

Fine-mapping, colocalization, & a bit more on pop gen and rare variation/sequencing

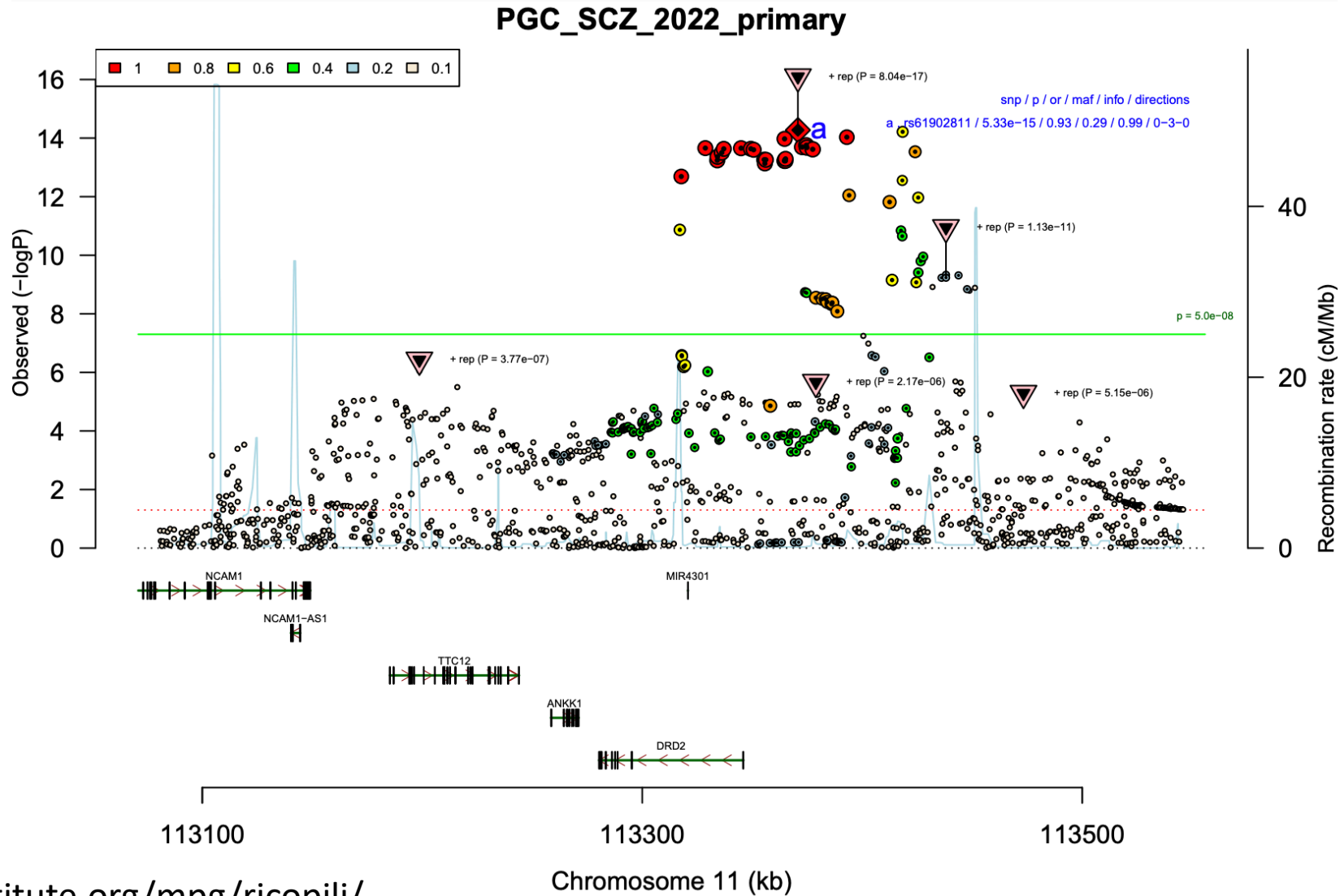
Ben Neale

March 10th

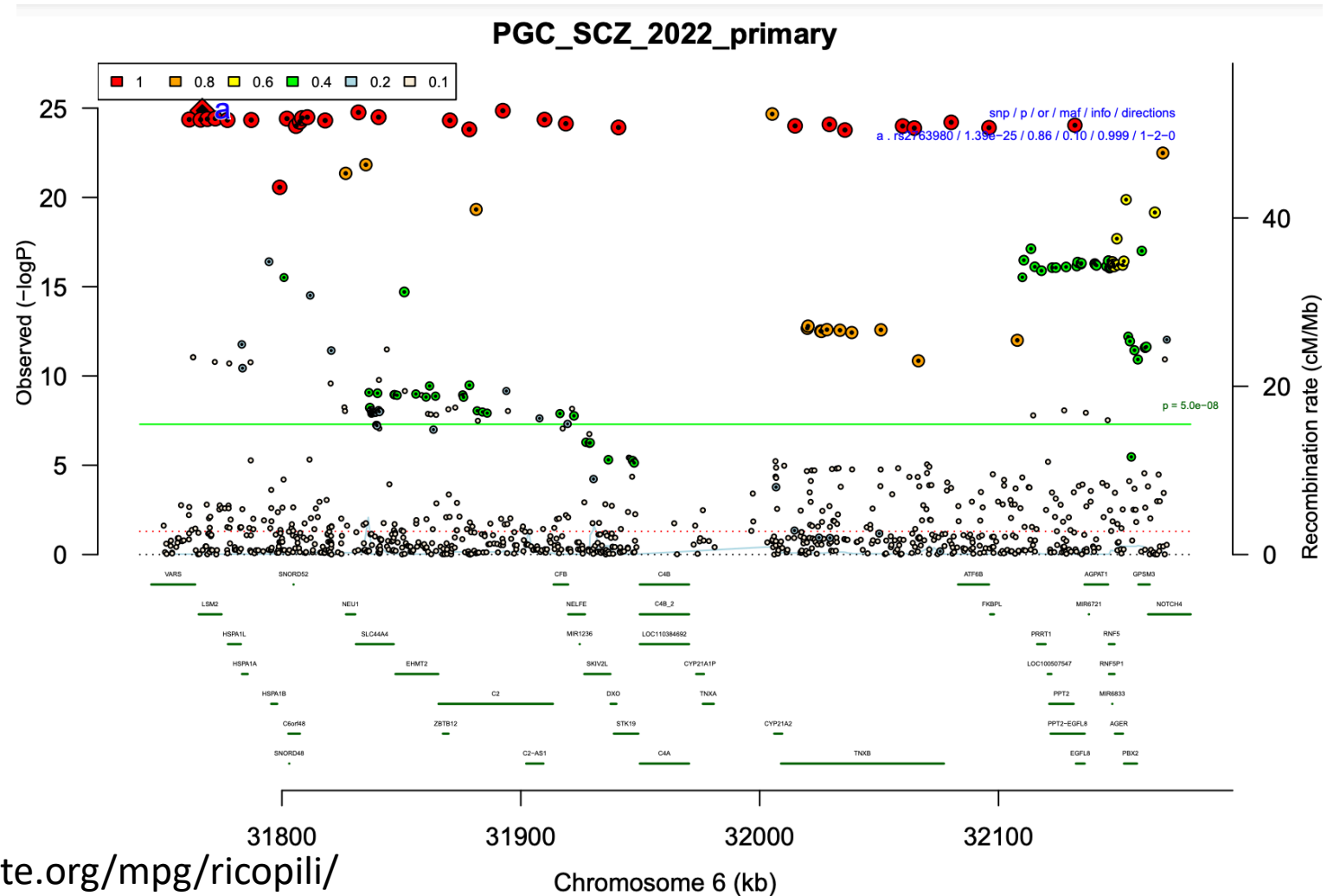
So you've done a GWAS and found some associations – what now?

- Some questions to contemplate:
- What are the causal variant/variants driving the signal?
- What are the proximal biological consequences of those base pair differences?
- What are genes, cells and biological processes that are perturbed by these differences?
- How do those perturbations influence physiology?
- Or as Lindon used to say “What does it all **mean**”

