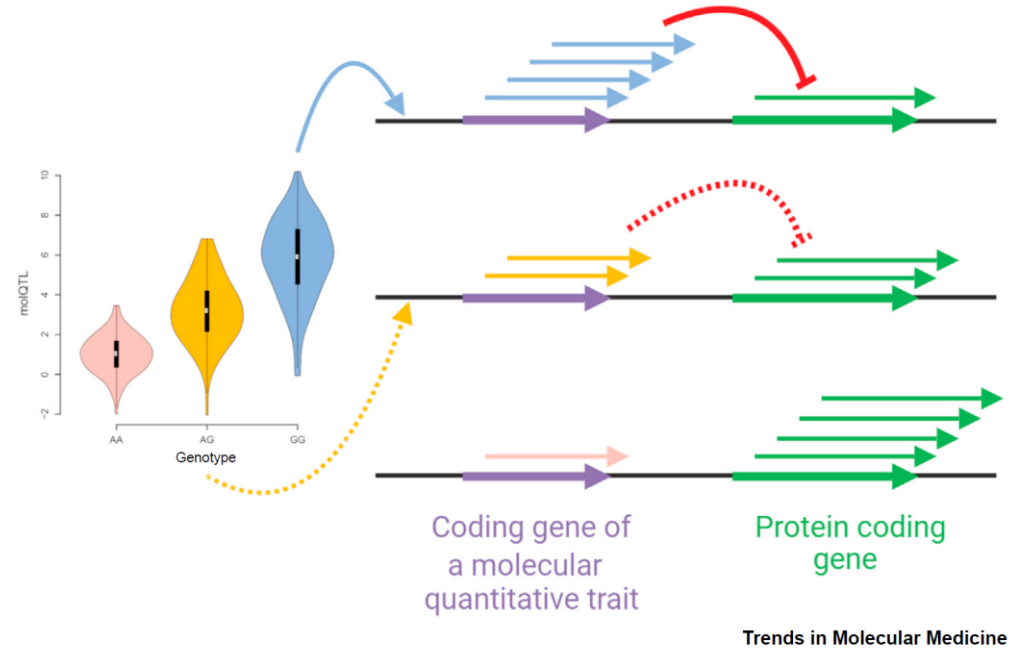
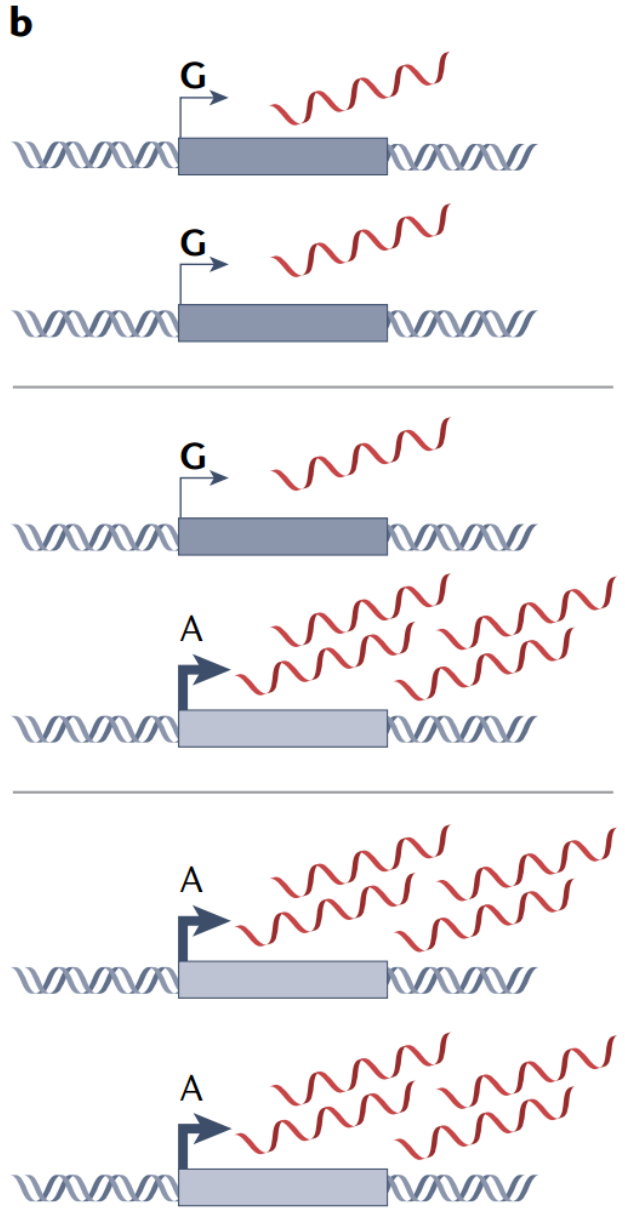
molQTL

孙健乐 (Jianle Sun)
Department of Bioinformatics & Biostatistics,
Shanghai Jiao Tong University

molQTL

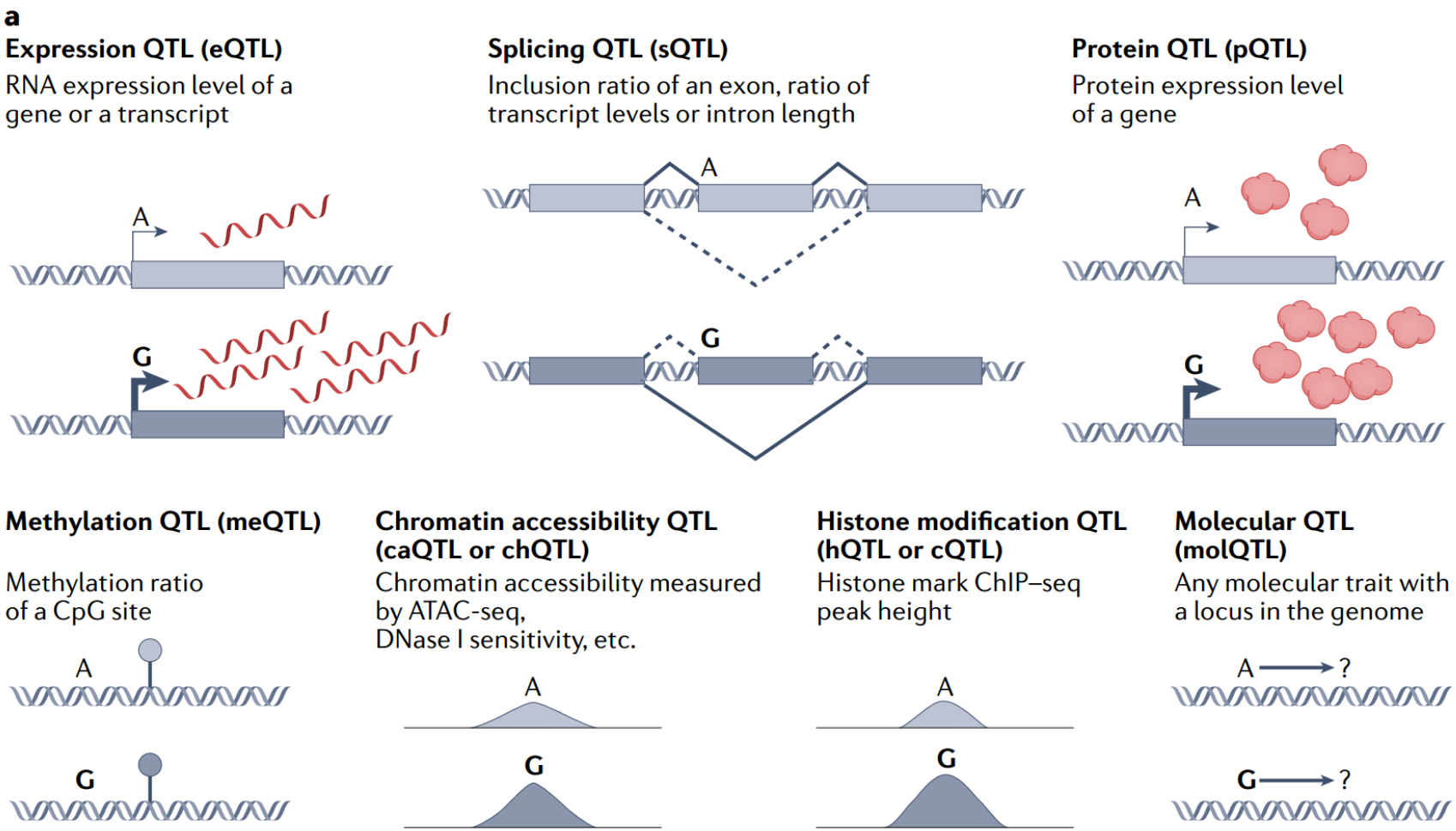
- Molecular quantitative trait locus (molQTL) is an umbrella term for loci with a genetic association for a quantitative level of a molecular trait





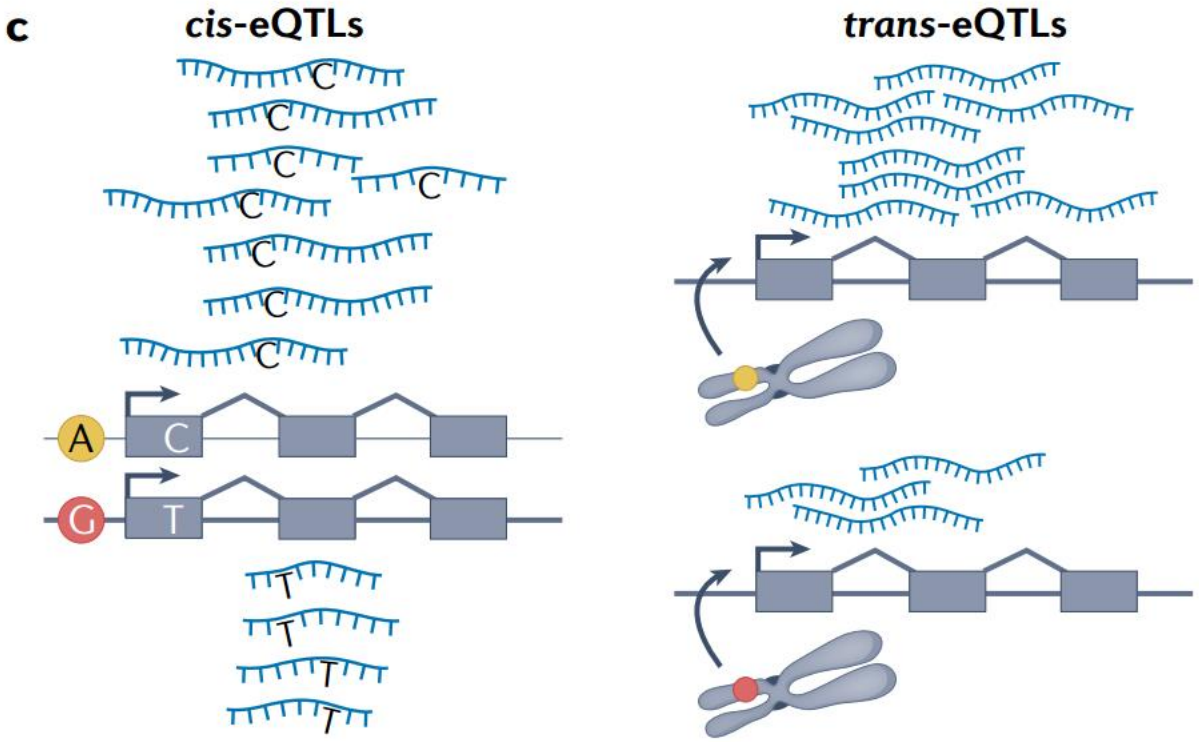
molQTL的类型

- 根据研究的分子性状不同



molQTL的类型

- 根据分析的范围：
  - *cis*-QTL：
    - 在所研究的分子特征（如基因）的附近（100kb-1Mb）的窗口内的位点
    - 通过同一染色质上的分子相互作用影响（启动子、增强子、沉默子等顺式作用元件*cis*-acting element）
  - *trans*-QTL：
    - 远端的、甚至其他染色体上的位点
    - 通过其他分子（如转录因子等反式作用因子*trans*-acting factor）的介导

