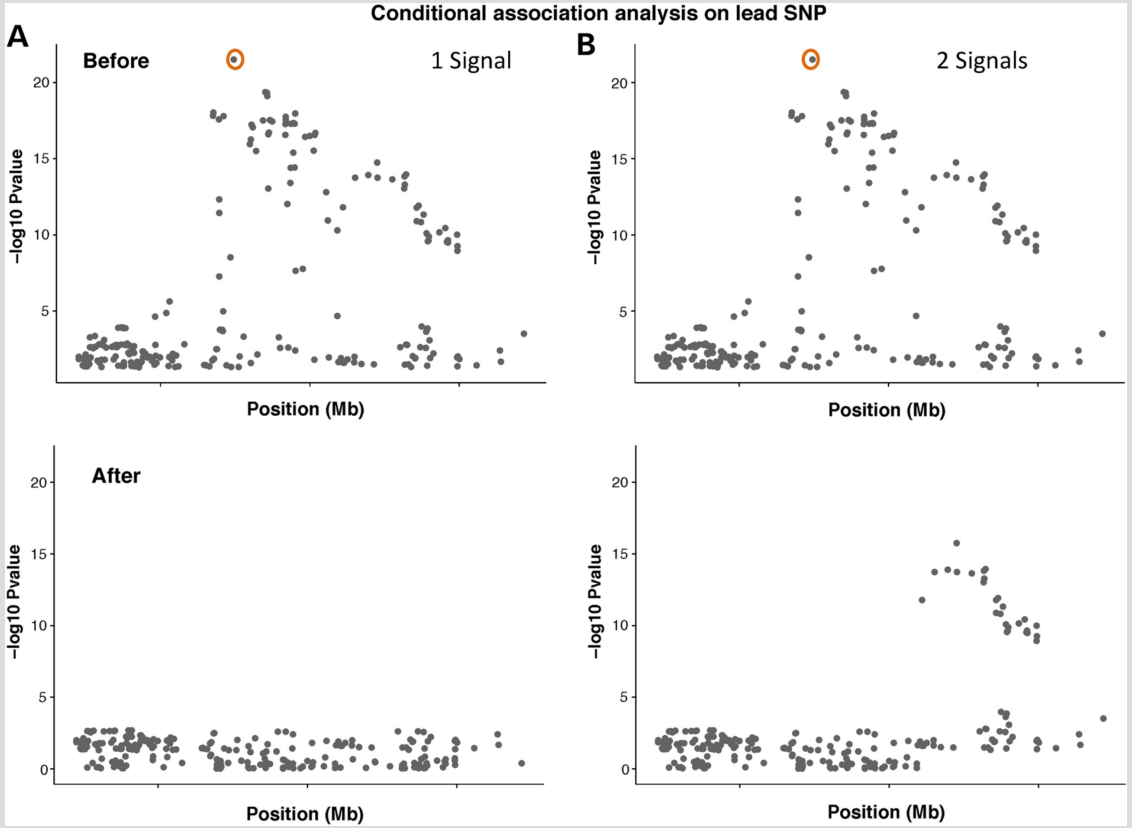
Viewing the data this way makes clear that the risk allele at the second SNP rs7687945 is more commonly found with the protective allele at rs356220 than would be expected were the SNPs in linkage equilibrium. As a result, the unconditional risk of the rs7687945 A allele (1.25) relative to the G allele (1.18) is 1.07 and not significantly different from 1.0

Shown in the centre of the table are estimates of odds ratio, 95% confidence limit (in brackets) and percentage frequency of the four haplotypes defined by the alleles at rs356220 and rs7687945. In the margins of the table is the risk of each of the alleles obtained by averaging the odds ratio of two haplotypes on which the allele can be found, weighting by the sample frequency. For example, the risk of carrying the G allele at rs356220 unconditional on the allele carried at rs7687945 is 1.11 (given in the top right) and is calculated as $(1 \times 20.8 + 1.16 \times 41.9)/(20.8 + 41.9)$.
By comparing the unconditional risks of the two alleles at each SNP, we recover the odds ratio estimated from a single SNP analysis.