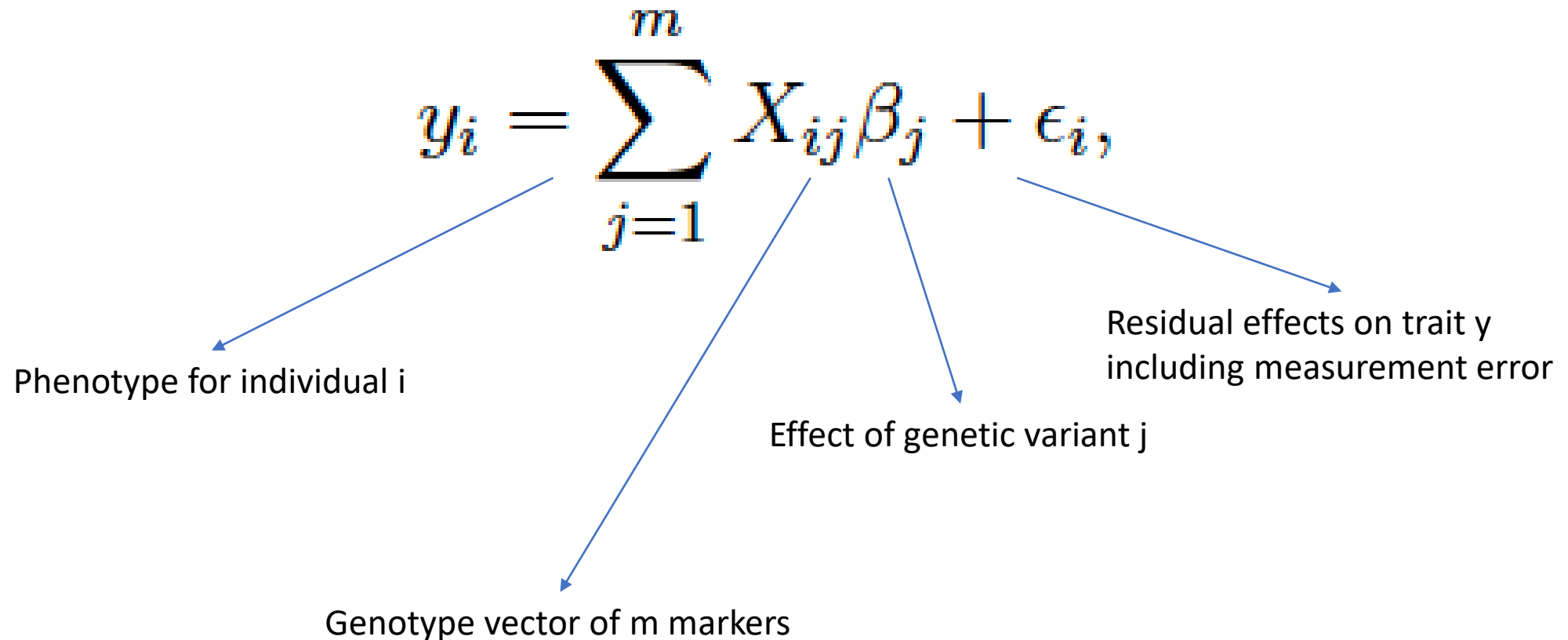
A typical locus zoom plot



A more complicated locus zoom plot



Statistical fine-mapping - some modeling considerations



We can also consider the binary case

$$\log \frac{p(y_i = 1)}{p(y_i = 0)} = \sum_{j=1}^m X_{ij} \beta_j + \alpha_i$$

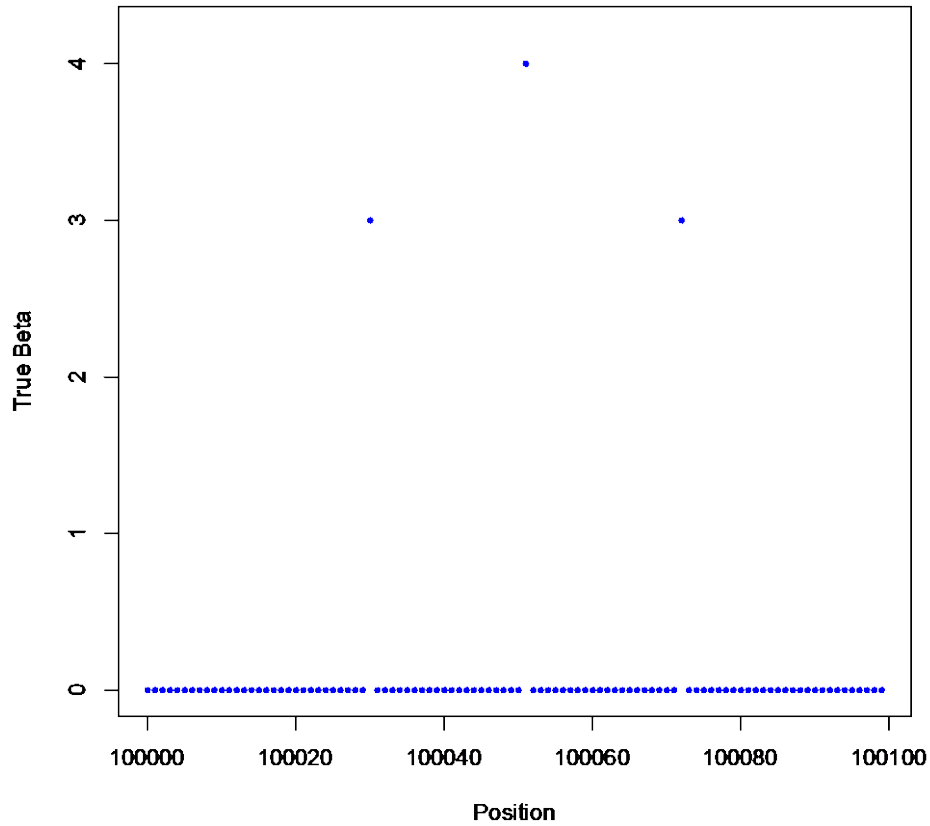
Phenotype y coded as 0/1 for individual i

Genotype vector of m markers

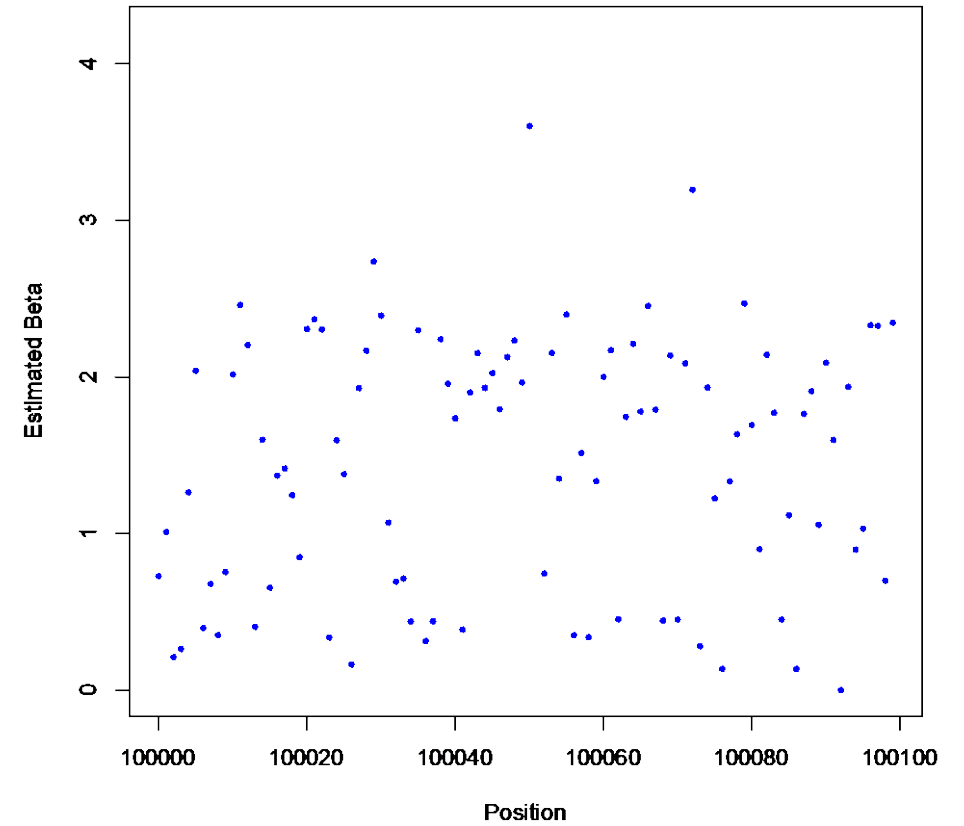
Effect of genetic variant j

Residual effects on trait y
including measurement error

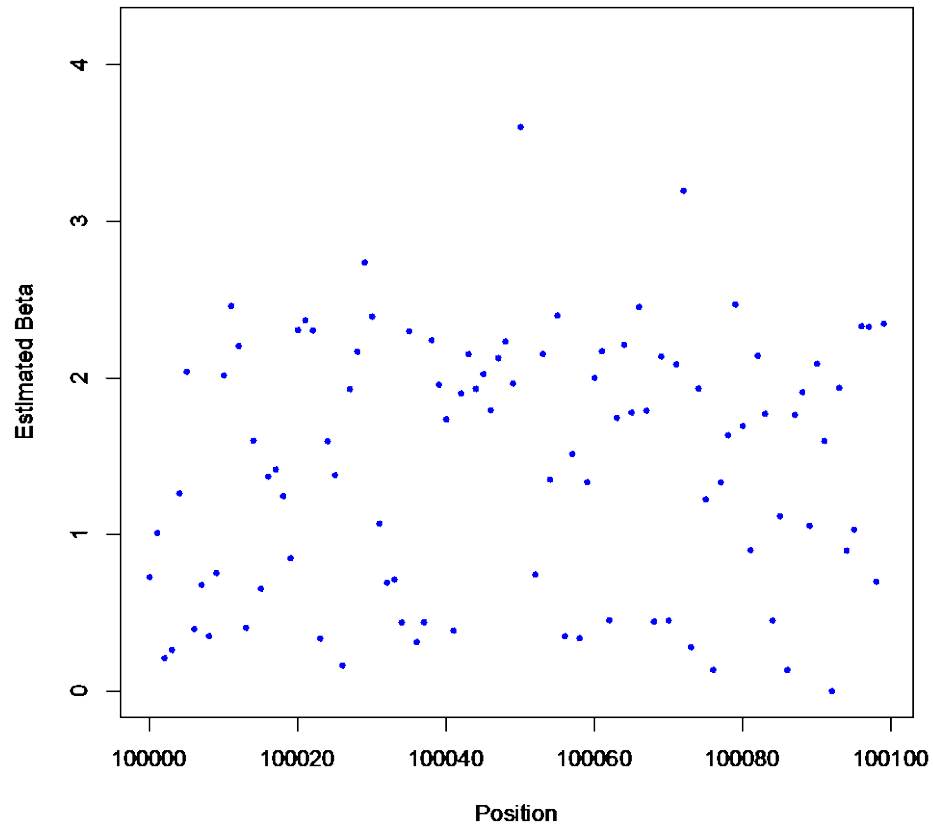
Statistical fine-mapping – conceptual introduction



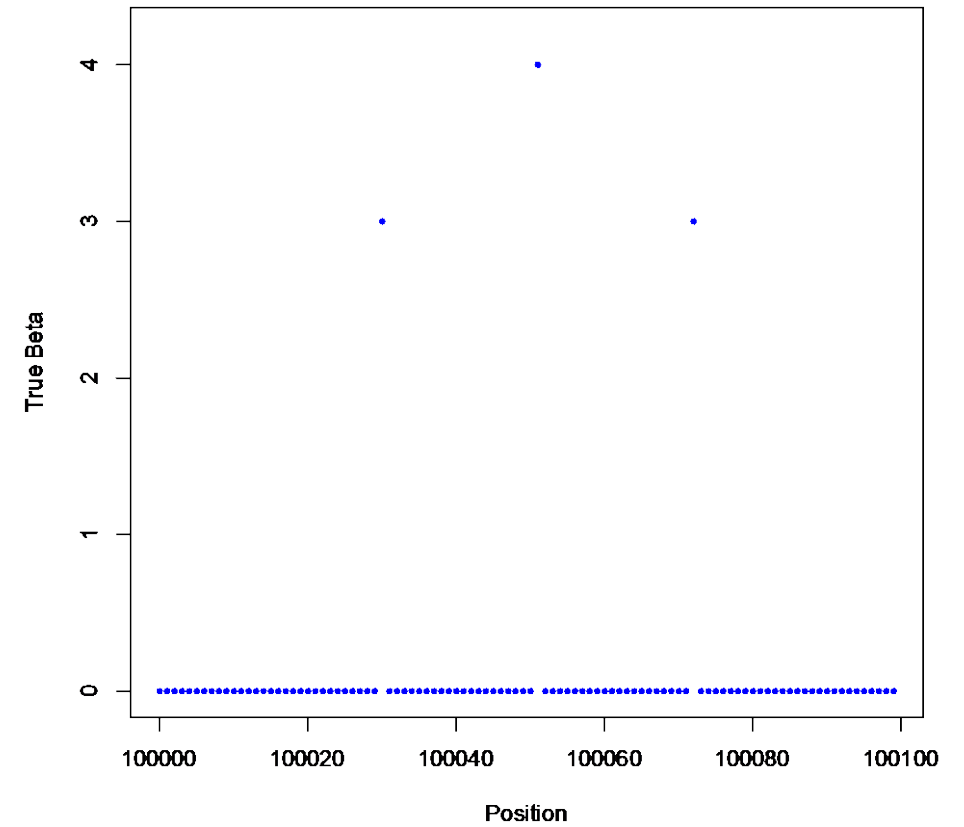
+ LD =



Statistical fine-mapping – conceptual introduction



- LD =



Among the simplest models – Maller et al.