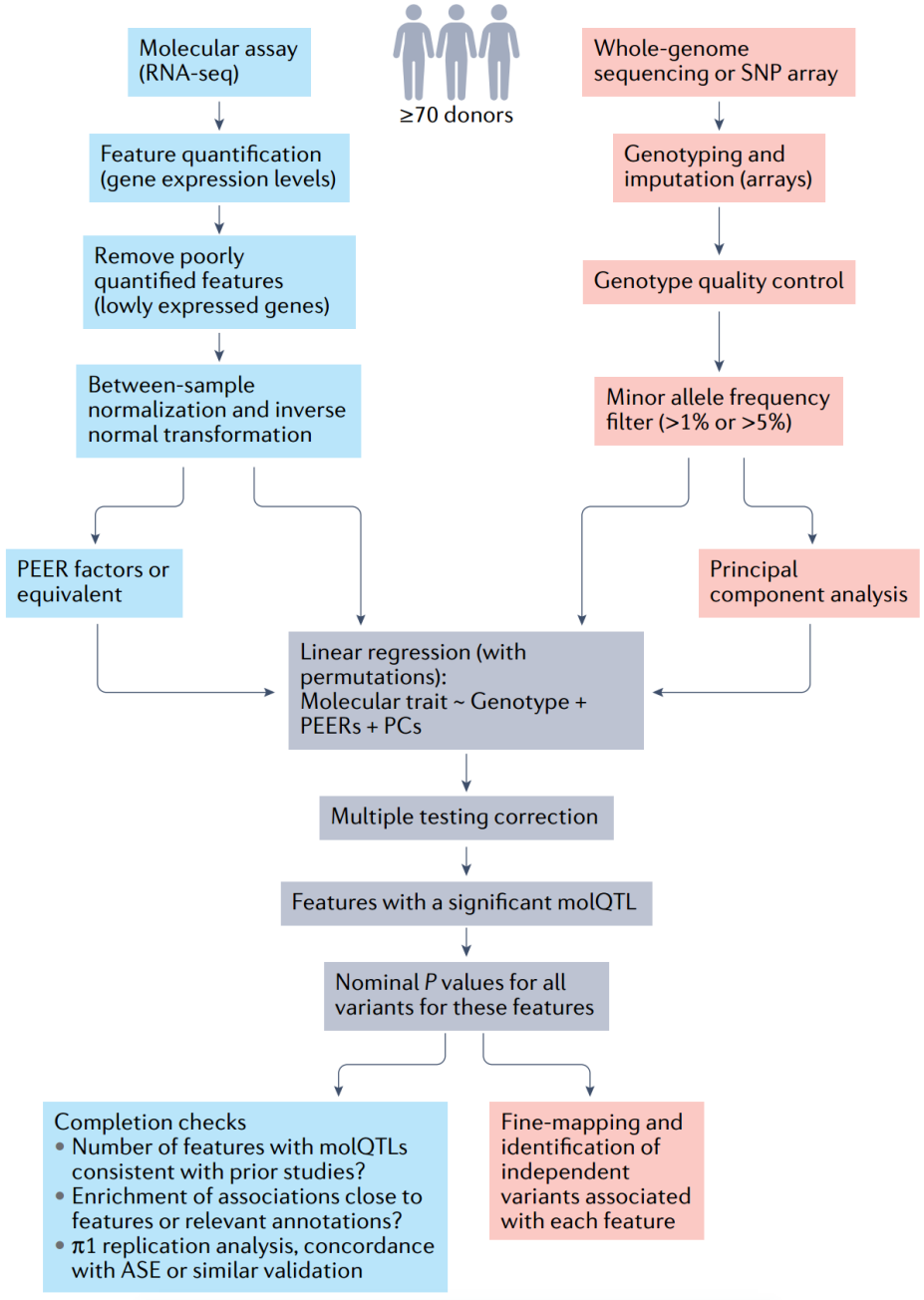


molQTL vs GWAS

- Nearly identical approaches that use regression to associate genetic variation with a quantitative phenotype in a population sample
- 区别：
  - QTL——分子性状，GWAS——宏观表型
  - QTL——置换检验（permutation）控制FDR，GWAS——设置固定的Bonferroni阈值
  - QTL——与分子性状相关的predefined loci（区分cis/trans），GWAS——没有明确的染色体区域
- 微妙关系

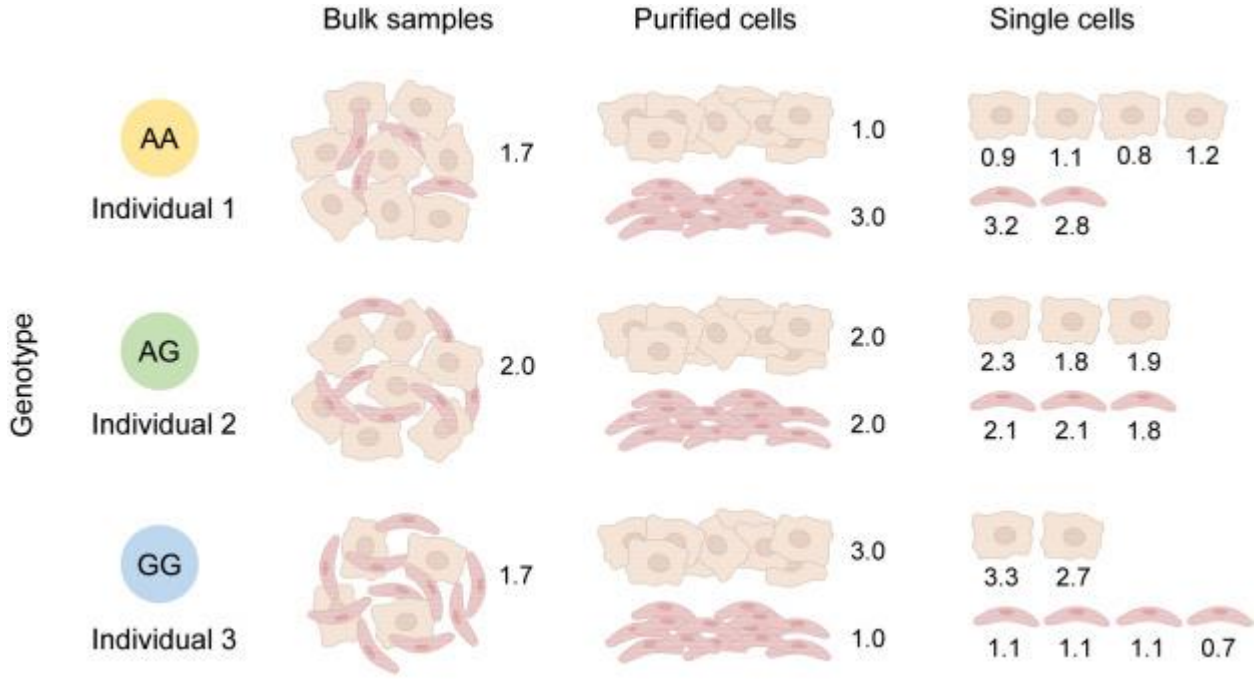
molQTL的步骤

- sample collection
- genotyping & quality control
- molecular phenotype assessment
- mapping & testing
- visualization
- extention



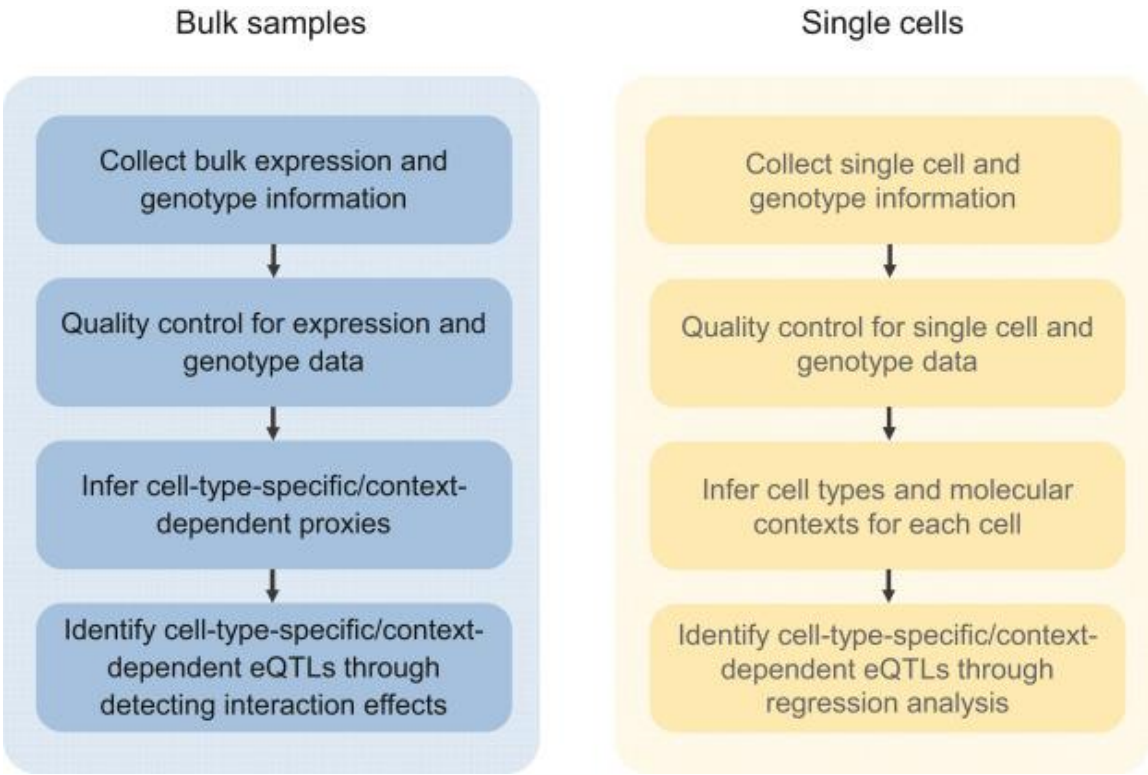
Sample collection

- data collection: selecting the study cohort from which genotype and molecular phenotyping data will be acquired
- resolution：
  - individual
  - tissue
  - single cell
    - cell type
    - cell state
- 样本的人口学特征



Bulk vs single-cell

Samples	Methods	Key ideas	Pros	Cons
Bulk	Westra et al. (2015) Zhernakova et al. (2017) Avila Cobos et al. (2020) Aguirre-Gamboa et al. (2020)	Detect interactions effects between candidate eQTL genotypes and cell-type-specific proxy markers (e.g., cell type proportions) on gene expression levels in bulk tissues	Applicable to large collection of eQTL studies based on bulk samples	Limited resolution for cell types and dependence on informative and robust cell-type-specific proxy markers
Single cells	Cuomo et al. (2022) Strober et al. (2022)	Detect differential effects of candidate eQTL genotypes on gene expression levels for different cell types and/or contexts inferred from single-cell expression data	High-resolution cell types and different molecular contexts	Limited number of subjects available and sparsity in single-cell gene expression data



Database

Table 1 | Notable recent molQTL studies and resources

Study/resource	Type	Number of donors	Population ancestries	Biospecimens	Molecular phenotypes
eQTL Catalogue[76]	Aggregated database of reanalysed data	73–948 per study, total 8,193	88.5% European ancestries	Diverse	Transcriptome phenotypes
GTEx[58]	Consortium with centralized data production and analysis	73–706	American; 85% European and 11% African ancestries	49 postmortem tissues	Gene expression and splicing, others in smaller scale
eQTLGen[5]	Consortium with federated analysis	31,684	Predominantly European	Whole blood	Gene expression
GoDMC[42]	Consortium with federated analysis	32,851	European ancestries	Whole blood	DNA methylation
Hawe et al.[43]	Research project	6,994	European and South Asian ancestries	Whole blood	DNA methylation
Ferkingstad et al.[50]	Single-cohort study	35,559	Icelandic	Plasma	Aptamer proteomics
Jerber et al.[158]	Research project	215	European ancestries	In vitro differentiated iPSCs	scRNA-seq
Yazar et al.[153]	Research project	982	European ancestries	PBMCs	scRNA-seq

iPSC, induced pluripotent stem cell; molQTL, molecular quantitative trait locus; PBMC, peripheral blood mononuclear cell; scRNA-seq, single-cell RNA sequencing.