UK Parkinson disease consortium & WTCCC2
Human Molecular Genetics,
Volume 20, Issue 2, 15 January 2011, Pages 345–353

STEPWISE FORWARD SEARCH

- Starts by conditioning on the lowest P-value
- Continues until no additional variant reaches pre-defined P-value threshold



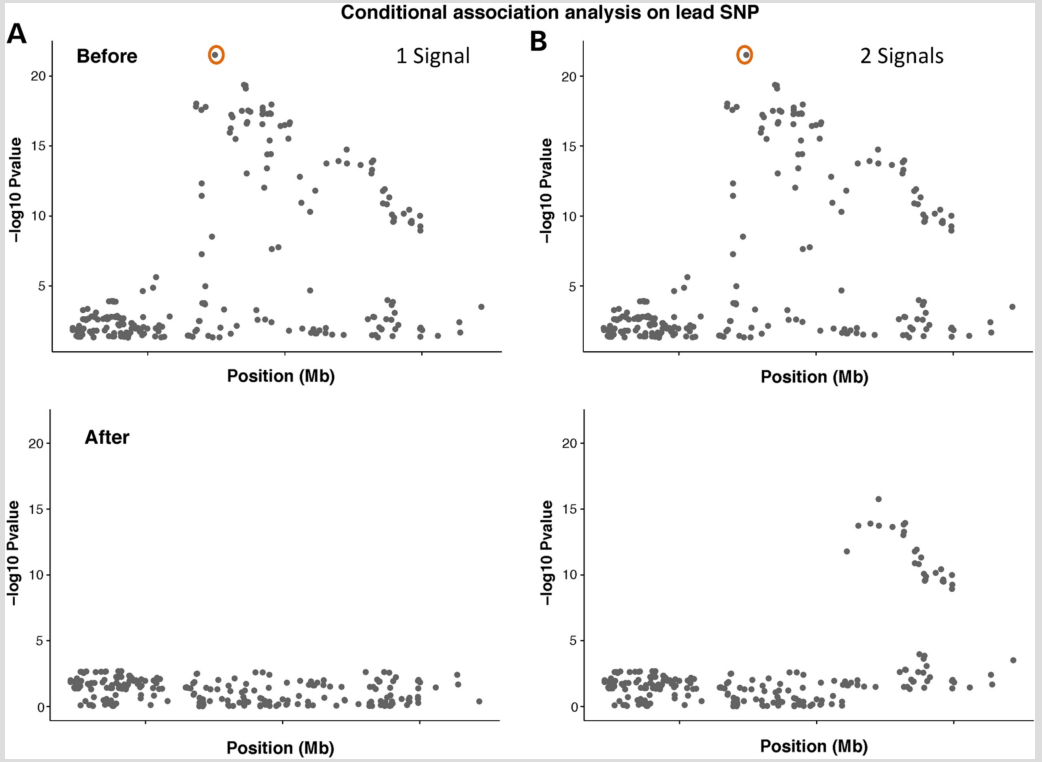
Spain & Barrett 2016

STEPWISE FORWARD SEARCH

+ Informs about multiple causal variants accounting for LD

− Does not necessarily find the optimal configuration

− Completely ignores the uncertainty of the causal configurations



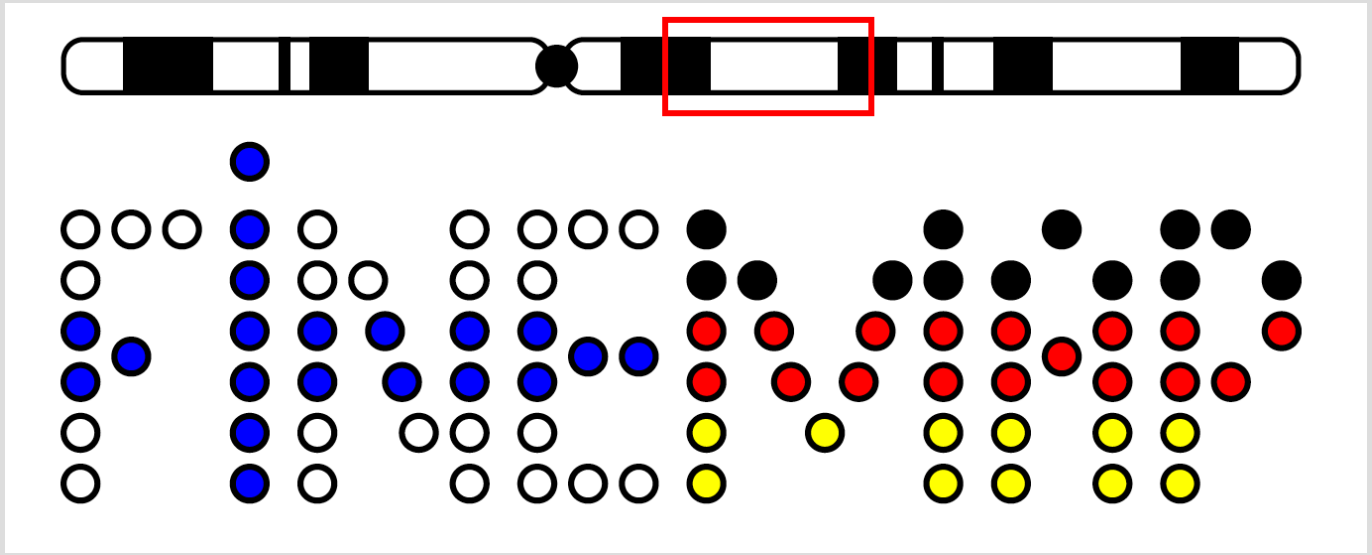
Spain & Barrett 2016

FINE-MAPPING

- Assign probabilities for each variant that the variant is (one of) the causal variant(s)
- Done in a Bayesian framework with prior assumptions
  - Prior probability of each variant being a causal variant?
    - Default: Each variant is equally likely to be causal
  - Prior distribution of the non-zero effects of the causal variants?
    - Default: $N(0, \tau^2)$ (See GWAS 4 for how to set τ^2)

FINE-MAPPING ASSUMING 1 CAUSAL VARIANT

- If there is exactly one causal variant in the region and it is among the genotyped variants, then the posterior probability of being causal is proportional to the single-SNP marginal Bayes factor of association (ABF from GWAS4)
- This idea can be extended to fine-mapping each independent signal of the region after we have conditioned on the other signals in the region when we have computed the GWAS statistics (betas and SEs) that are used in calculating ABFs
- For multiple causal variants, we use methods such as FINEMAP or SuSiE



Bioinformatics, 32(10), 2016, 1493–1501
doi: 10.1093/bioinformatics/btw018
Advance Access Publication Date: 14 January 2016
Original Paper