ARTICLES

nature
genetics

Bayesian refinement of association signals for 14 loci in 3 common diseases

The Wellcome Trust Case Control Consortium^{1,2}

The authors of this paper are:

Julian B Maller^{1,2}, Gilean McVean^{1,2}, Jake Byrnes¹, Damjan Vukcevic¹, Kimmo Palin³, Zhan Su¹, Joanna M M Howson^{4,5}, Adam Auton¹, Simon Myers^{1,2}, Andrew Morris¹, Matti Pirinen¹, Matthew A Brown^{6,7}, Paul R Burton^{8,9}, Mark J Caulfield¹⁰, Alastair Compston¹¹, Martin Farrall¹², Alistair S Hall¹³, Andrew T Hattersley^{14,15}, Adrian V S Hill¹, Christopher G Mathew¹⁶, Marcus Pembrey¹⁷, Jack Satsangi¹⁸, Michael R Stratton^{3,19}, Jane Worthington²⁰, Nick Craddock²¹, Matthew Hurles³, Willem Ouwehand^{3,22,23}, Miles Parkes²⁴, Nazneen Rahman¹⁹, Audrey Duncanson²⁵, John A Todd⁵, Dominic P Kwiatkowski^{1,3}, Nilesh J Samani^{26,27}, Stephen C L Gough^{28,29}, Mark I McCarthy^{1,28,29}, Panagiotis Deloukas³ & Peter Donnelly^{1,4}

Formalized Marchini et al. 2007 and
Servin & Stephens PLoS Genet 2007

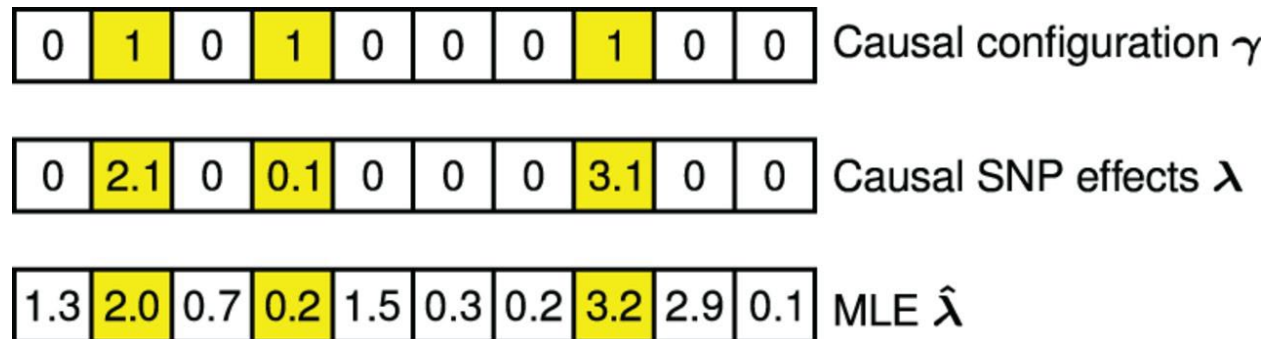
Assumes a single causal variant

$$\begin{aligned} \text{BF}_i &= \frac{\Pr(\mathbf{X}|M_i)}{\Pr(\mathbf{X}|M_0)} \\ &= \frac{\Pr(\mathbf{X}_i|M_i)\Pr(\mathbf{X}_{-i}|\mathbf{X}_i, M_i)}{\Pr(\mathbf{X}|M_0)} \\ &= \frac{\Pr(\mathbf{X}_i|M_i)\Pr(\mathbf{X}_{-i}|\mathbf{X}_i, M_0)}{\Pr(\mathbf{X}|M_0)} \\ &= \frac{\Pr(\mathbf{X}_i|M_i)\Pr(\mathbf{X}_{-i}|\mathbf{X}_i, M_0)\Pr(\mathbf{X}_i|M_0)}{\Pr(\mathbf{X}|M_0)\Pr(\mathbf{X}_i|M_0)} \\ &= \frac{\Pr(\mathbf{X}_i|M_i)\Pr(\mathbf{X}|M_0)}{\Pr(\mathbf{X}|M_0)\Pr(\mathbf{X}_i|M_0)} \\ &= \frac{\Pr(\mathbf{X}_i|M_i)}{\Pr(\mathbf{X}_i|M_0)}. \end{aligned}$$

$$\begin{aligned} \text{BF}_{\text{reg}} &= \frac{\Pr(\mathbf{X} | M)}{\Pr(\mathbf{X} | M_0)} \\ &= \frac{\sum_{i=1}^k \Pr(\mathbf{X} | M_i) \Pr(M_i | M)}{\Pr(\mathbf{X} | M_0)} \\ &= \sum_{i=1}^k \text{BF}_i \Pr(M_i | M). \end{aligned}$$

Maybe there is more than one causal SNP?

Fig. 1. The binary indicator vector γ determines which SNPs have non-zero causal effects (λ). The corresponding causal ...



$$p(\lambda|\gamma) = \mathbf{N}(\lambda|\mathbf{0}, s_{\lambda}^2 \sigma^2 \mathbf{\Delta}_{\gamma}),$$

User-specified prior variance for causal SNPs

Variance of the trait

Diagonal matrix with elements of gamma on the diagonal

JOURNAL ARTICLE

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

Christian Benner, Chris C.A. Spencer, Aki S. Havulinna, Veikko Salomaa, Samuli Ripatti, Matti Pirinen

Many methods for multiple causal variants

Identifying Causal Variants at Loci with Multiple Signals of Association

Farhad Hormozdiari,^{*1} Emrah Kostem,^{*1} Eun Yong Kang,^{*} Bogdan Pasaniuc,^{1,2,3} and Eleazar Eskin^{*1,2,3}
^{*}Department of Computer Science, ¹Department of Human Genetics, and ²Department of Pathology and Laboratory Medicine, University of California, Los Angeles, California 90095

Caviar - 2014

Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics

Wenan Chen,^{*} Beth R. Larrabee,^{*} Inna G. Ovsyannikova,[†] Richard B. Kennedy,[†] Iana H. Haralambieva,[†] Gregory A. Poland,[†] and Daniel J. Schaid^{*1}
^{*}Division of Biostatistics and [†]Mayo Clinic Vaccine Research Group, Mayo Clinic, Rochester, Minnesota 55905

Caviarbf - 2015

Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies

Gleb Kichaev, Wen-Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L. Price, Peter Kraft, Bogdan Pasaniuc 

JOURNAL ARTICLE

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

Christian Benner , Chris C.A. Spencer, Aki S. Havulinna, Veikko Salomaa, Samuli Ripatti, Matti Pirinen  Author Notes

PAINTOR - 2015

FINEMAP - 2016

JOURNAL ARTICLE







A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping

Gao W

ARTICLE

DOI: 10.1093/biostatistics/19.1.1 OPEN

Causal associations between risk factors and common diseases inferred from GWAS summary data

Zhihong Zhu¹, Zhili Zheng^{1,2}, Futao Zhang¹, Yang Wu¹, Maciej Trzaskowski¹, Robert Maier , Matthew R. Robinson , John J. McGrath , Peter M. Visscher , Naomi R. Wray  & Jian Yang 

SuSIE - 2020

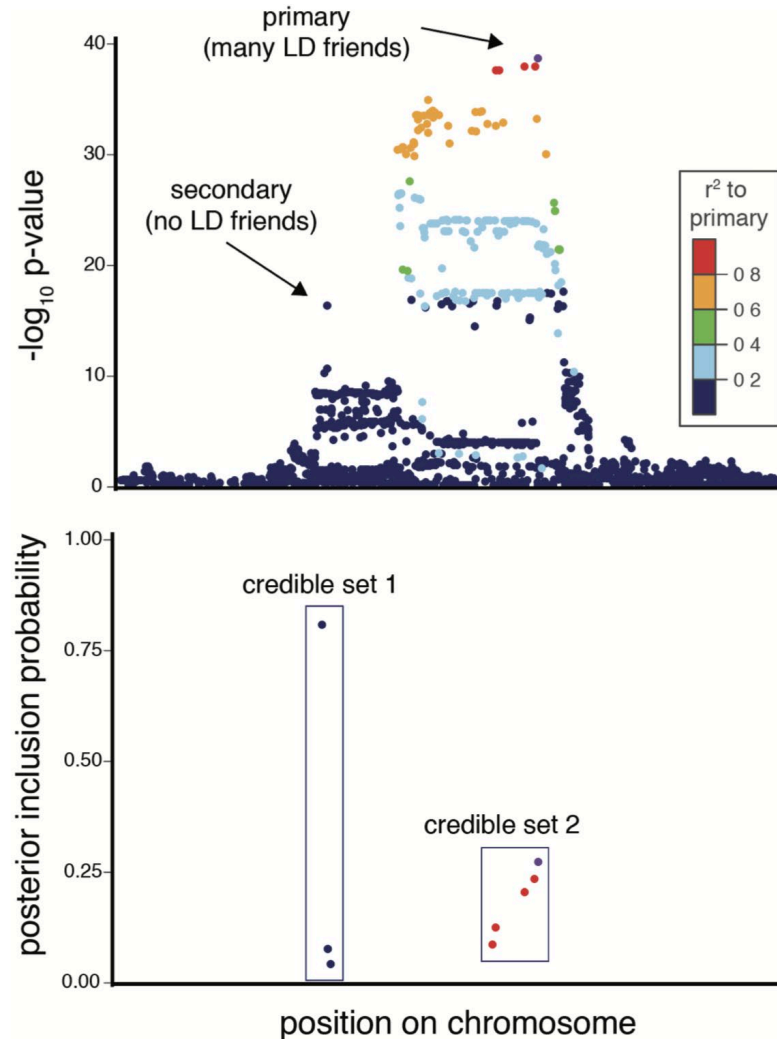
COJO and COJO-ABF - 2018

Handy guide to assumptions differences in methods

	①	②
y	quantitative	binary
X	genotype	genotype and heterozygote
β	fixed	random

Paper	Assumptions
BIMBAM [Servin and Stephens, 2007]	$y^{①}, X^{②}, \beta^{②}$
Maller et al. [Maller et al., 2012]	$y^{②}, X^{②}, \beta^{②}$
fgwas [Pickrell, 2014]	$y^{①②}, X^{①}, \beta^{②}$
CAVIAR [Hormozdiari et al., 2014]	$y^{①②}, X^{①}, \beta^{①②}$
CAVIARBF [Chen et al., 2015]	$y^{①②}, X^{②}, \beta^{②}$
PAINTOR [Kichaev et al., 2014]	$y^{①②}, X^{①}, \beta^{①}$
FINEMAP [Benner et al., 2016]	$y^{①②}, X^{①}, \beta^{②}$
SuSiE [Wang et al., 2018]	$y^{①}, X^{①}, \beta^{②}$

Fine-mapping quantifies causal probability



Posterior inclusion probability (PIP)

- Probability under the model, given the data, that the variant is causal for the disease

95% credible set (CS)

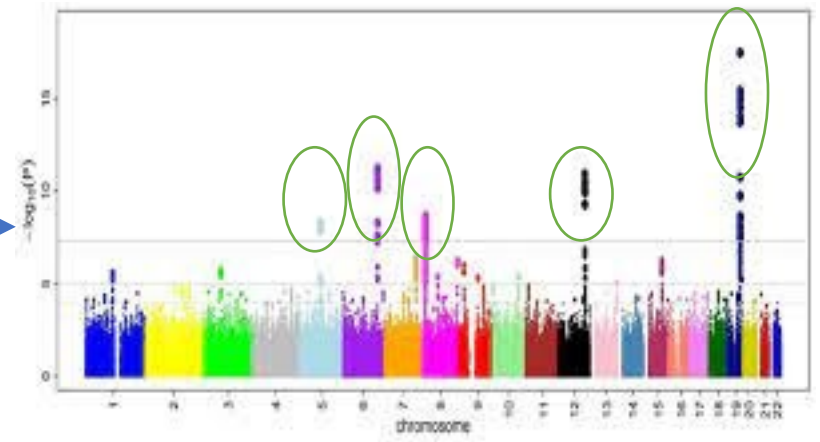
- Smallest set of variants for a single effect that together have > 95% probability of being causal for the disease