Table 2 | Commonly used repositories for different molQTL data types

Open access	Controlled access	Summary statistics
European Nucleotide Archive (ENA)	European Genome–Phenome Archive (EGA)	Zenodo
Sequence Read Archive (SRA)	Database of Genotypes and Phenotypes (dbGaP)	Synapse
ArrayExpress (for microarray data)	Synapse (supports both access modes)	eQTL Catalogue (if processed with uniform workflows)
Synapse		

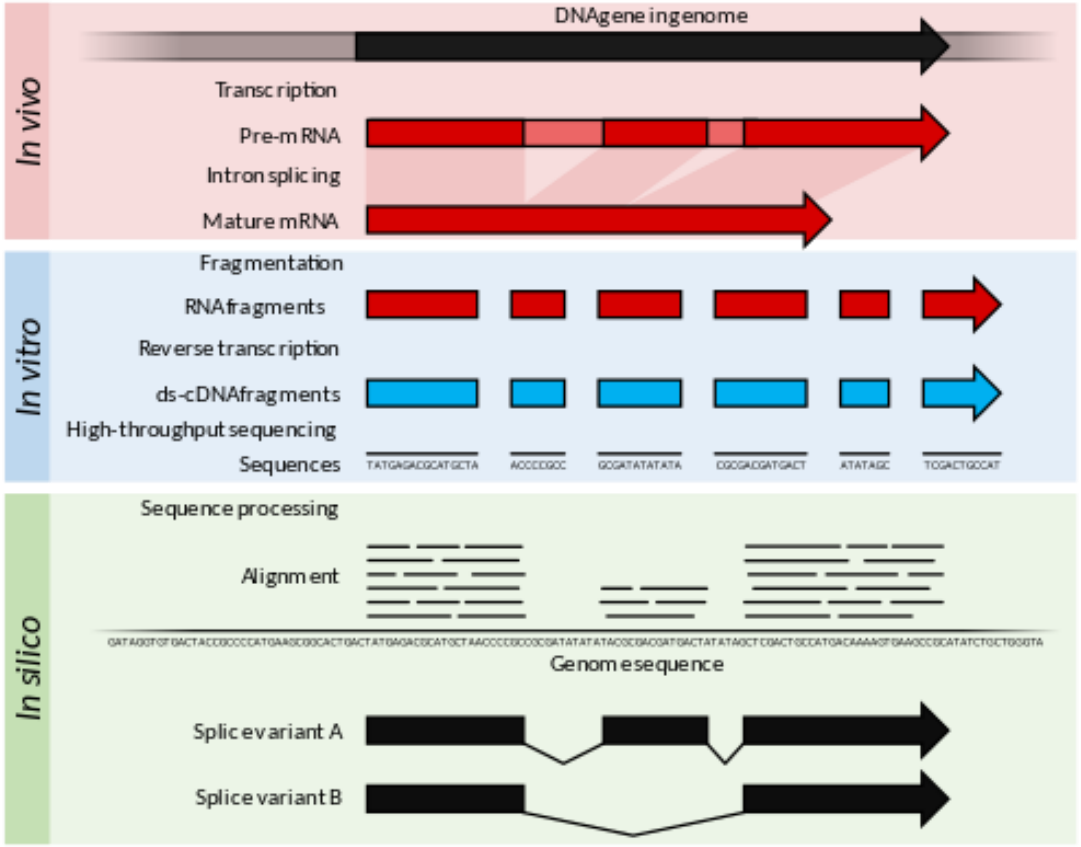
molQTL, molecular quantitative trait locus.

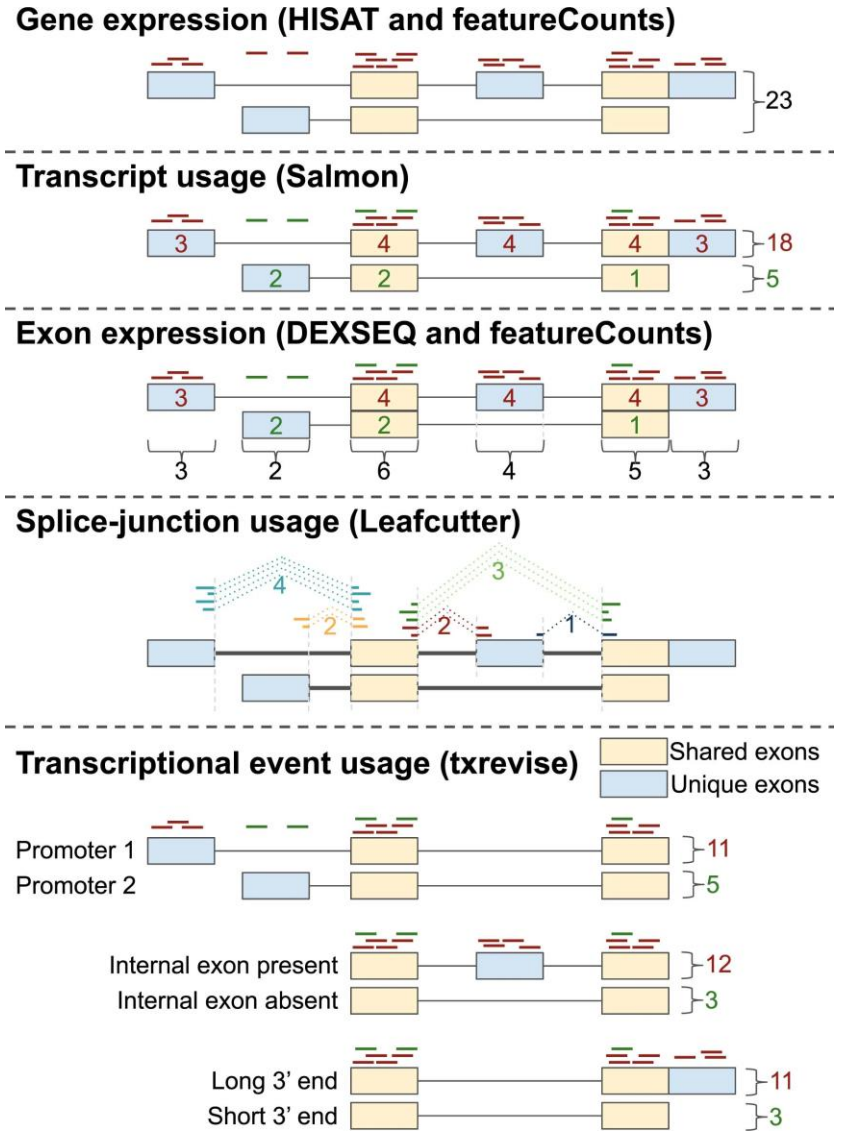
Genotyping

- genotyping:
  - whole-genome sequencing: including non-coding regions, increased power for identifying causal variants, more complex genetic variants (short tandem repeats, short indels, structural variants)
  - exome sequencing / genotype calling from RNA-seq
  - SNP arrays + imputation: common variants
- quality control:
  - exclusion of samples: poor genotyping quality
  - exclusion of variants: low complexity regions of genome, high missingness, failing Hardy-Weinberg equilibrium and imputation quality
- variants in X, Y chromosomes and mitochondrial DNA

Molecular phenotype assessment

- transcriptomic phenotypes: RNA-seq





Molecular phenotype assessment

- other molecular phenotypes
  - methylation of cytosines at CpG sites: arrays
  - chromatin accessibility: DNase-seq, ATAC-seq
  - histone modification: ChIP-seq
  - protein: proteomic assays

Mapping

- Quality control of samples
- Data normalization
- Selection of covariates
- Computation of the association
- FDR control to identify significant association

Quality control

- Quality control to exclude problematic samples
  - RNA-seq: quality of the RNA, sequencing data, alignment and quantification (the fraction of reads originating from exonic regions)
  - RIN (RNA Integrity Number), ROS (RNA Quality Score)
- Sample or label swaps
- Inspected for outlier samples and batch effect

Normalization

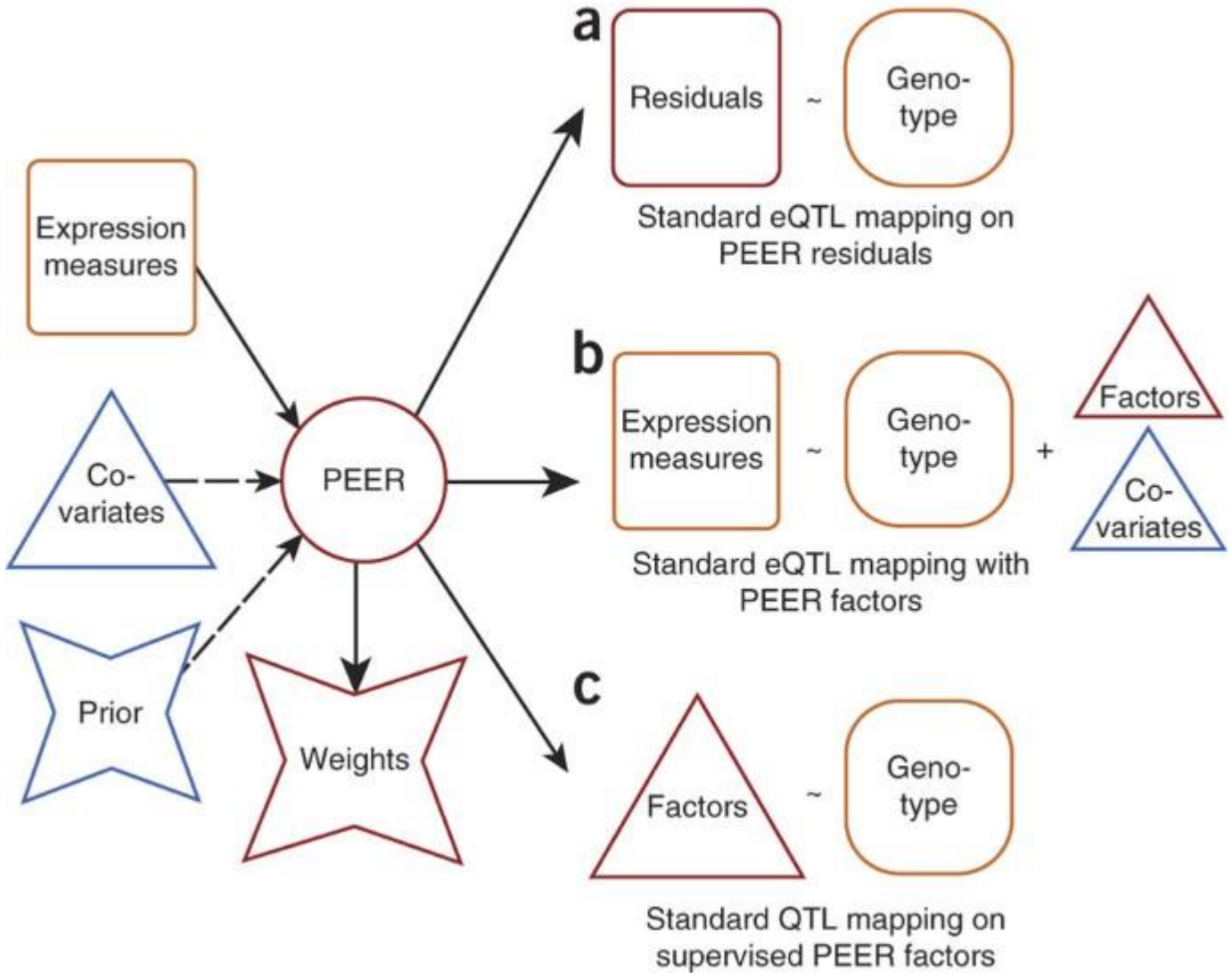
- Read depth: quantile normalization, TMM (trimmed mean of M values), size factor from DESeq
- Feature selection: low or undetectable levels, limited affinity or mapping accuracy
- Phenotypes (molecular features) must be transformed to conform to the assumption of the regression model
  - Homoscedasticity, Gaussian residuals
  - Inverse normal transformation

Confounding & Covariates

- Biological heterogeneity across samples:
  - Differences in cell type composition
  - Technical heterogeneity arising during collection and processing of samples
  - Population stratification
- Population structure
  - principal components derived from genotype data
- Latent variables computed from the normalized phenotypes
  - Principal components or probabilistic estimation of expression residuals (PEER) factors computed from the molecular data

PEER

- PEER stands for "probabilistic estimation of expression residuals". It is a collection of Bayesian approaches to infer hidden determinants and their effects from gene expression profiles using factor analysis methods.