Genetics and population analysis

FINEMAP: efficient variable selection using summary data from genome-wide association studies

Christian Benner[1,2]*, Chris C.A. Spencer[3], Aki S. Havulinna[4], Veikko Salomaa[4], Samuli Ripatti[1,2,5] and Matti Pirinen[1]*

Christian Benner

BAYESIAN MODEL FOR FINE-MAPPING

- Define causal configuration γ as a binary vector for variants

$\gamma =$ 

- This configuration says that variants 3 and 7 are causal and the others have no effects.

BAYESIAN MODEL FOR FINE-MAPPING

- Define causal configuration γ as a binary vector for variants

$$\gamma = \begin{array}{|c|c|c|c|c|c|c|c|} \hline 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ \hline \end{array}$$

- In total there are 2^p configurations on p variants, but we will assume that only **sparse** configurations are plausible, say those with < 10 causal variants per a region.
- Ultimate goal is to compute probability for each configuration, given the observed GWAS data in the region.

BAYESIAN MODEL FOR FINE-MAPPING

- Define causal configuration γ as a binary vector for variants

- Each causal variant picks its effect from $N(0, s^2)$

$$p(\boldsymbol{\lambda}|\boldsymbol{\gamma}) = \mathcal{N}\left(\mathbf{0}, s^2\boldsymbol{\Delta}_{\boldsymbol{\gamma}}\right)$$

Causal configuration $\boldsymbol{\gamma}$

1	0	0	0	0	0	0	0	0	0

$$\boldsymbol{\Delta}_{\boldsymbol{\gamma}} = \text{diag}(\boldsymbol{\gamma}) = \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}$$

BAYESIAN MODEL FOR FINE-MAPPING

- Define causal configuration γ as a binary vector for variants
- Each causal variant picks its effect from $N(0, s^2)$
- For each configuration compute the Bayes factor (BF), i.e., how well the configuration explains the data relative to the null model