How robust are these methods? Replication Failure Rate



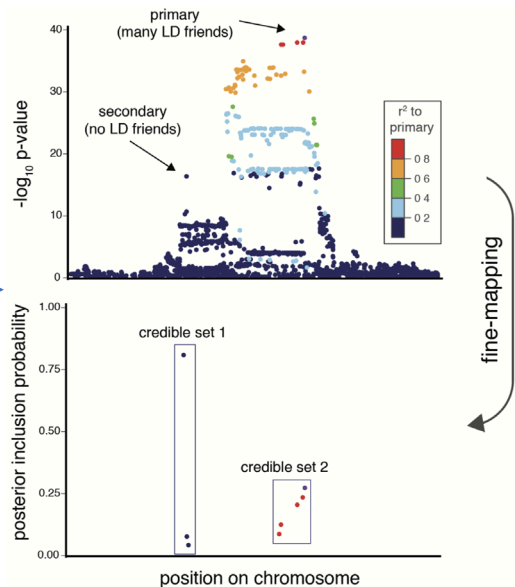
Take large well-powered GWAS study

Downsample to 100K individuals
Run a GWAS

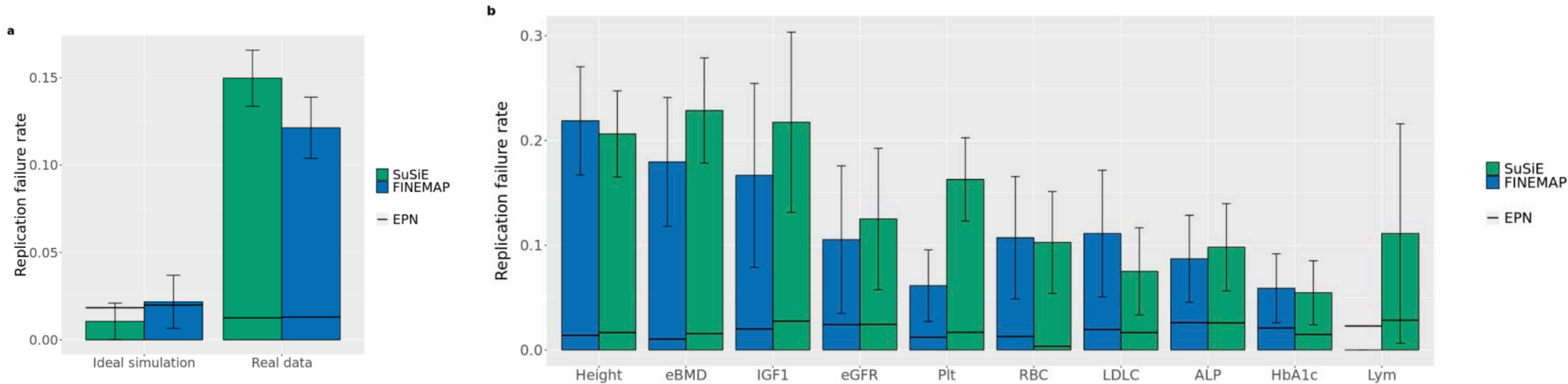


Run fine-mapping on genome-wide significant loci

Later rinse repeat to inspect PIP calibration



Replication Failure Rate



What about modeling an ‘infinitesimal’ contribution?

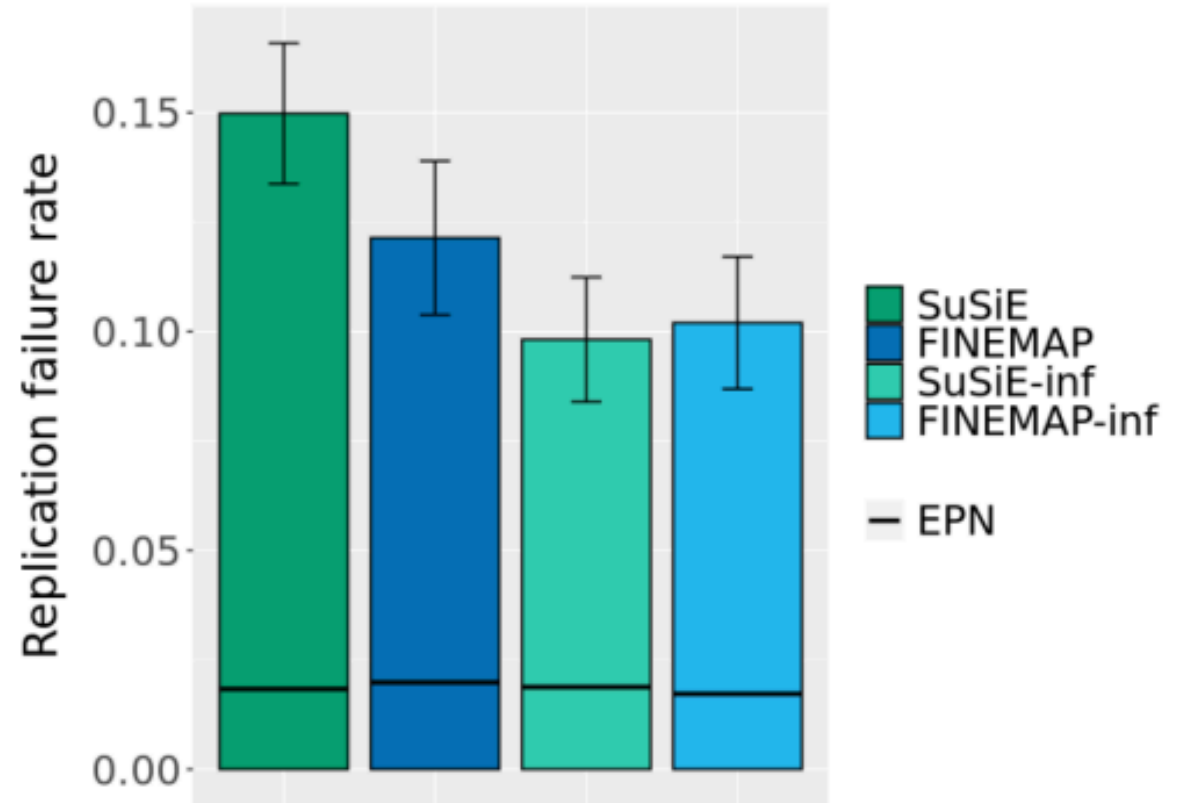
Take this model

$$p(\lambda|\gamma) = \mathbf{N}(\lambda|\mathbf{0}, s_\lambda^2 \sigma^2 \mathbf{\Delta}_\gamma),$$

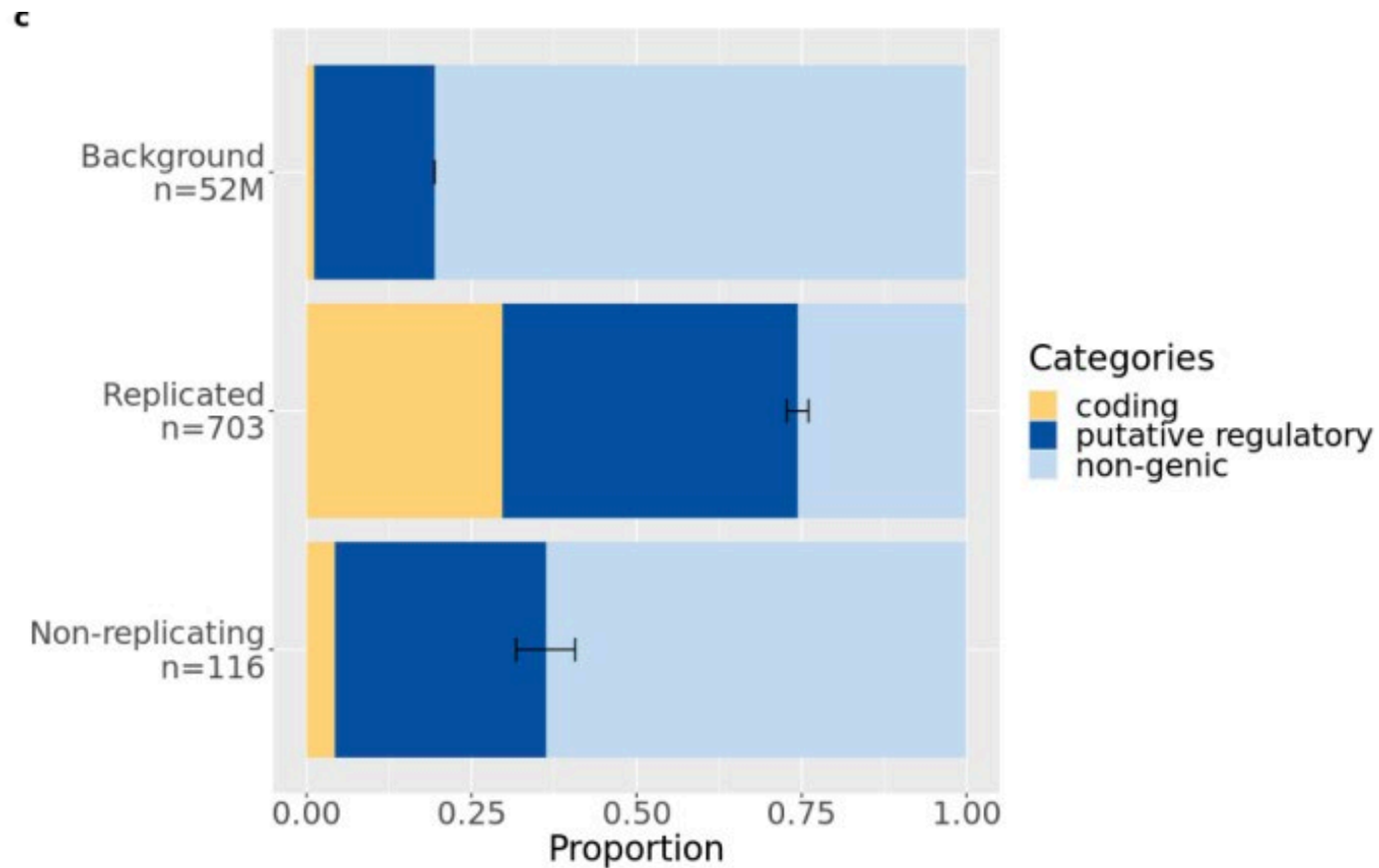
And add this consideration:

$$\pi_0 \mathbf{N}(0, \mathbf{s}^2) + (1 - \pi_0) \delta_0.$$

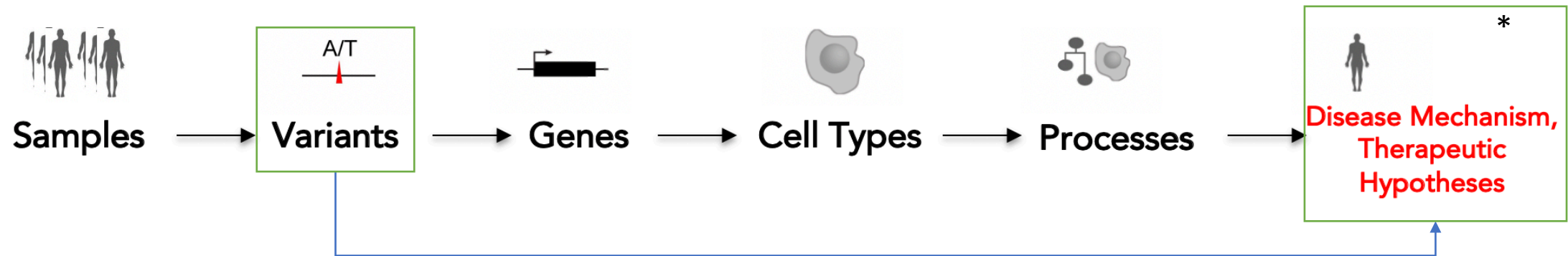
a



Adding functional information improves but does not guarantee replicability!