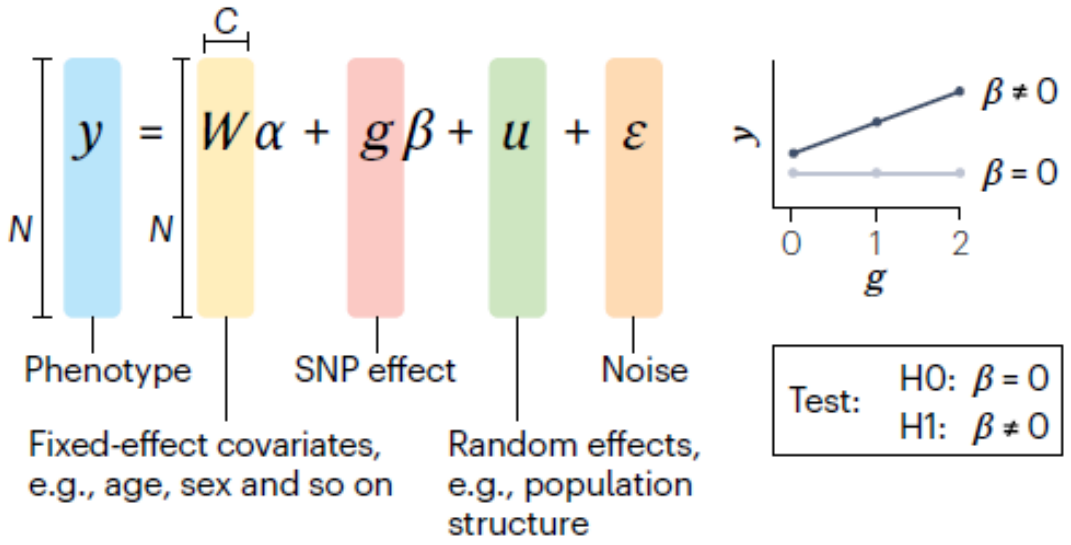
Association

- Bulk: $y_{ng} = \beta_g + \beta_{gs} x_{ns} + \varepsilon_{ngs},$
  - the errors are assumed to be Gaussian, is reasonable for microarray-based gene expression measurements.
  - with gene expression data collected through RNA sequencing, the measured gene expression level is the total number of sequence reads mapped to a specific gene, which needs to be adjusted for total sequencing depth and other factors. These data may be better modeled by other distributions, for example, negative binomial, while accounting for factors that may impact the observed sequencing reads.
  - allelic-specific expression to identify cis-eQTLs: TReCASE, RASQUAL, and mixQTL. (杂合子两条染色体转录本的不平衡性)
  - context-dependent eQTLs: sex-biased eQTLs, population-biased eQTLs, including an interaction term between the context variable and the SNP genotype in the regression model

association

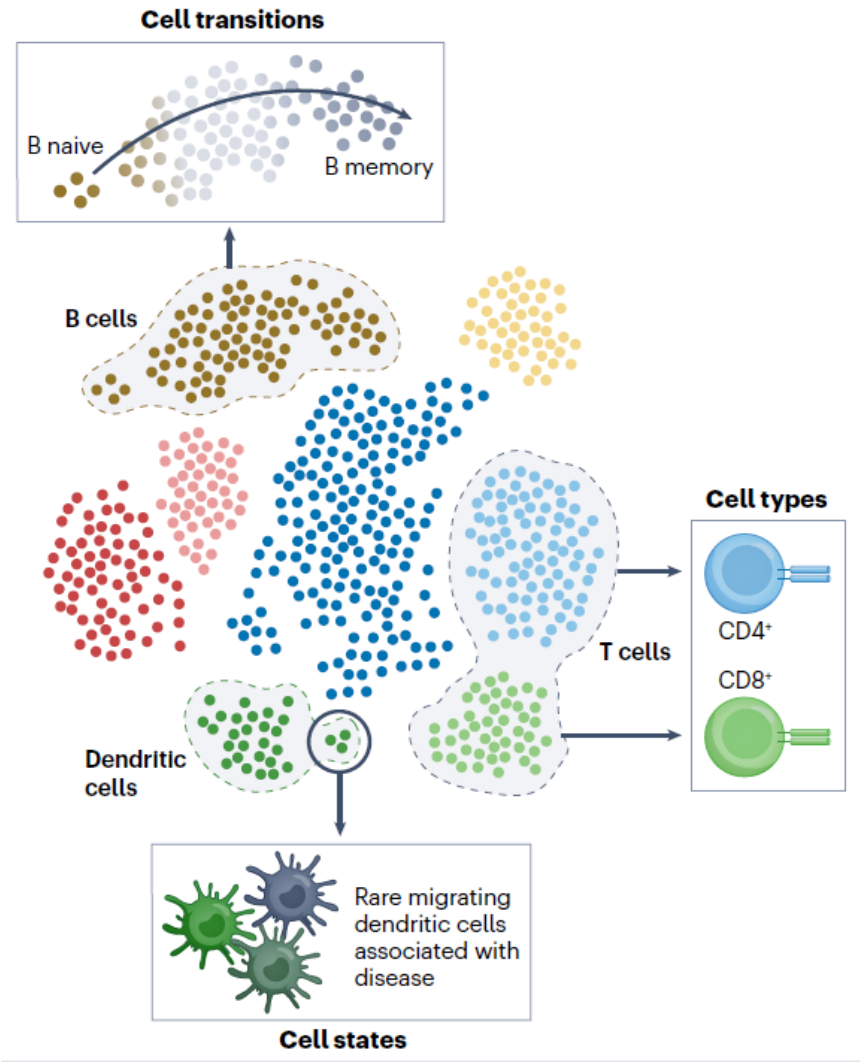
- covariates and confounding:
  - add covariates
  - linear mixed-effect model:

Molecular trait ~ Genotype + PEERs + PCs

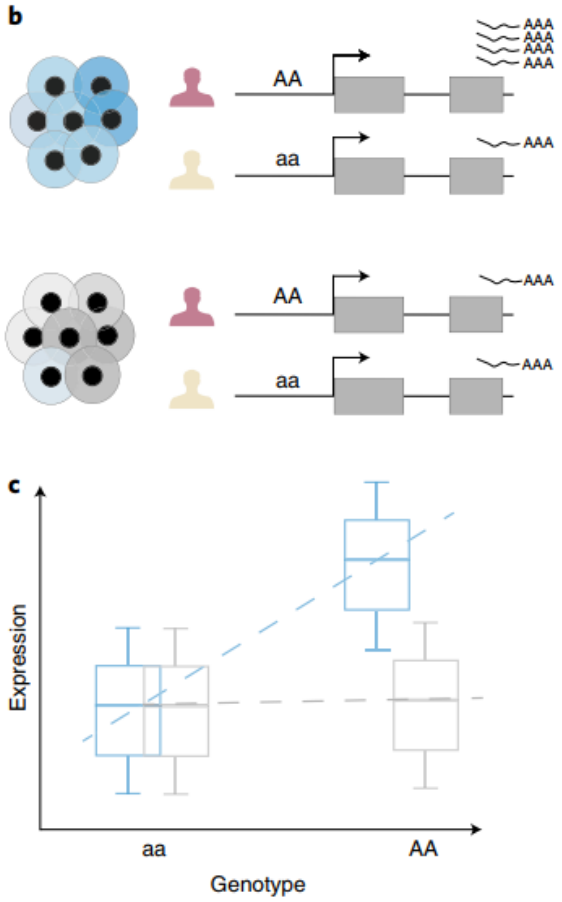


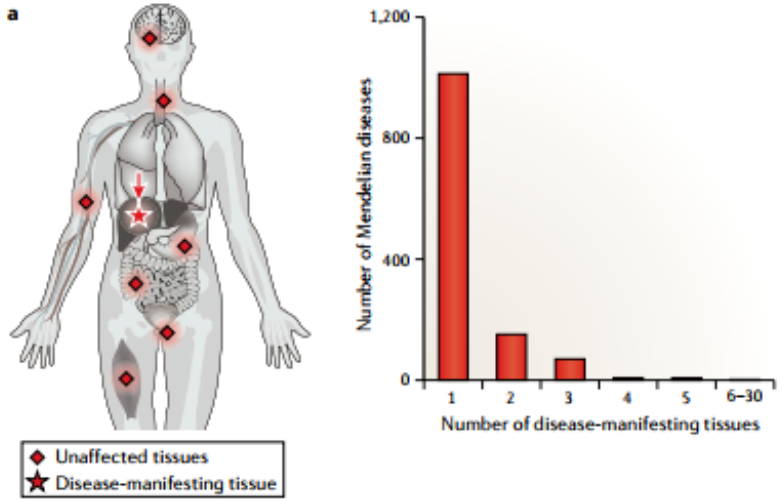
Cell type & cell state

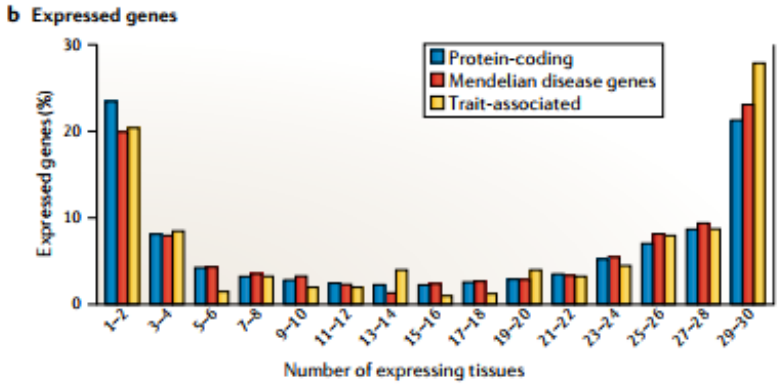
- cell type: groups of cells from distinct, irreversible developmental lineages. → discrete
- cell state: functionally specialized, often plastic, subpopulations of cells. These states can be discrete (for example, T helper cells) or continuous (for example, developmental states).