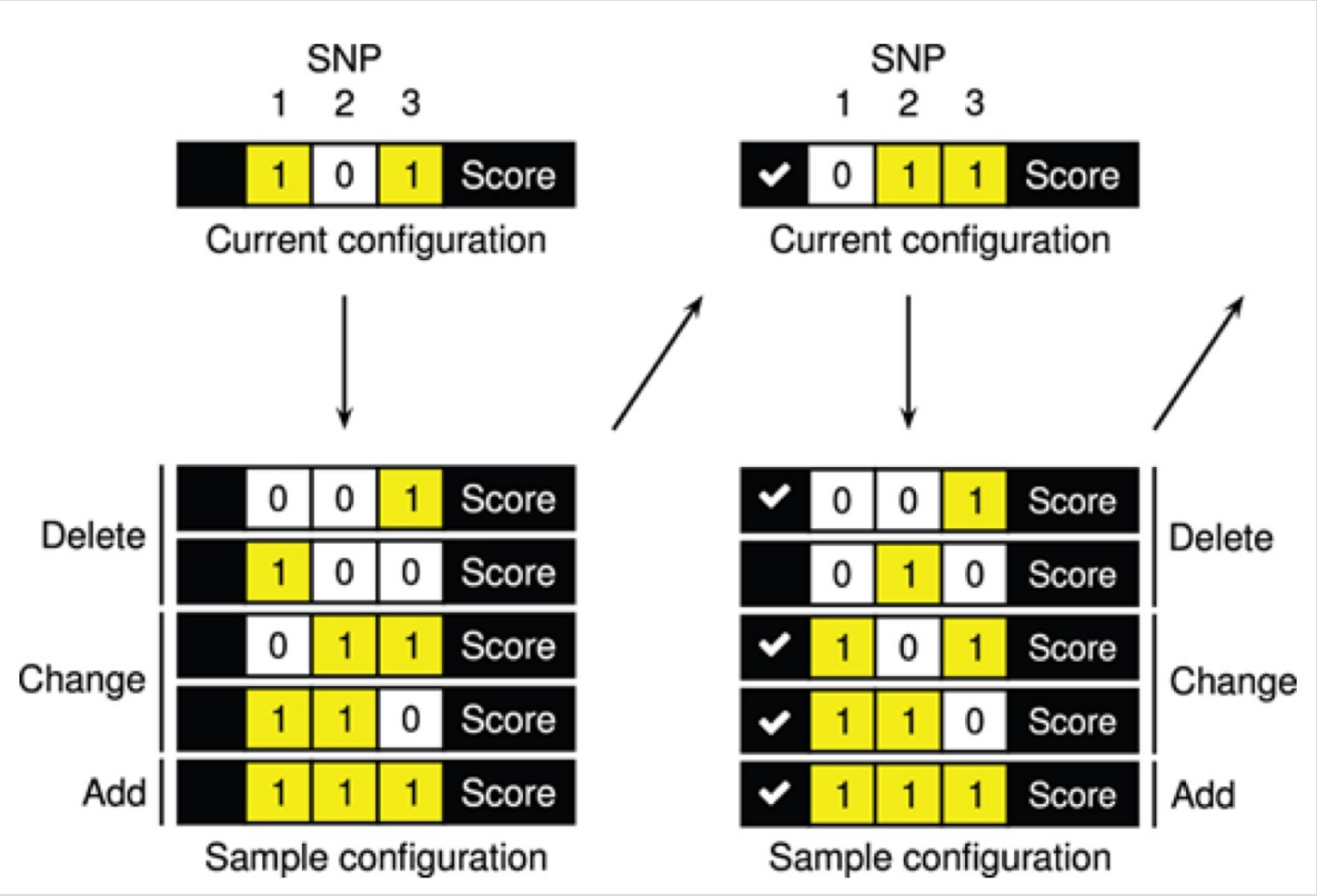
$$\mathrm{BF}_\gamma = \frac{P(\mathrm{DATA}|\gamma)}{P(\mathrm{DATA}|\mathrm{NULL})} = \frac{\mathcal{N}(\boldsymbol{z}|0, \mathbf{R} + s^2 \mathbf{R}\, \boldsymbol{\Delta}_\gamma\, \mathbf{R})}{\mathcal{N}(\boldsymbol{z}|0, \mathbf{R})}$$

BAYESIAN MODEL FOR FINE-MAPPING

- Define causal configuration γ as a binary vector for variants
- Each causal variant picks its effect from $N(0, s^2)$
- For each configuration compute the Bayes factor (BF), i.e., how well the configuration explains the data relative to the null model
- By combining BFs with prior probabilities of configurations we get the posterior probabilities

$$p_\gamma = P(\gamma|\mathrm{DATA}) \propto \mathrm{prior}_\gamma \times \mathrm{BF}_\gamma$$