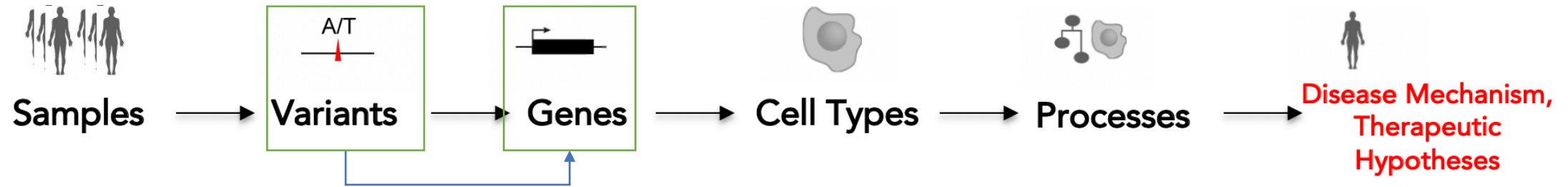


How does the associated variant impact function?



This is the GWAS!

How does the associated variant impact function?



This is the an eQTL study

eQTL = expression quantitative trait locus

Exploring eQTL resources



Navigate to <https://gtexportal.org/home>

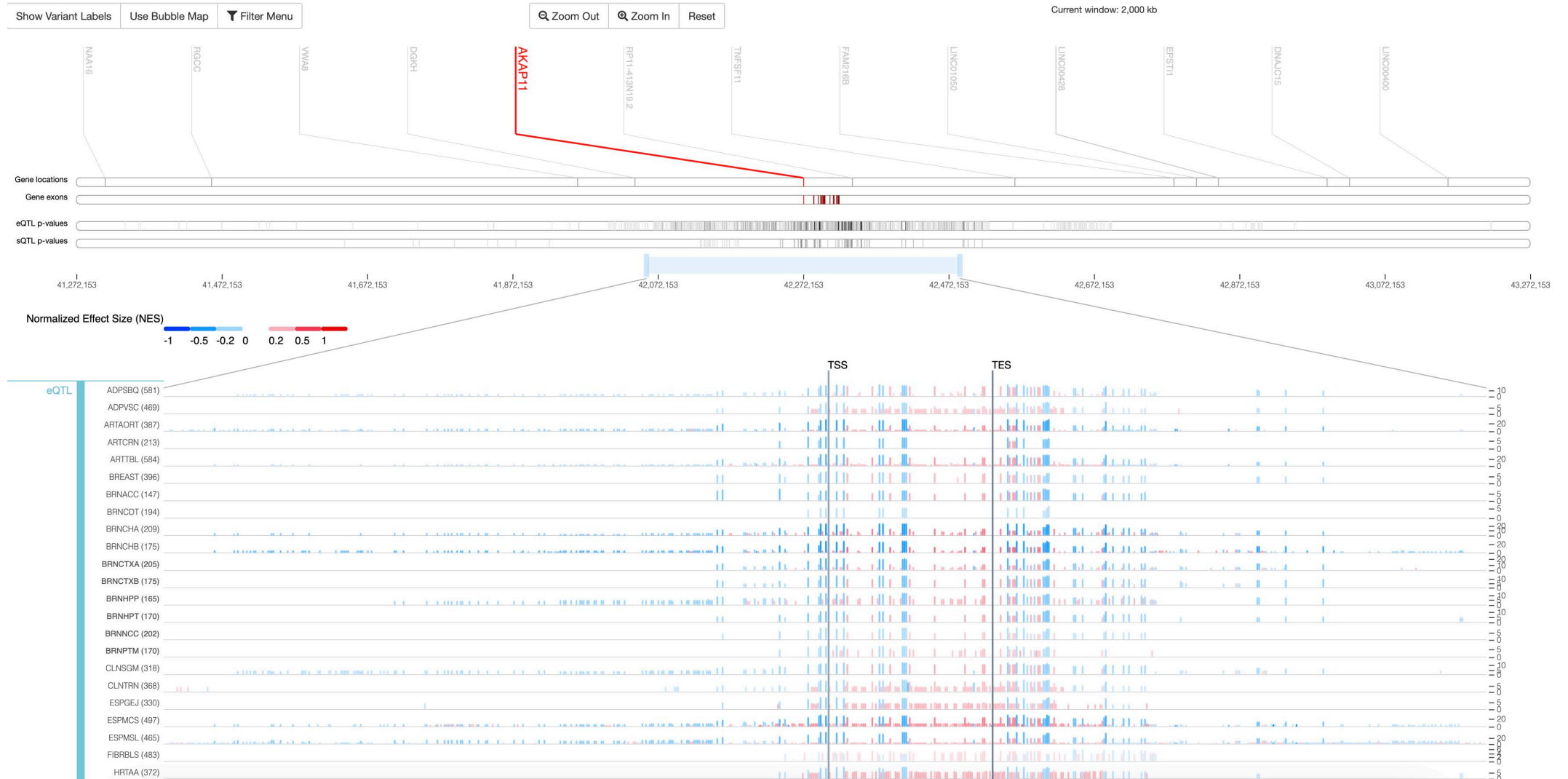
Click QTL browser

Click Locus Browser (Gene-centric)

A screenshot of the GTEx Portal website. The top navigation bar includes the GTEx Portal logo on the left and links for "About GTEx", "Publications", "Access Biospecimens", "FAQs", and "Contact" on the right. Below the navigation bar is a search bar labeled "Search Gene or SNP ID...". The main navigation menu is open, showing options: "Home", "Downloads", "Expression", "Single Cell", "QTL", "IGV Browser", "Tissues & Histology", and "Documentation". The "QTL" menu is expanded, and the "Locus Browser (Gene-centric)" option is circled in blue. Other options in the QTL menu include "Locus Browser (variant-centric)", "eQTL Dashboard", and "eQTL Calculator". A notification banner is visible on the left side of the page, dated "2023-02-23", with the text "GTEx API V2 Released, GTEx API...". At the bottom of the page, there are two blue buttons: "Resource Overview" and "Explore GTEx".

Explore your favorite gene!

Example – AKAP11



Can we test whether these associations are the same?

COLOC

- H_0 : No association with either trait
- H_1 : Association with trait 1, not with trait 2
- H_2 : Association with trait 2, not with trait 1
- H_3 : Association with trait 1 and trait 2, two independent SNPs
- H_4 : Association with trait 1 and trait 2, one shared SNP

$$P(H_h|D) \propto \sum_{S \in S_h} P(D|S)P(S)$$

PP4

$$= P(H_4|D)$$

$$= \frac{P(H_4|D)}{P(H_0|D) + P(H_1|D) + P(H_2|D) + P(H_3|D) + P(H_4|D)}$$

$$= \frac{\frac{P(H_4|D)}{P(H_0|D)}}{1 + \frac{P(H_1|D)}{P(H_0|D)} + \frac{P(H_2|D)}{P(H_0|D)} + \frac{P(H_3|D)}{P(H_0|D)} + \frac{P(H_4|D)}{P(H_0|D)}}$$

The screenshot shows the PLOS Genetics article header with navigation links (BROWSE, PUBLISH, ABOUT, SEARCH), the article title, authors (Claudia Giambartolomei et al.), and a statistics table.

977 Save	1,459 Citation
52,240 View	9 Share

Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types

Sung Chun¹⁻³, Alexandra Casparino⁴, Nikolaos A Patsopoulos^{3,5,6}, Damien C Croteau-Chonka^{2,7}, Benjamin A Raby^{2,7}, Philip L De Jager^{2,3,5,6}, Shamil R Sunyaev^{1-3,8} & Chris Cotsapas^{3,4,9}

JLIM model – maximum likelihood treatment rather than Bayesian

Managing multiple variants

Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types

Sung Chun¹⁻³, Alexandra Casparino⁴, Nikolaos A Patsopoulos^{3,5,6}, Damien C Croteau-Chonka^{2,7}, Benjamin A Raby^{2,7}, Philip L De Jager^{2,3,5,6}, Shamil R Sunyaev^{1-3,8} & Chris Cotsapas^{3,4,9}

Naturally fits into the methodology

PLOS GENETICS BROWSE PUBLISH ABOUT SEARCH advanced search

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

A more accurate method for colocalisation analysis allowing for multiple causal variants

Chris Wallace

Version 2 Published: September 29, 2021 • <https://doi.org/10.1371/journal.pgen.1009440>

140 Save	42 Citation
6,668 View	38 Share

Coloc multivariate extension

Alternate approaches – with potential specificity challenges

ARTICLES

nature
genetics

Integrative approaches for large-scale transcriptome-wide association studies

Alexander Gusev¹⁻³, Arthur Ko^{4,5}, Huwenbo Shi⁶, Gaurav Bhatia¹⁻³, Wonil Chung¹, Brenda W J H Penninx⁷, Rick Jansen⁷, Eco J C de Geus⁸, Dorret I Boomsma⁸, Fred A Wright⁹, Patrick F Sullivan¹⁰⁻¹², Elina Nikkola⁴, Marcus Alvarez⁴, Mete Civelek¹³, Aldons J Lusis^{4,13}, Terho Lehtimäki¹⁴, Emma Raitoharju¹⁴, Mika Kähönen¹⁵, Ilkka Seppälä¹⁴, Olli T Raitakari^{16,17}, Johanna Kuusisto¹⁸, Markku Laakso¹⁸, Alkes L Price¹⁻³, Päivi Pajukanta^{4,5} & Bogdan Pasaniuc^{4,6,19}

TWAS

TECHNICAL REPORTS

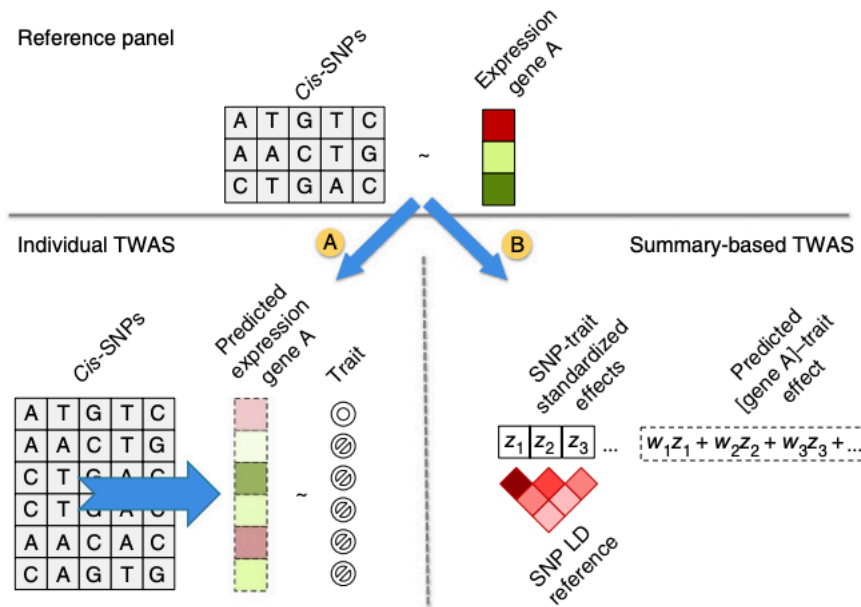
nature
genetics

A gene-based association method for mapping traits using reference transcriptome data

Eric R Gamazon^{1,2,9}, Heather E Wheeler^{3,9}, Kanaan P Shah^{1,9}, Sahar V Mozaffari⁴, Keston Aquino-Michaels¹, Robert J Carroll⁵, Anne E Eyler⁶, Joshua C Denny⁵, GTEx Consortium⁷, Dan L Nicolae^{1,4,8}, Nancy J Cox^{1,2,4} & Hae Kyung Im¹