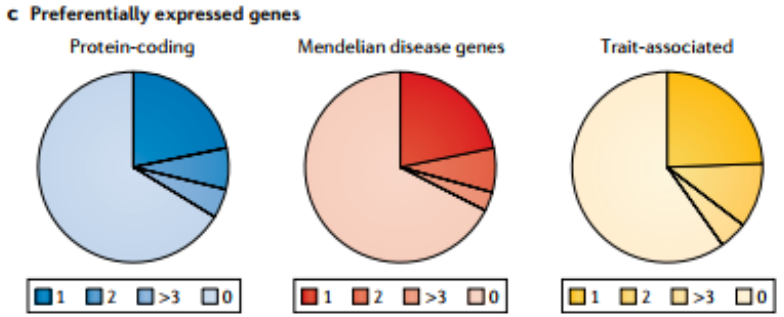


Cell-type specific eQTL









Association

- Cell-type-specific eQTL (ct-eQTLs) with bulk data:
  - identifying ct-eQTLs by investigating whether there is an interaction effect between the surrogate score for a cell type and candidate SNP's genotype on bulk gene expression levels from the collected samples.

$$y_{ng} = \beta_g + \beta_{gs} x_{ns} + \beta_{gm} m_n + \beta_{g,sm} (x_{ns} \times m_n) + \varepsilon_{ngs},$$

m_n is a proxy marker for the cell type of interest in the nth individual

  - Instead of deriving cell-type-specific proxy markers or enrichment scores, the estimated cell type proportions can also be used as a proxy for a given cell type.

$$y_{ng} = \beta_g + \beta_{gs} x_{ns} + \beta_{gk} \pi_{nk} + \beta_{g,sk} (x_{ns} \times \pi_{nk}) + \varepsilon_{ngs}.$$

π_{nk} denotes the estimated proportion of the kth cell type for this individual, where there is a total of K cell types

Association

- Cell-type-specific eQTL (ct-eQTLs) with bulk data:
  - takes into account all cell types simultaneously

$$y_{ng} = \beta_g + \beta_{gs} x_{ns} + \sum_{k=1}^{K} \beta_{gk} \pi_{nk} + x_{ns} \left(\sum_{k=1}^{K} \beta_{g,sk} \times \pi_{nk} \right) + \varepsilon_{ng},$$

  - Another way to parametrize this model

$$y_{ng} = \sum_{k=1}^{K} \left(\beta_{gk} + \beta_{g,sk} \times x_{nk} \right) \pi_{nk} + \varepsilon_{ng}.$$

Association

- Cell-type-specific eQTLs with single-cell data
  - Pseudo-bulk (aggregation): single-cell data are first annotated to distinct cell types, and the cells annotated to the same cell types from a specific subject are combined to derive cell-type-specific gene expression levels. eQTL methods for bulk samples can then be applied to detect ct-eQTLs.
  - Single-cell individually: Poisson mixed effects regression to model the effects of SNPs, cell states (which can be both discrete and continuous), batch structure, and other covariates (such as sex, age, genotype principal components and gene expression principal components, and percentage of mitochondrial unique molecular identifiers (UMIs)) on the observed gene expression level measured by UMI counts at the single-cell level.

Association

- More
  - CellRegMap

  we have *N* subjects, with m_n cells collected from the *n*th subject, and a total of *C* different cellular contexts are defined for each cell.

$$y_{ngi} = \beta_g + \beta_{gs} x_{ns} + \beta_{g,si} x_{ns} + u_{ng} + c_{ngi} + \varepsilon_{ngi},$$

$$\beta_{g,si} \sim N\left(0, \sigma^2_{S\times C}\Sigma\right), u_{ng} \sim N\left(0, \sigma^2_R\Sigma\right), c_{ngi} \sim N\left(0, \sigma^2_C\Sigma\right) \qquad \varepsilon_{ngi} \sim N\left(0, \sigma^2_e\right)$$

  - single-cell unsupervised regulation of gene expression (SURGE)

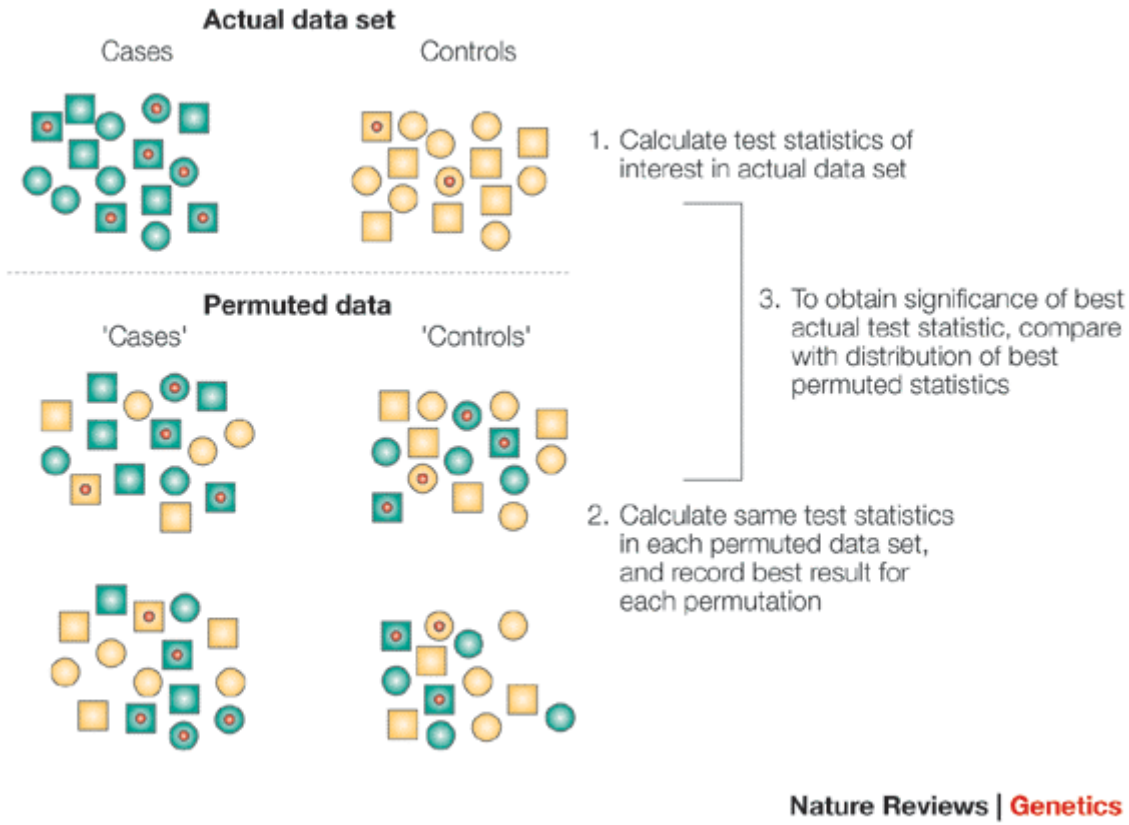
$$y_{ngi} = \beta_g + \beta_{gs} z_{ns} + \sum_{c=1}^{C} h_{nic}\beta_{gsc} z_{ns} + u_{ng} + \varepsilon_{ngi},$$

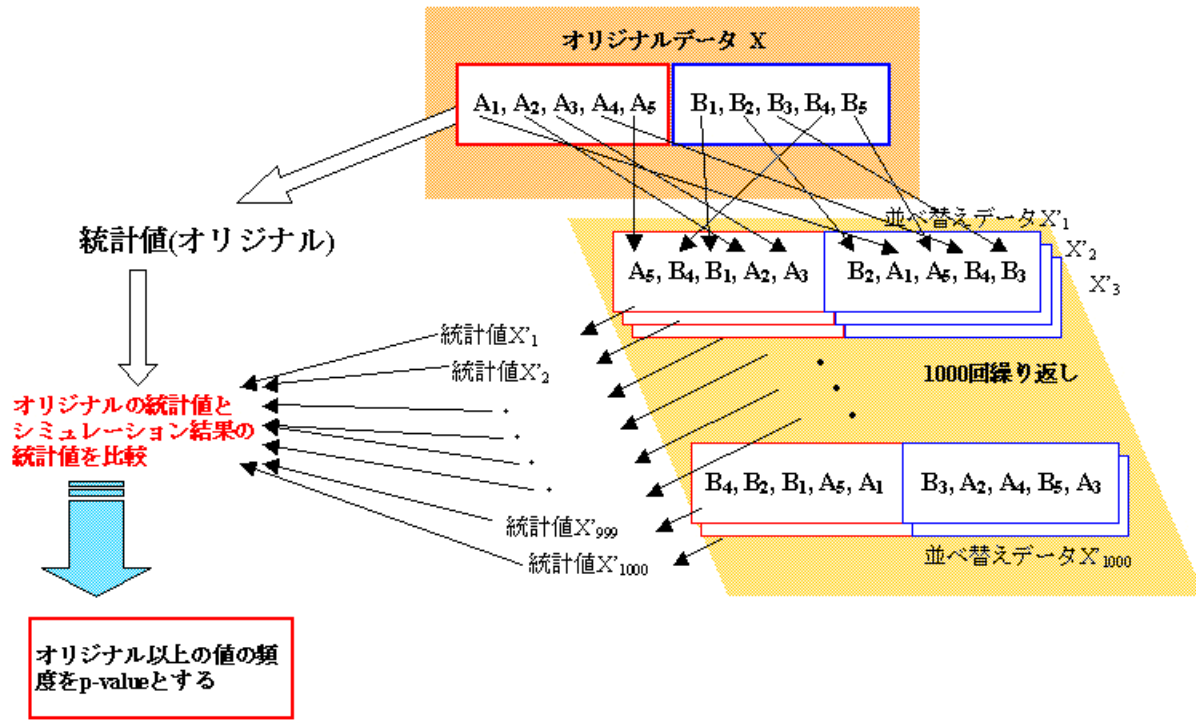
Multiple test correction: FDR control

Two types of multiple test correction
- Bonferroni adjustment (GWAS)
- FDR control (eQTL)

- Two levels:
  - Multiple variants per molecular trait
  - Multiple molecular traits cross the genome
- Permutation test
  - the permutations determine the probability of observing the lead association observed for a feature by chance among all variants tested for the feature, accounting for the first layer of multiple testing correction.
- FDR: false discovery rate
  - Storey q values
  - Benjamini-Hochberg

Permutation test





Rethinking *P* values

- Please mark each of the statements below as “true” or “false”.

  我研究两个变量之间是否存在效应（effect），以没有effect作为零假设H_0，得到了一个$P<0.05$的检验结果，那么我能得到以下哪些结论：

  1. 零假设成立的概率小于0.05
  2. 观察到的效应（effect）仅仅是由随机性产生的概率小于0.05
  3. 在零假设成立时，得到当前的观测样本的概率小于0.05
  4. 这两个变量之间存在效应的概率大于0.95
  5. 如果我拒绝原假设，犯错（I型错误）的概率小于0.05
  6. 我如果重复这项研究，得到同样结果（$P<0.05$）的概率大于0.95

以上说法都不对！

Rethinking *P* values

- The hybrid of two theories: Fisher & Neyman-Pearson

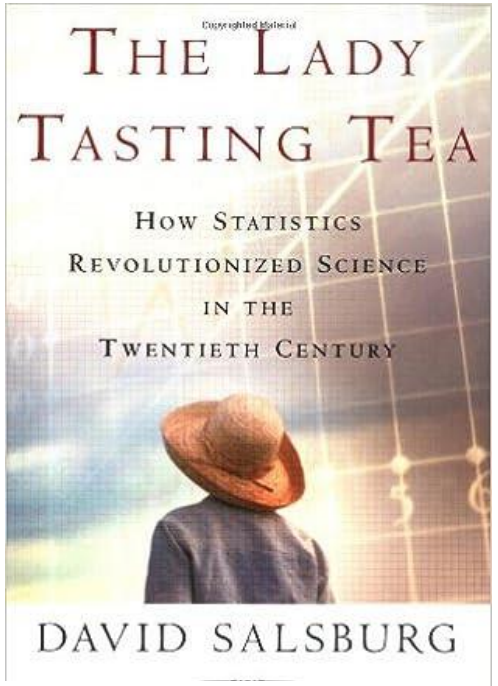
For all the *P* value's apparent precision, Fisher intended it to be just one part of a fluid, non-numerical process that blended data and background knowledge to lead to scientific conclusions. But it soon got swept into a movement to make evidence-based decision-making as rigorous and objective as possible. This movement was spearheaded in the late 1920s by Fisher's bitter rivals, Polish mathematician Jerzy Neyman and UK statistician Egon Pearson, who introduced an alternative framework for data analysis that included statistical power, false positives, false negatives and many other concepts now familiar from introductory statistics classes. They pointedly left out the *P* value.