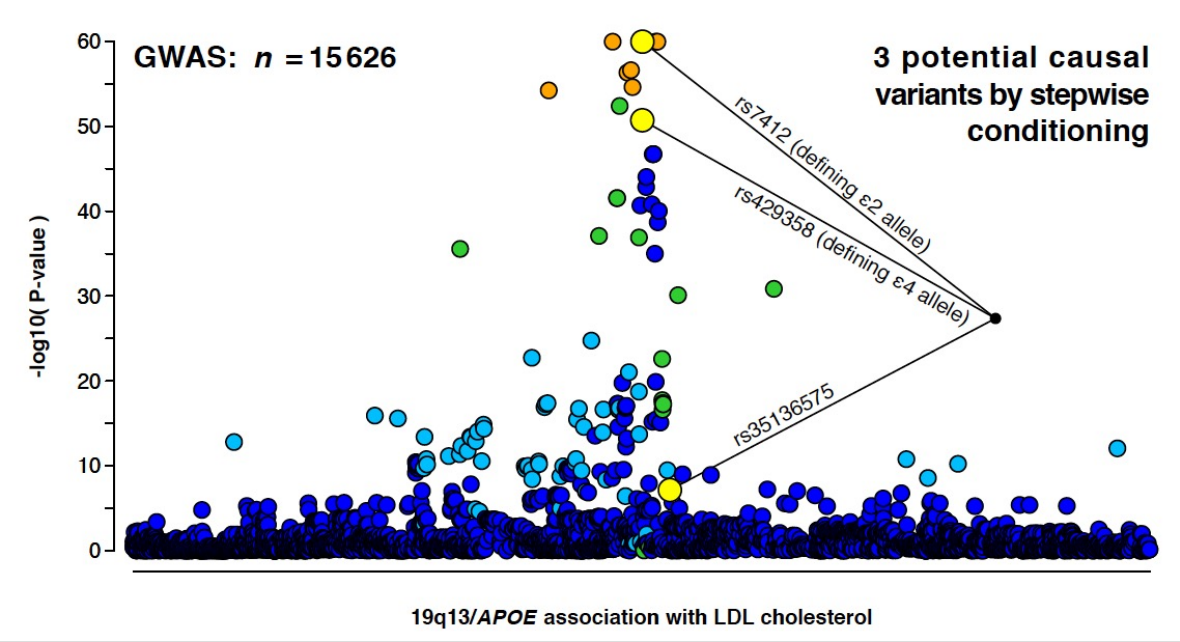
FINEMAP ALGORITHM



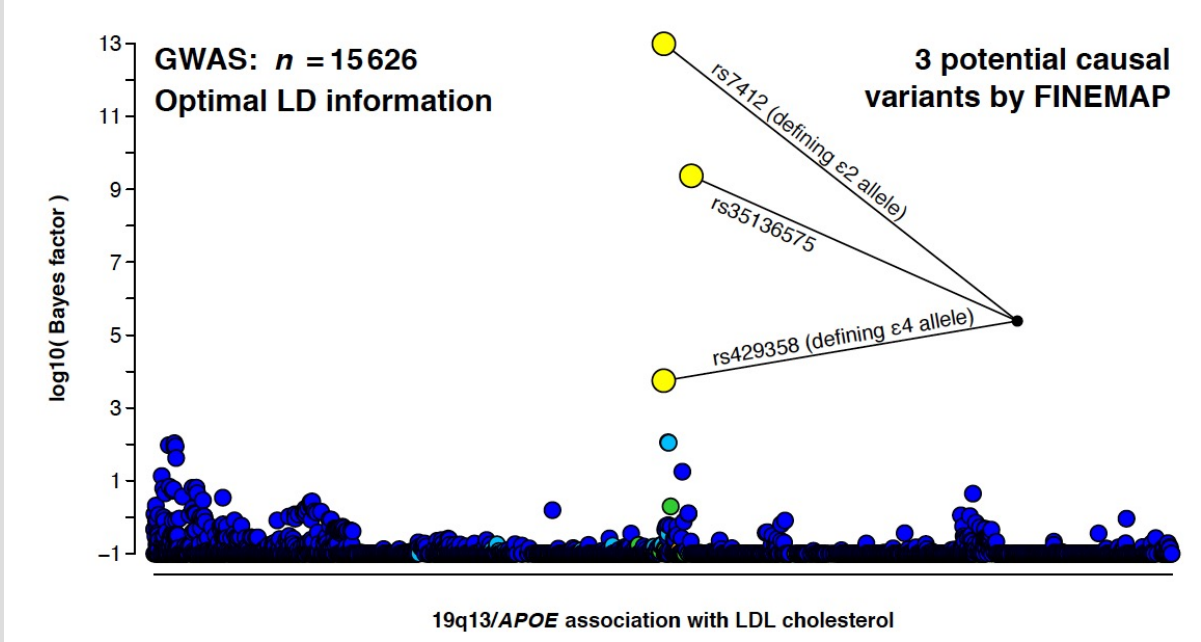
- Collect configurations from a high probability region using Shotgun stochastic search (Hans et al. 2007)
  - Memorize BFs of all those configurations seen during the search
  - Stop once not much new probability mass is found
  - Renormalize posteriors with respect to the configurations visited