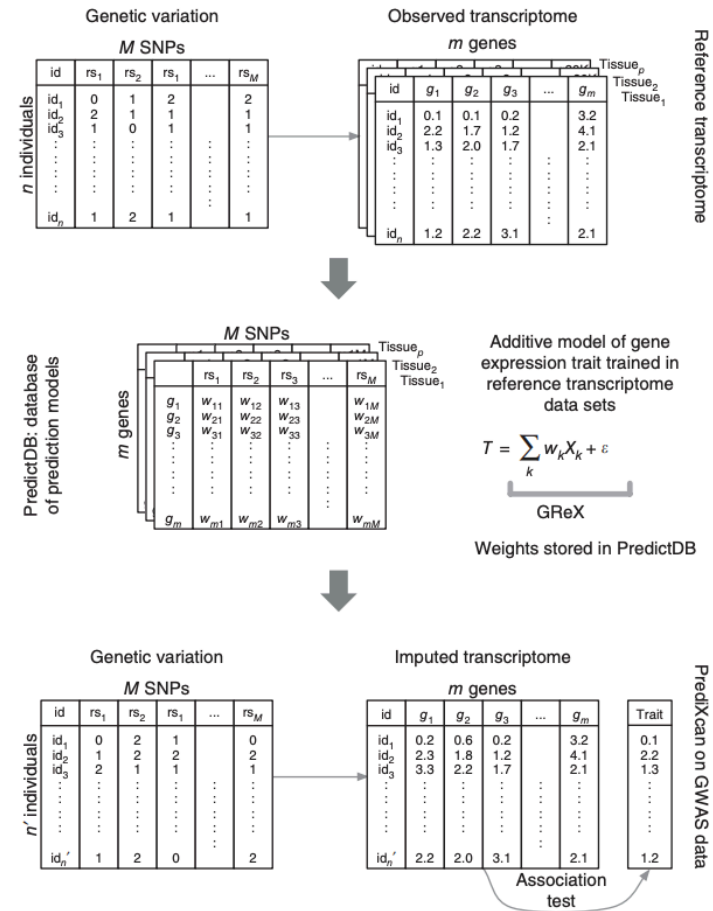
PrediXcan

TWAS and PrediXcan descriptions



TWAS

Why ought we be concerned?



PrediXcan

Wright-Fisher

EVOLUTION IN MENDELIAN POPULATIONS

SEWALL WRIGHT

University of Chicago, Chicago, Illinois

Received January 20, 1930

THE
GENETICAL THEORY OF
NATURAL SELECTION

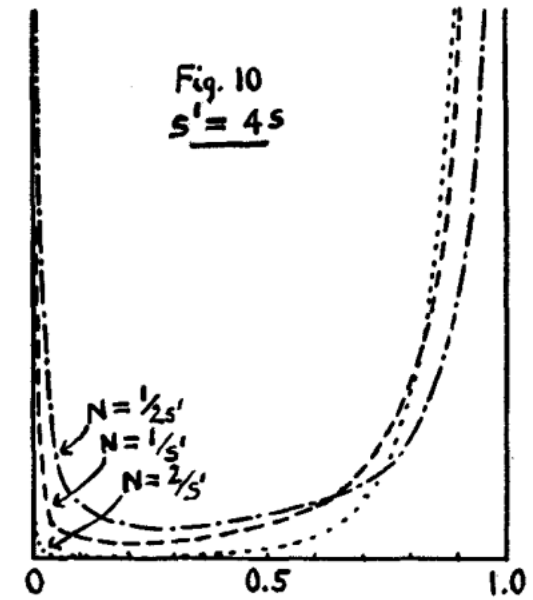
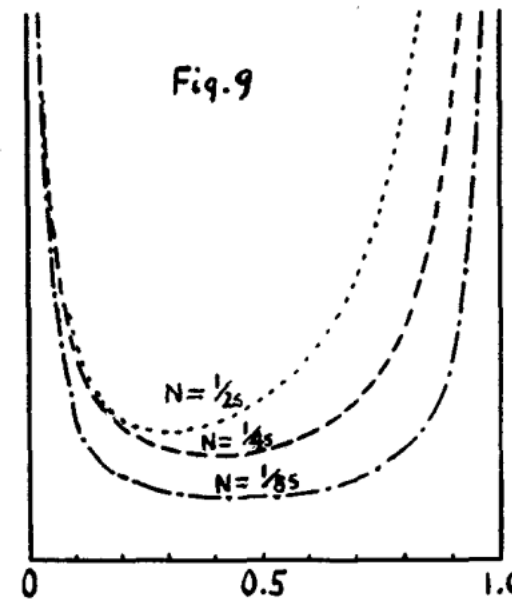
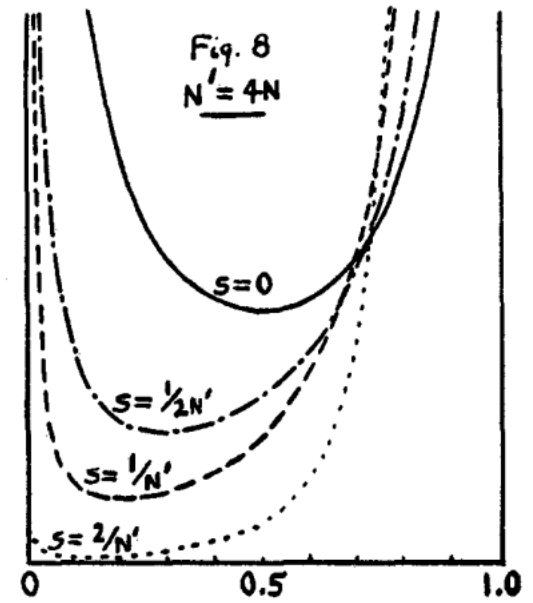
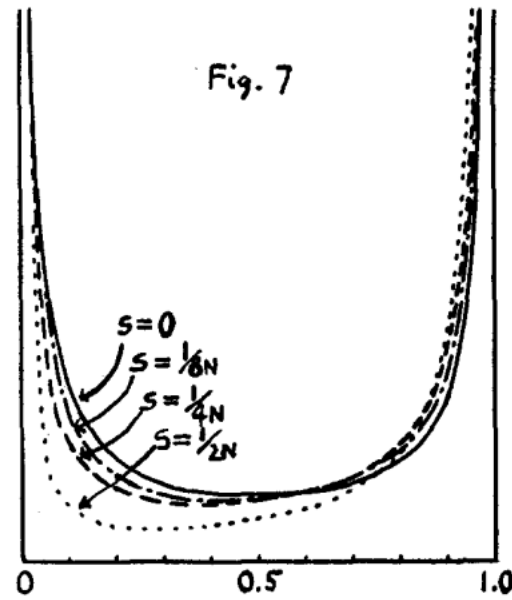
BY