But while the rivals feuded — Neyman called some of Fisher's work mathematically "worse than useless"; Fisher called Neyman's approach "childish" and "horrifying [for] intellectual freedom in the west" — other researchers lost patience and began to write statistics manuals for working scientists. And because many of the authors were non-statisticians without a thorough understanding of either approach, they created a hybrid system that crammed Fisher's easy-to-calculate *P* value into Neyman and Pearson's reassuringly rigorous rule-based system. This is when a *P* value of 0.05 became enshrined as 'statistically significant', for example. "The *P* value was never meant to be used the way it's used today," says Goodman.

Nuzzo, R. Scientific method: Statistical errors. *Nature* **506**, 150–152 (2014). https://doi.org/10.1038/506150a

The lady tasting tea

- Fisher 给一位名叫 Muriel Bristol 的女士倒了一杯茶，这位女士号称能够分辨先倒茶和先倒牛奶的区别。Fisher 当然想用实验检验一下：这位女士的味觉是否有这么敏锐？Fisher 倒了 8 杯奶茶：其中 4 杯“先奶后茶”，其余 4 杯“先茶后奶”。随机打乱次序后，Fisher 请 Bristol 品尝，并选出“先奶后茶”的 4 杯，看她是否能分辨奶和茶的顺序。



	Bristol “先奶后茶”	Bristol “先茶后奶”	总数
Fisher “先奶后茶”	k	$4-k$	4
Fisher “先茶后奶”	$4-k$	k	4
总数	4	4	8

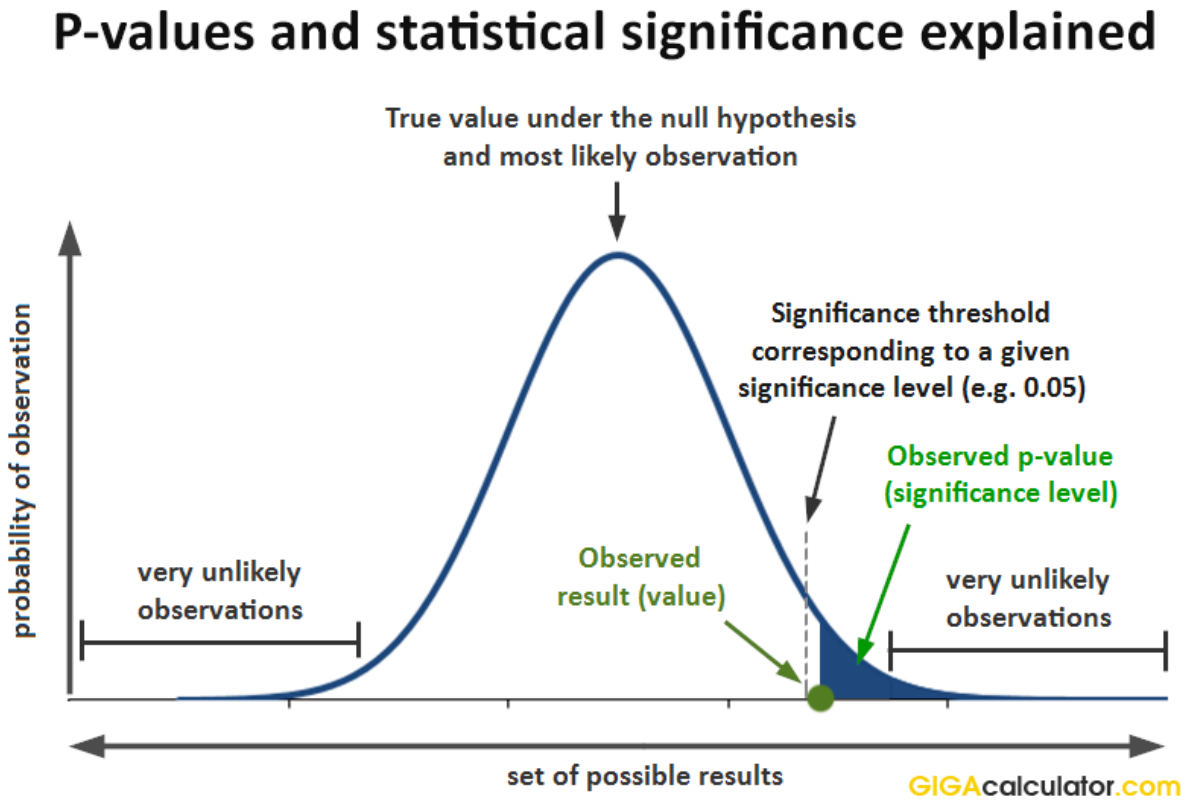
各种正确次数对应的组合数。

正确次数	组合数
0正确	$\binom{4}{0}\times\binom{4}{4-0}=\frac{4!}{4!\times 0!}\times\frac{4!}{0!\times 4!}=1$
1正确	$\binom{4}{1}\times\binom{4}{4-1}=\frac{4!}{3!\times 1!}\times\frac{4!}{1!\times 3!}=16$
2正确	$\binom{4}{2}\times\binom{4}{4-2}=\frac{4!}{2!\times 2!}\times\frac{4!}{2!\times 2!}=36$
3正确	$\binom{4}{3}\times\binom{4}{4-3}=\frac{4!}{1!\times 3!}\times\frac{4!}{3!\times 1!}=16$
4正确	$\binom{4}{4}\times\binom{4}{4-4}=\frac{4!}{0!\times 4!}\times\frac{4!}{4!\times 0!}=1$
总和	70

Fisher's significance test

1. Identify the null hypothesis.
2. Determine the appropriate test statistic and its distribution under the assumption that the null hypothesis is true.
3. Calculate the test statistic from the data.
4. Determine the achieved significance level that corresponds to the test statistic using the distribution under the assumption that the null is true.
5. Reject H_0 if the achieved significance level is sufficiently small. Otherwise reach no conclusion.

- Null hypothesis only
- In null-hypothesis significance testing, the *p*-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct. $p = \Pr(T(X) \geq T(x) \,|\mathrm{H}_0),)$
- *p*-value is the minimal significance level that would result in a rejection of the null hypothesis.



Fisher's significance test

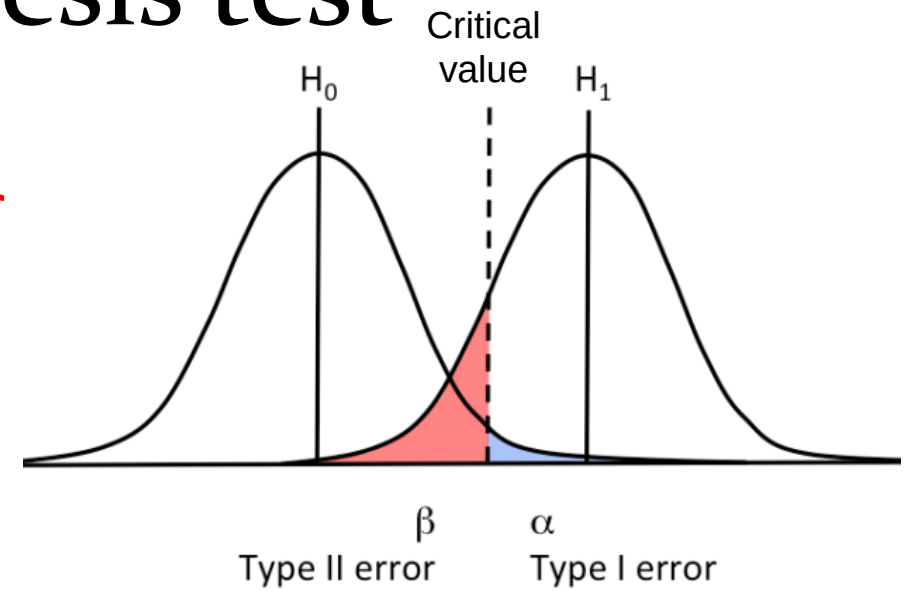
- The *p*-value is a measure of (im)plausibility of observed as well as unobserved more extreme results, assuming a true null hypothesis.
- A very small *p*-value:
  - either an exceptionally rare chance has occurred, or the theory of random distribution [H_0] is not true [i.e., strong evidence against H_0]
- A large *p*-value:
  - a significant result provides evidence against H_0, whereas a non-significant result simply suspends judgment—nothing can be said about H_0.
- In Fisher's view, the *p*-value is an epistemic measure of evidence from a single experiment and not a long-run error probability, and he also stressed that 'significance' depends strongly on the context of the experiment and whether prior knowledge about the phenomenon under study is available. (**quasi-Bayesian interpretation**)

Fisher's significance test: *P* values

- *P*-values can indicate **how incompatible the data are with a specified statistical model**.
- *P*-values **do not** measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- A *p*-value, or statistical significance, **does not** measure the size of an effect or the importance of a result.
- A relatively large *p*-value **does not** imply evidence in favor of the null hypothesis.

Neyman-Pearson hypothesis test

- Two competing hypotheses: Type I error & type II error
- Confidence level & Significance Levels: control the type I error
- Power function



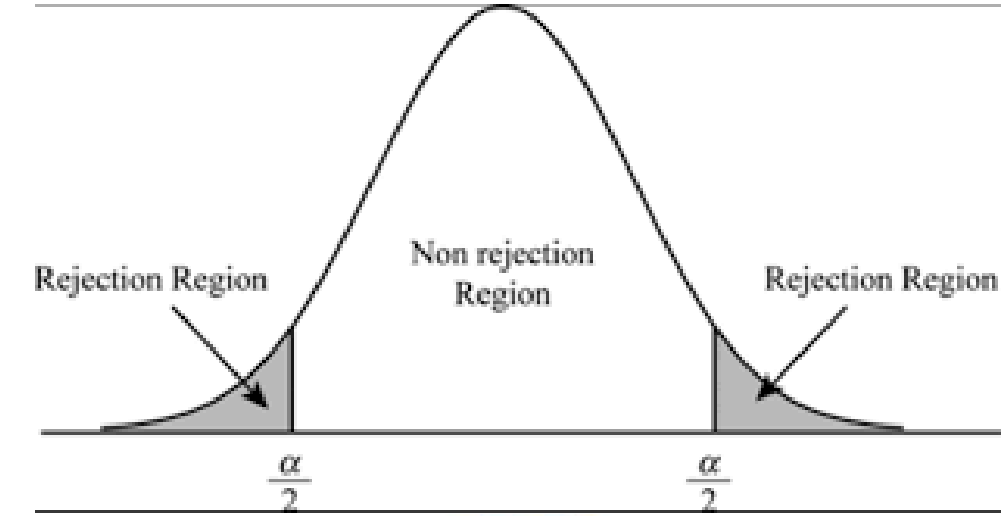
1. Identify a hypothesis of interest, Θ_B, and a complementary hypothesis, Θ_A.
2. Determine the appropriate test statistic and its distribution under the assumption that Θ_A is true.
3. Specify a significance level (α), and determine the corresponding critical value of the test statistic under the assumption that Θ_A is true.
4. Calculate the test statistic from the data.
5. Reject Θ_A and accept Θ_B if the test statistic is further than the critical value from the expected value of the test statistic (calculated under the assumption that Θ_A is true). Otherwise accept Θ_A.