Benner et al. 2016

FINEMAP RESULTS

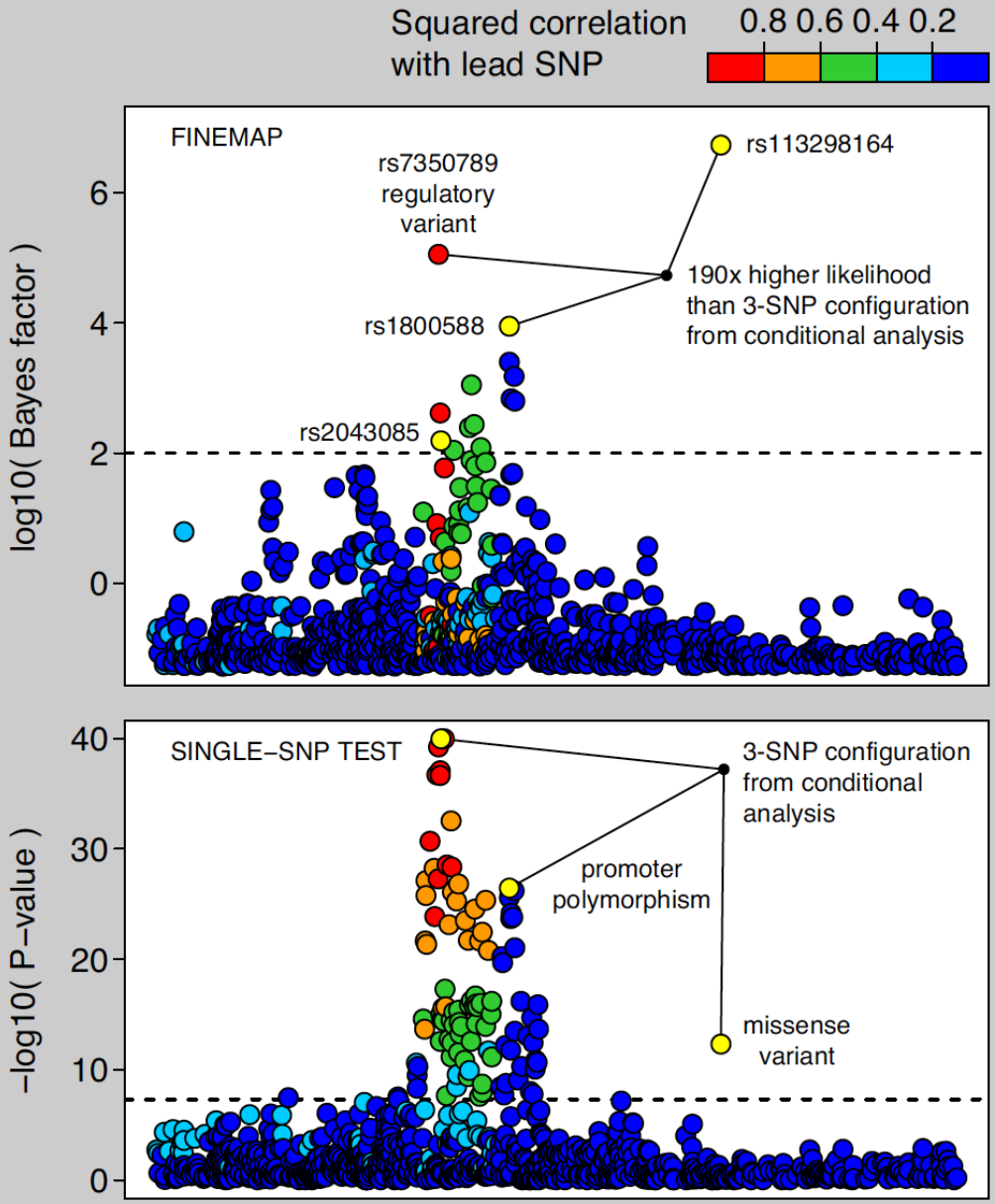


GWAS



FINEMAP

15q21/*LIPC* association with HDL cholesterol



Christian Benner

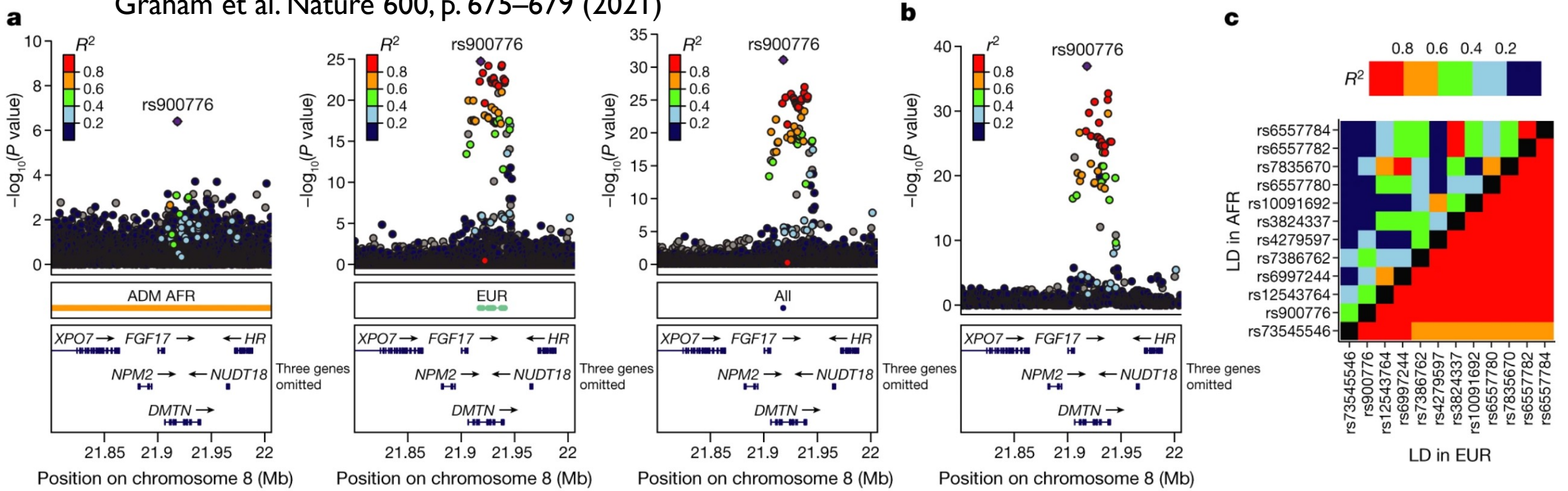
Surakka et al.
Nat. Genet. 2015

NATIONAL INSTITUTE FOR HEALTH AND WELFARE
FINLAND

FINRISK STUDY
20000 indivduals

Inclusion of multiple ancestries drives improved fine-mapping of LDL-C levels in *DMTN* locus.

Graham et al. Nature 600, p. 675–679 (2021)



Association of the *DMTN* intron variant rs900776 with LDL-Cholesterol in the admixed African, European, or multi-ancestry meta-analysis (**a**) or with *DMTN* gene expression (**b**).
The region spanned by the 99% credible sets (assuming a single causal variant) are shown in the centre box. The LDL-C association signal significantly colocalizes with the expression signal of *DMTN* in liver.
c, The LD patterns for variants in the European ancestry 99% credible set differ greatly between African (AFR) and European ancestry individuals in 1000 Genomes. The lead variant has a posterior probability of 0.86 in the admixed African analysis, 0.51 in the European analysis and >0.99 in the multi-ancestry analysis.