R. A. FISHER, Sc.D., F.R.S.

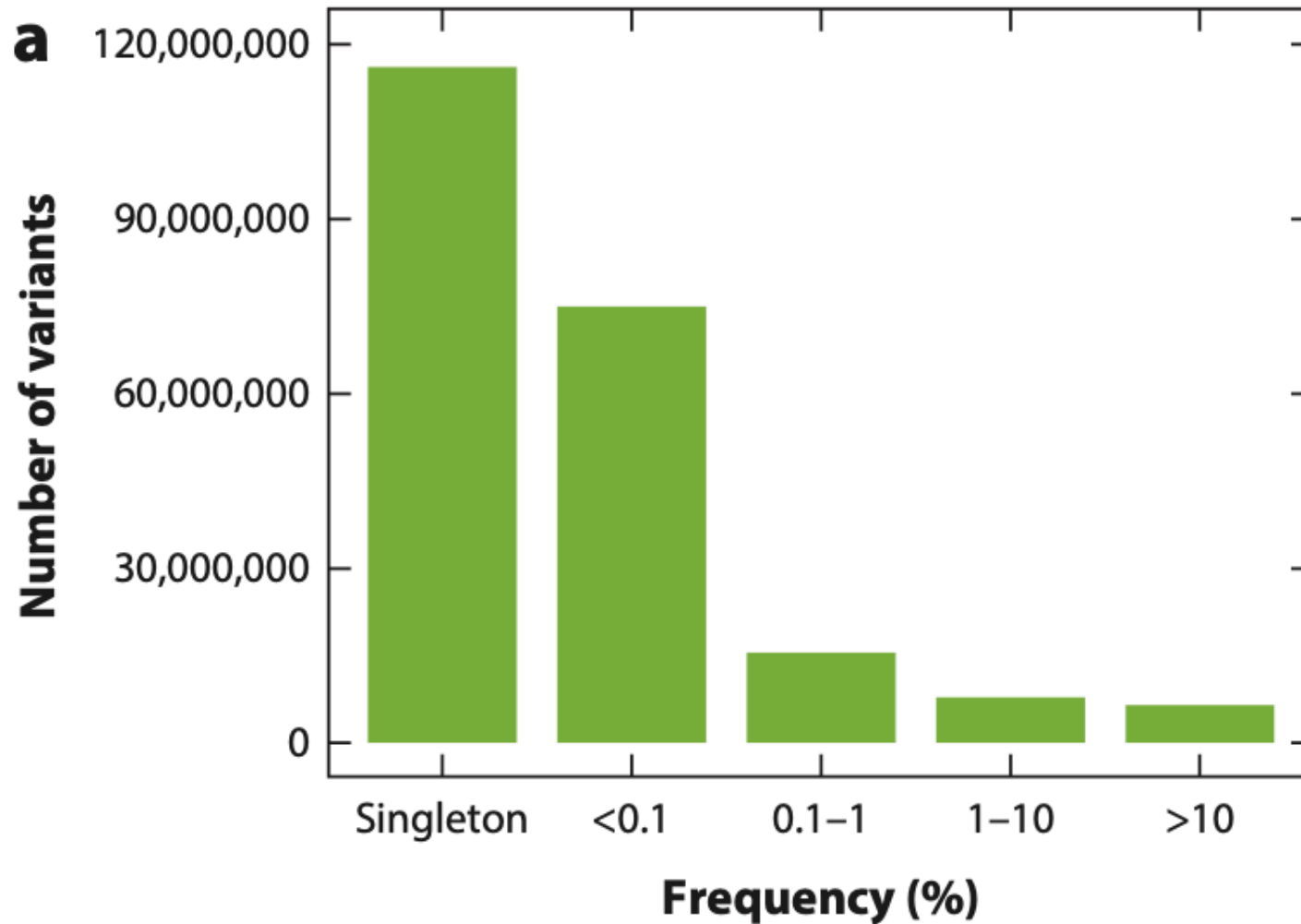
OXFORD
AT THE CLARENDON PRESS
1930

Wright 1930

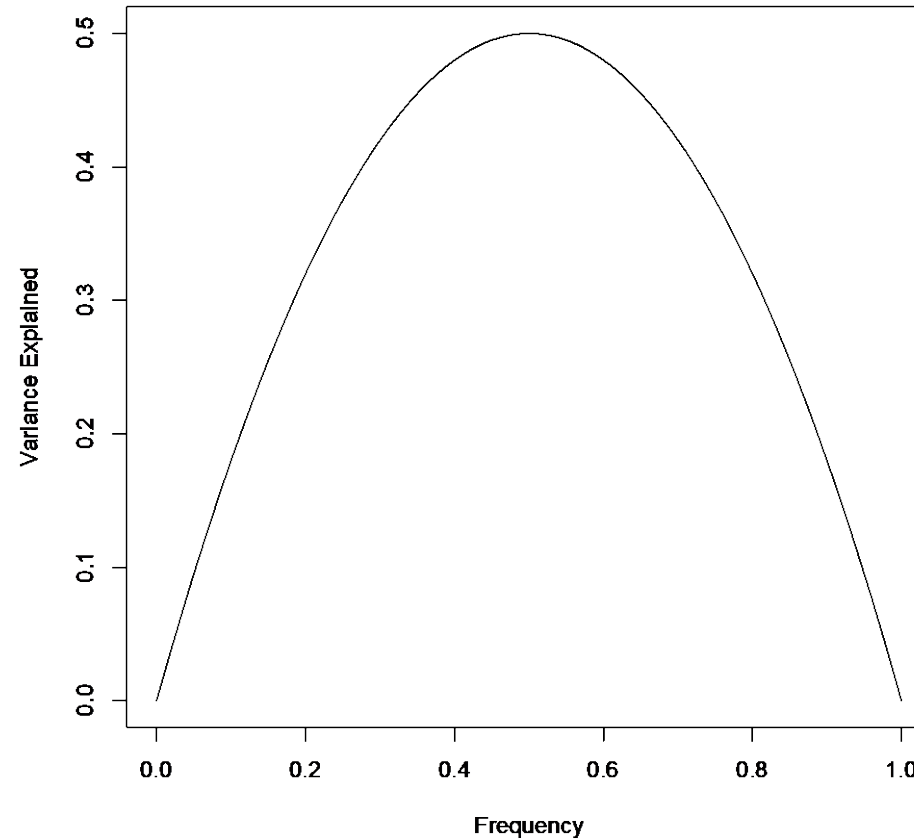
Site frequency spectra



Empirical SFS – very close to drift

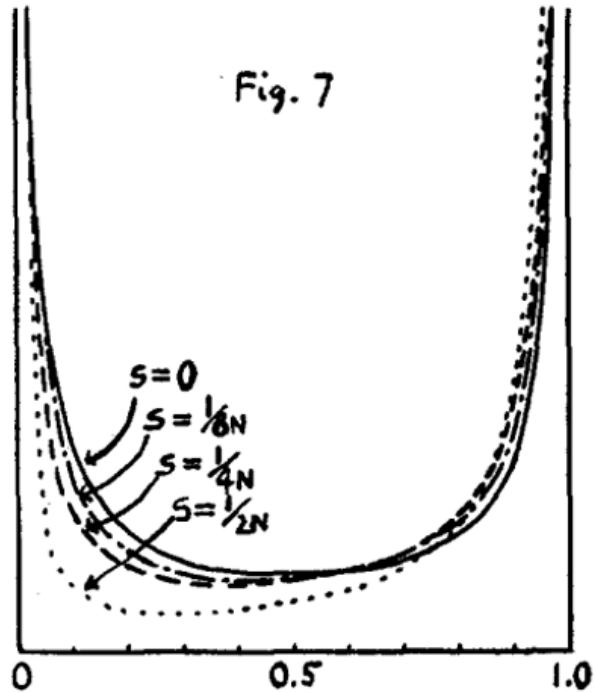


Variance explained (alpha =0) – effect size is constant across frequency spectrum

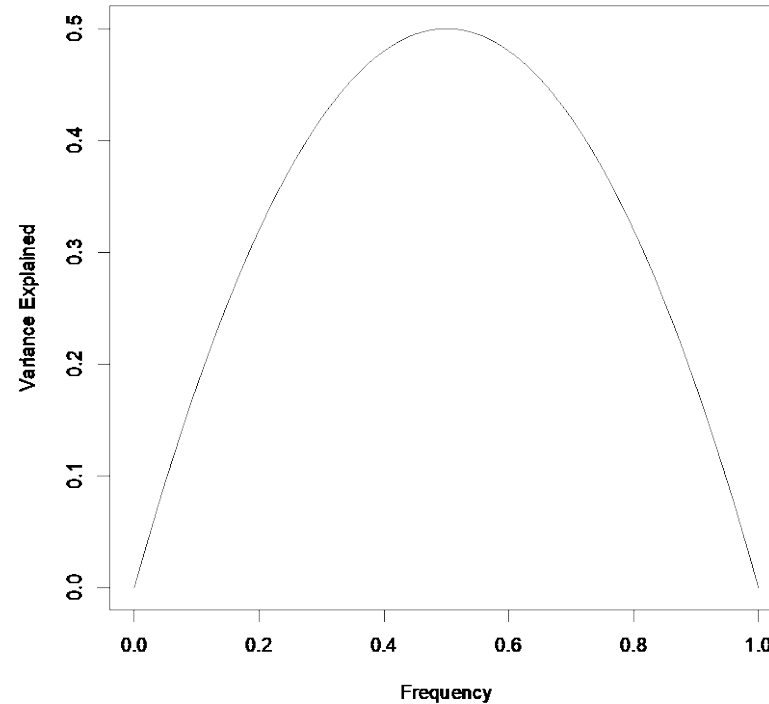


Recall variance explained is $2pqa^2$

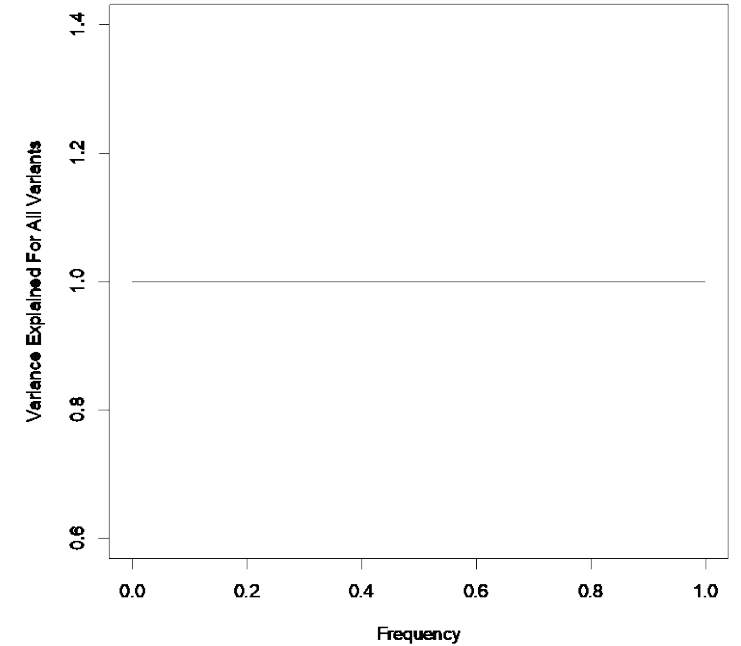
When we add the SFS + variance explained when alpha = 0



*



||



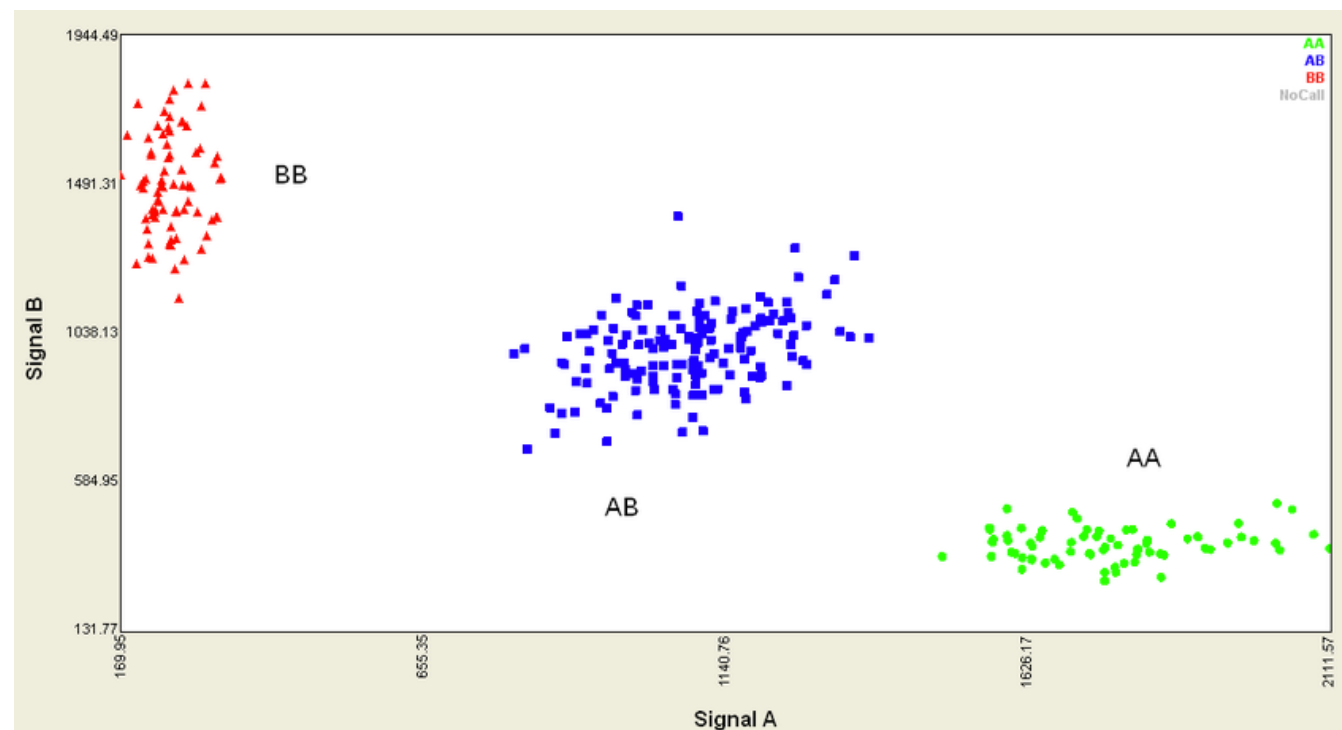
Few preparatory remarks for
the **haxl** session

Microarray (genotyping)

QC removes loci with bad binding chemistry

Samples can be assigned high-confidence genotype calls at remaining loci

PLINK BED files include AA/AB/BB/NA call states, no probability information.