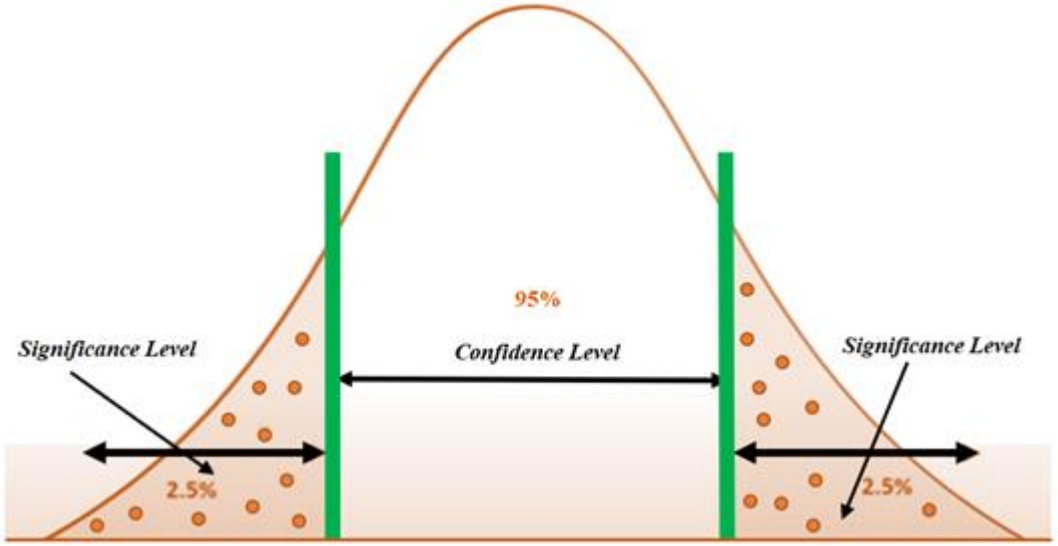
Study findings	Truth	
	Null hypothesis is true	Null hypothesis is false
Null hypothesis is not rejected	True negative	Type II error (β) (false negative)
Null hypothesis is rejected	Type I error (α) (false positive)	True positive

α and β represent the probability of Types I and II errors, respectively.

Neyman-Pearson hypothesis test

- Two competing hypotheses: Type I error & type II error
- Confidence level & Significance Levels: control the type I error
- Power function





Neyman-Pearson hypothesis test

- Two competing hypotheses: Type I error & type II error
- Confidence level & Significance Levels: control the type I error
- Power function

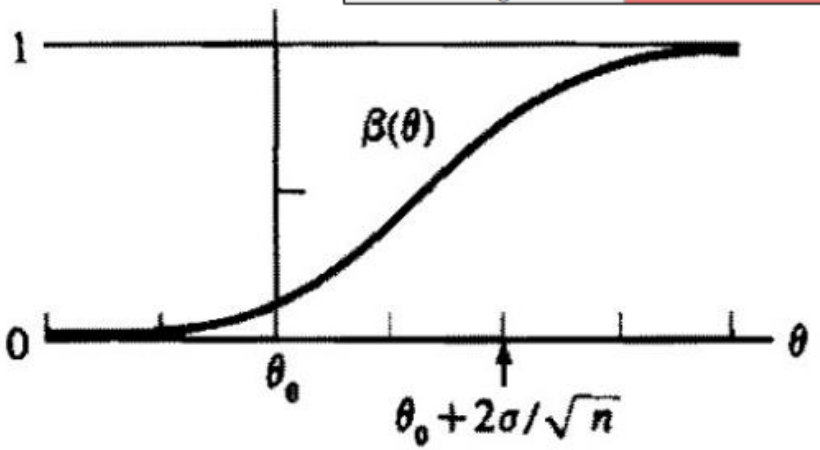
$$H_0 = \{P_\theta : \theta \in \Theta_0 \subset \Theta\} \qquad H_1 = \{P_\theta : \theta \in \Theta_1 = \Theta \setminus \Theta_0\}$$

The power function of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(X \in R)$

$$P_\theta(X \in R) = \begin{cases} \text{probability of a Type 1 error} & \text{if } \theta \in \Theta_0 \\ \text{one minus the probability of a Type 2 error} & \text{if } \theta \in \Theta_0^c \end{cases}$$

	True state of the world	
Decision	H_0 is true	H_0 is false
Fail to reject H_0	0.95	0.67179
Reject H_0	0.05	0.32821

$$\begin{aligned} \beta(\theta) &= P_\theta\left(\frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}} > c\right) \\ &= P_\theta\left(\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z > c + \frac{\theta_0 - \theta}{\sigma/\sqrt{n}}\right) \end{aligned}$$



Z test

Neyman-Pearson hypothesis test

- Neyman–Pearson's model is only about rules of behavior in the long-run, so that "we shall reject H when it is true not more than say, once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false"
- Accordingly, 'hypothesis tests' are concerned with minimizing Type II errors subject to a bound on Type I errors, and α is a prescription for 'inductive behaviors' and not evidence for a specific result.
- error control is a pre-selected fixed measure; α is therefore a rigidly fixed level, not a random variable based on the actual data, and α applies only to infinitely random selections from the same finite population, not to an actual result in a single experiment.

Two theories

- the underlying philosophy and the interpretation of the [Fisher versus N–P] results is profoundly different

Points of contrast	Fisher approach	Neyman–Pearson approach
Arrangement of hypotheses	A single *null hypothesis* (Fisher's term) is formulated. This hypothesis serves for the conceptual interpretation of experimental results (e.g., in terms of group differences) and the mathematical specification of the distribution ("population") under which the experimental data are assessed.	A dichotomous decision-making situation is postulated in which the preferred course of action is contingent on the (unknown) distribution of an observed variable. Two *alternative hypotheses* (Neyman and Pearson's term) are formulated in correspondence with the actions.
Testing procedure	A *test of significance* (Fisher's term) is applied to evaluate the discrepancy between the observed data and the null hypothesis. If the probability of the data under the hypothesis, *P{D|H}*, is sufficiently small (e.g., $p < .05$), the hypothesis is rejected. If *P{D|H}* is not small enough, the hypothesis is not rejected (but also not accepted).	A test of statistical hypotheses or *rule of inductive behavior* (Neyman's term) is applied to one of the hypotheses. If P{D|H} is sufficiently small (i.e., $< \alpha$), then the tested hypothesis is rejected and the alternative hypothesis is accepted by implication. If *P{D|H}* is $> \alpha$, then the tested hypothesis is accepted and the alternative rejected.
Interpretation of outcome	Fisher proposed that the outcome of a successful test of significance can be interpreted in terms of the following disjunction: Either the null hypothesis is false, or an unlikely event has occurred. If it is concluded that the hypothesis is false, the corresponding substantive interpretation is that the experiment has demonstrated a positive result (e.g., a difference between groups). Fisher's interpretation of nonsignificant outcomes is ambiguous.	Neyman and Pearson did not interpret the outcome of a test epistemically but in terms of the relative frequency of errors in the long term, or *Type I error* (α) and *Type II error* ($1 - \beta$), where β designates the *statistical power* of a test (Neyman and Pearson's terms). The substantive interpretation of the test is to adopt one specified course of action or the other, corresponding to which hypothesis has been accepted (i.e., a decision).

Two theories

'Significance test' (R. A. Fisher)	'Hypothesis test' (Neyman and Pearson)
p value—a measure of the evidence against H_0	α and β levels—provide rules to limit the proportion of decision errors
Calculated *a posteriori* from the observed data (random variable)	Fixed values, determined *a priori* at some specified level
Applies to any single experiment (short run)	Applies only to ongoing, identical repetitions of an experiment, not to any single experiment (long-run)
Roots in inductive philosophy: from particular to general	Roots in deductive philosophy: from general to particular
'Inductive inference': guidelines for interpreting strength of evidence in data (subjective decisions)	'Inductive behavior': guidelines for making decisions based on data (objective behavior)
Based on the concept of a 'hypothetical infinite population'	Based on a clearly defined population
Evidential, i.e., based on the evidence observed	Non-evidential, i.e., based on a rule of behavior

It is often overlooked that 'significance tests' as well as 'hypotheses tests' were specifically developed for controlled experimental settings (like in Fisher's case agricultural research), and *not* studies based on observational data. Paramount to experimental settings and frequentist tests is randomization (i.e., random assignment and probability sampling).

Confusion: "significance level"

Table 1. Differences between *p*-values and α levels as measures of 'statistical significance'.

P-values	α levels
Fisher's significance level	Neyman–Pearson's significance level
Inductive philosophy – from particular to general	Deductive philosophy—from general to particular
Only the null hypothesis, H_0	Null hypothesis, H_0, and alternative hypothesis, H_A
Empirical evidence against H_0	Type I error—erroneous rejection of H_0
Inductive inference –framework for evaluating strength of evidence in data	Inductive behaviour –prescriptions for making decisions between H_0 and H_A based on data
Data-dependent random variable with uniform distribution over the interval [0–1] under the null hypothesis	Predetermined fixed value
Characteristic of data	Characteristic of test
Power of test only implicit	Power of test plays central role
Short-run – applicable to each specific study	Long-run – applicable only to ongoing, identical repetitions of original study, not to each specific study

A typical misuse: 'roving alphas'

- the Type I error is a conditional probability which can be written as $\alpha = p(\text{reject } H_o | H_{o,\text{ true}})$. It is a pre-selected fixed measure that applies only to infinitely random selections from the same finite population, and not to an actual result in a single experiment.
- *p* values are conditional probabilities of the data, so they do not apply to any specific decision to reject H_o because any particular decision to do so is either right or wrong (the probability is either 1 or 0). Only with sufficient replication could one determine whether a decision to reject H_o in a particular study was correct.

Debate

- Neyman–Pearson's model is considered to be theoretically consistent and is generally accepted as 'frequentist orthodoxy' in mathematical statistics. However, the emphasis upon decision rules with stated error rates in infinitely repeated trials may be applicable to quality control in industrial settings, but seems less relevant to assessment of scientific hypotheses (Fisher, 1955).
- On the other hand, the supposed 'objective' evidential nature of p values was also questioned early on, especially the fact that p values only test one hypothesis and are based on tail area probabilities was early on considered a serious deficiency (Jeffreys, 1961).
  - Dependence on tail area probabilities means that the calculation of p values is not only based on the observed results but also on 'more extreme results', i.e., results that have not occurred.
  - 'stopping rule' paradox: what is 'more extreme results' depends on the actual sampling plan in a study

ASA Statement on *p*-values (2016)

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Multiple test burden

- Multiple test: family-wise error rate (FWER) — the probability of at least one type I error

$$\bar{\alpha} = 1 - \left(1 - \alpha_{\{\text{per comparison}\}}\right)^m.$$

- Boole's inequality

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

- Bonferroni correction

$$\alpha_{\{\text{per comparison}\}} = \alpha/m.$$

- False discovery rate (FDR)

$$\text{FDR} = E(V/R).$$

	Null hypothesis is true (H_0)	Alternative hypothesis is true (H_A)	Total
Test is declared significant	V	S	R
Test is declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

- m is the total number hypotheses tested
- m_0 is the number of true null hypotheses, an unknown parameter
- $m - m_0$ is the number of true alternative hypotheses
- V is the number of false positives (Type I error) (also called "false discoveries")
- S is the number of true positives (also called "true discoveries")
- T is the number of false negatives (Type II error)
- U is the number of true negatives
- $R = V + S$ is the number of rejected null hypotheses (also called "discoveries", either true or false)

The distribution of *P*-values

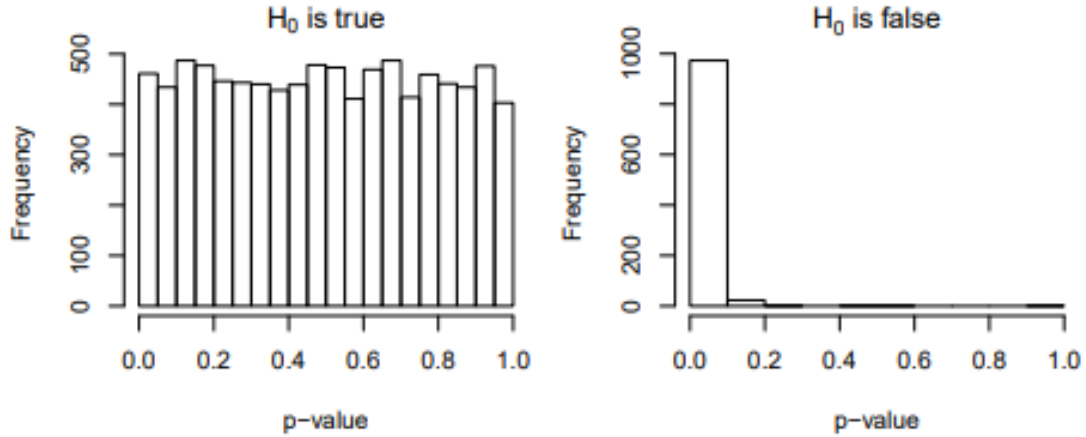
- *P*-values are random variables
- Under null hypothesis: uniform distribution $\mathcal{U}(0,1)$

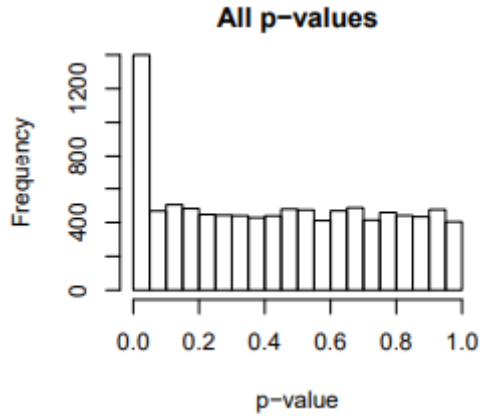
consider a left-sided one-tailed hypothesis test,

$F_T(t)$ is the cumulative distribution function of the test statistic under the null hypothesis.

$$
\begin{aligned}
P &= F_T(T) \\
p &= F_T(t_{\text{obs}}) \\
F_P(p) &= \Pr(P < p) \\
&= \Pr(F_T(T) < p) \\
&= \Pr(T < F_T^{-1}(p)) \\
&= F_T(F_T^{-1}(p)) \\
&= p
\end{aligned}
$$

- When H_0 is false: a good T will tend to be larger under H1, so p will be smaller.