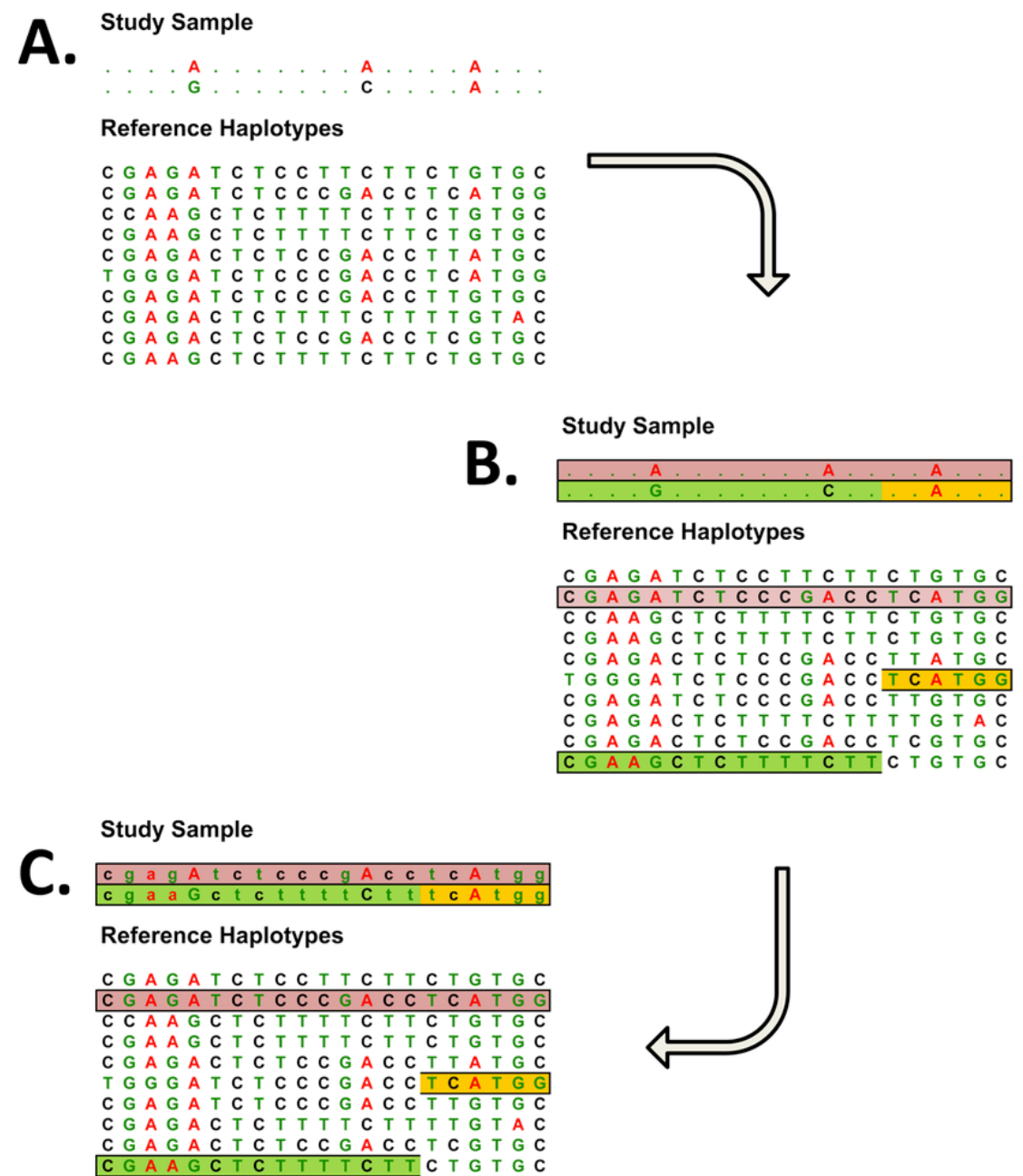


Microarray + imputation

Genotypes at directly measured SNPs are “hard calls”

Genotypes at imputed SNPs are **probability distributions** over possible genotype states, or a **genotype dosage**.

Oxford BGEN files contain probabilities for each genotype configuration.



So where do sequencing
data originate?

High-throughput sequencing

Also called:

- Shotgun sequencing
- Short read sequencing
- Next generation sequencing

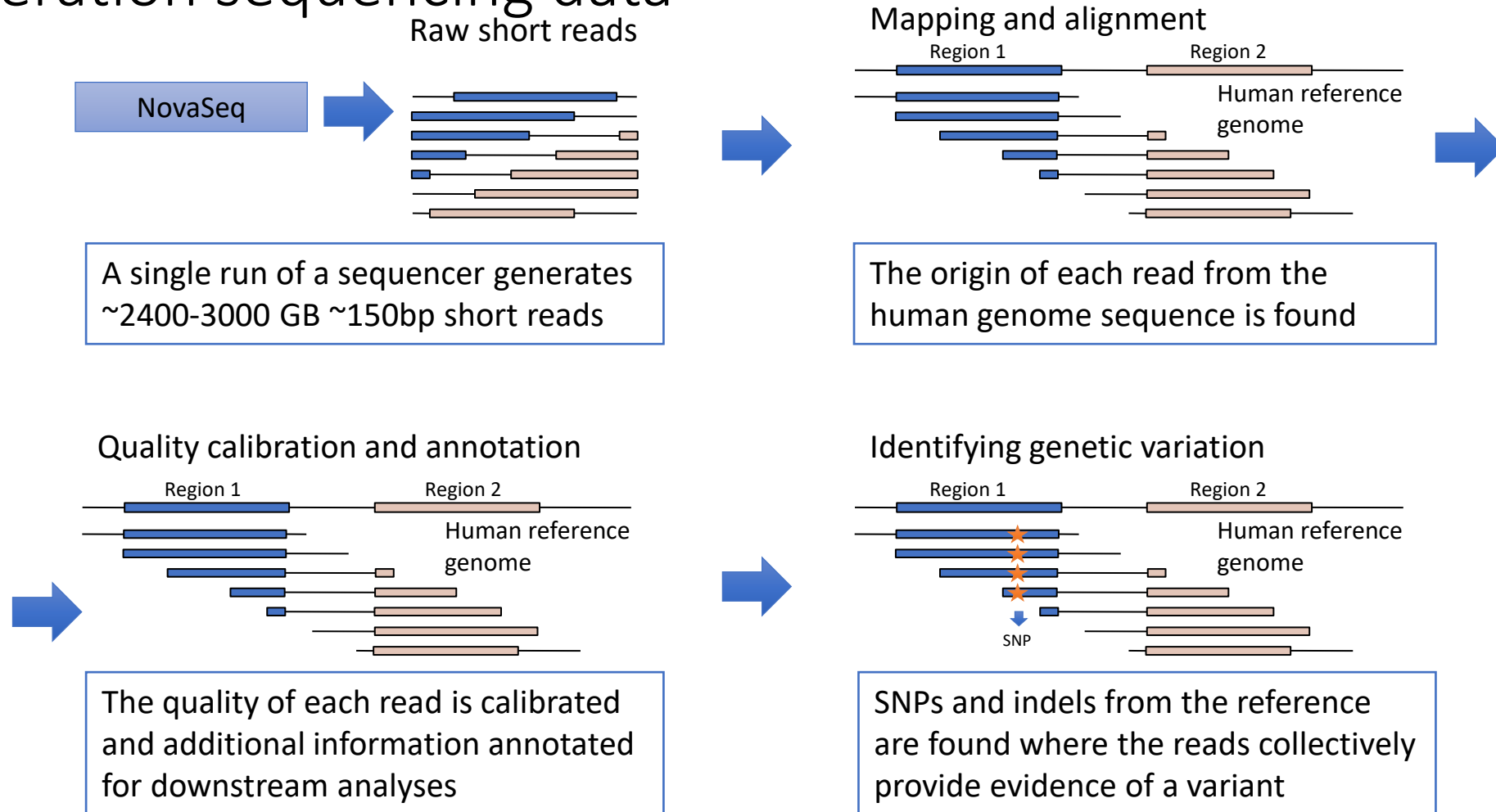
Genome is broken down into small segments, and many segments are read in parallel

Segments are computationally assembled into a complete sequence using overlaps (***de novo assembly***) or aligned against an existing reference genome (**alignment**).



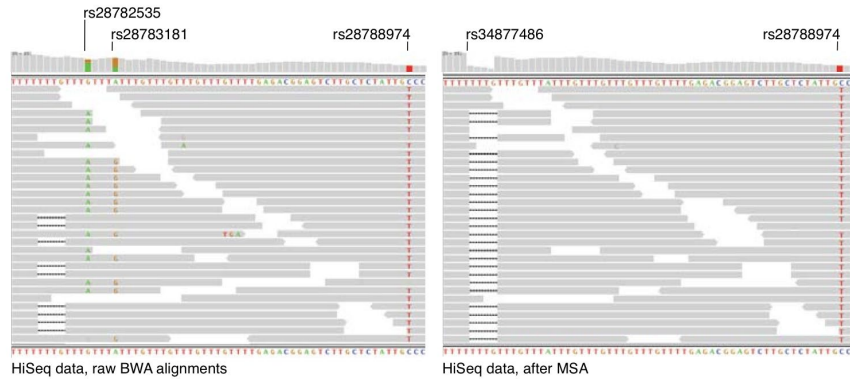
Illumina NovaSeqX
3 Terabases per flow cell run

From unmapped reads to true genetic variation in next-generation sequencing data

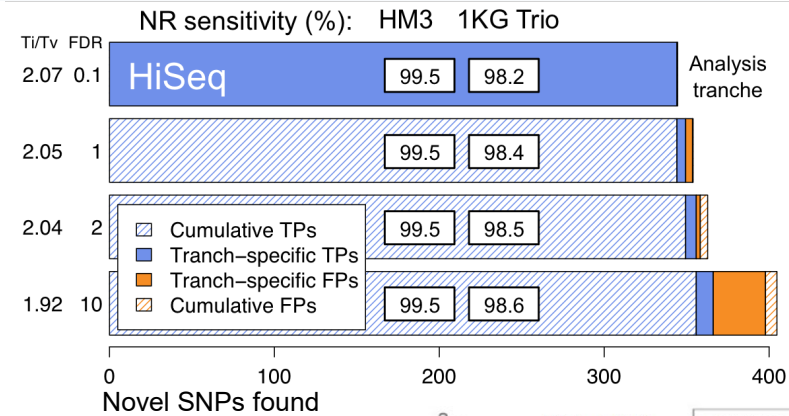


Data processing and analysis methods

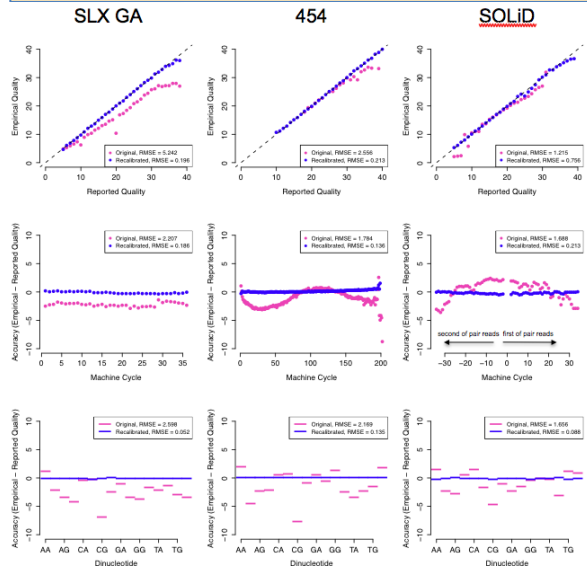
Local realignment



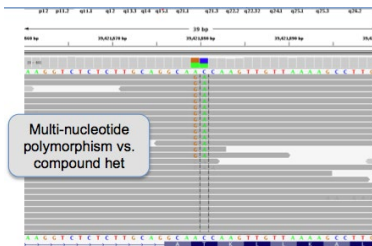
Variation discovery and genotyping



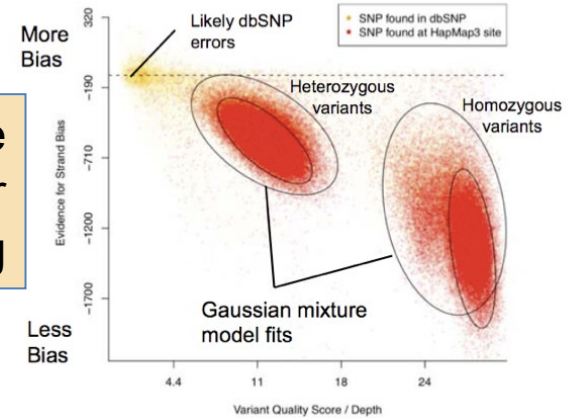
Base quality score recalibration



Read-backed phasing



Adaptive error modeling



VariantEval

Validation results of de novo mutation calls on NA12878

	Germline	Somatic	False Positive
Hard Filtered	49	952	1304
Variant Recalibrated	49	929	85
Complete Genomics	47	886	255

Sequenced variant calls

- GT - best-guess genotype.
 - 0/0 for homozygous reference, 1/2 for heterozygous non-reference, ./ for missing
- GQ - conditional genotype quality.
 - GQ 10, 20, 30 indicate 90%, 99%, 99.9% confidence in GT.
- DP - total read depth
- AD - read depth by allele.
 - AD = [10, 8] indicates 10 reads from the reference, 8 from first alternate.
- PL - scaled likelihoods of each genotype configuration.

Sequencing data QC...

...is hard. For a few:

- depth is important: contamination and mapping errors can cause spurious heterozygous calls
- Low-complexity regions are filled with insertions and deletions that defy a fixed reference genome
- Handling multiallelic sites is complicated, and often necessary