Benjamini-Hochberg

- 将p-value从小到大排好序后，选定前L个满足 $p_L \leq L\frac{\alpha}{M}$的点：$\alpha$是FDR阈值，M是总检验次数

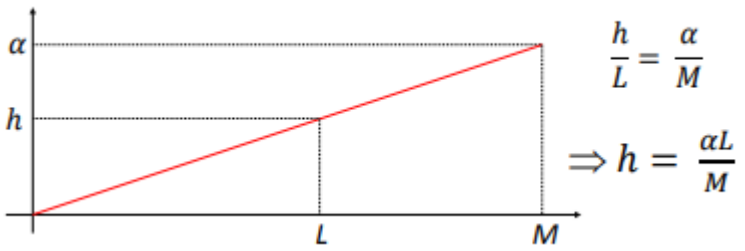
$$FDR = \frac{M_0 h}{L} \leq \frac{M_0 \alpha L}{ML} = \frac{M_0}{M}\alpha \leq \alpha$$

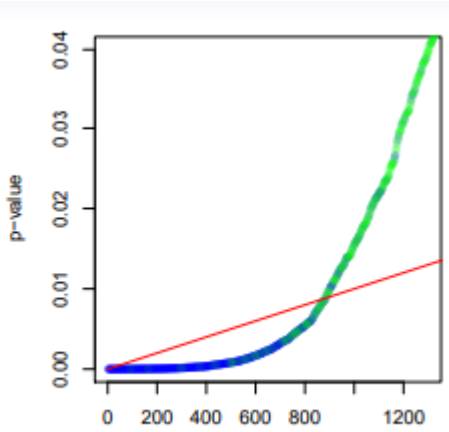
H_0为真时，p值服从[0,1]均匀分布，M_0个H_0为真的检验，落在$[0,h]$内的p值数量为$M_0 h$。现在拒绝掉L个p值最小的检验，对应的最大p值为$h = p_L$，则h不超过$\frac{\alpha L}{M}$。

Benjamini-Hochberg Method

To control FDR $\leq \alpha$:

1. Let $p_{(1)} \leq \cdots \leq p_{(M)}$ be ordered p-values.
2. Define $L = \max\{j : p_{(j)} < \alpha j/M\}$.
3. Reject all hypotheses H_{0j} for which $p_j \leq p_{(L)}$.





Storey q values

- the q-value in the Storey-Tibshirani procedure provides a means to control the positive false discovery rate (pFDR).

$FDR = E(V/R \mid R>0)P(R>0)$ $pFDR = E(V/R \mid R > 0)$

- The q-value is defined as the minimum pFDR at which the feature can be called significant.

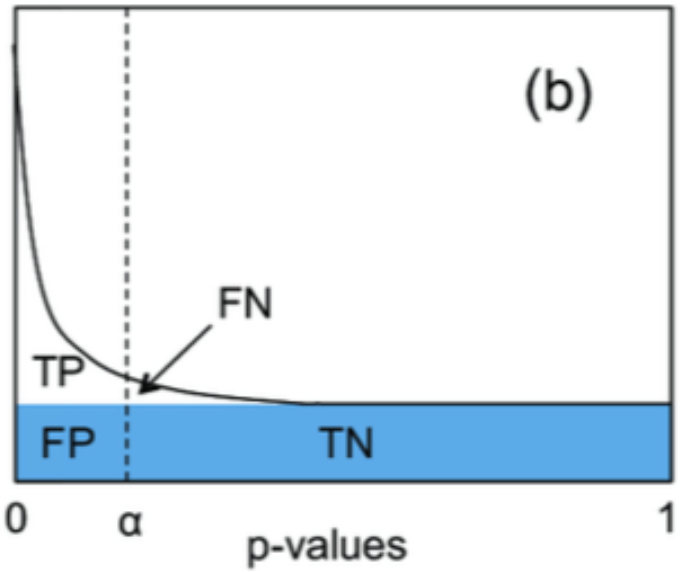
Large m: $P(R > 0) \approx 1, pFDR \approx FDR \approx \frac{E[V]}{E[R]}$

sig阈值为t时：$E[R] = \#\{p_i \le t\}; E[V] = m_0 t = m\pi_0 t$

Estimate $\pi_0 = \frac{m_0}{m}$ $\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, \ldots, m\}}{m(1-\lambda)}$,

$\widehat{\mathrm{FDR}}(t) = \frac{\hat{\pi}_0 m \cdot t}{\#\{p_i \le t\}}$. $\hat{q}(p_i) = \min_{t \ge p_i} \widehat{\mathrm{FDR}}(t)$.

观测到大于λ的p值个数/全部m个检验都是H_0成立时大于λ的p值个数



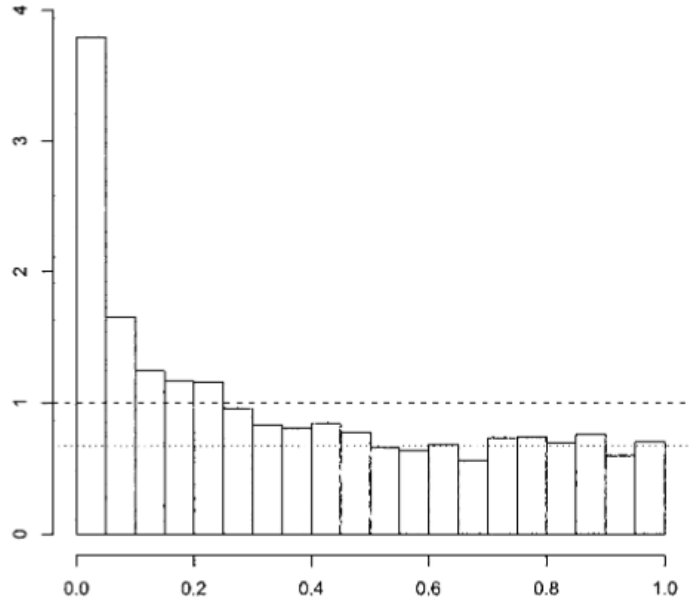


Fig. 1. A density histogram of the 3,170 *p* values from the Hedenfalk *et al.* (14) data. The dashed line is the density histogram we would expect if all genes were null (not differentially expressed). The dotted line is at the height of our estimate of the proportion of null *p* values.

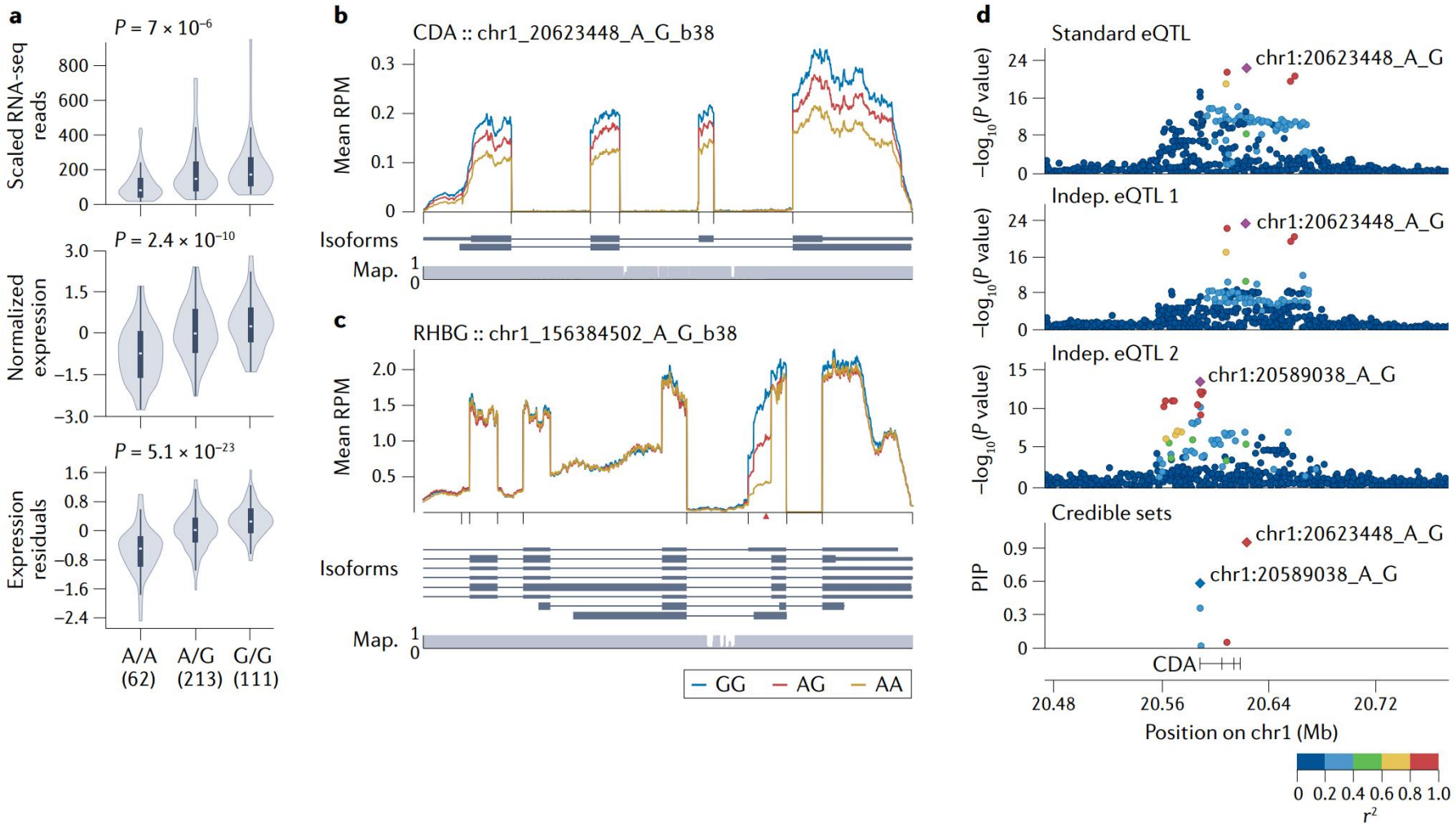
Notes

- MAF threshold: 低MAF会增加假阳性风险
- trans-QTL: 更容易出现artefacts
  - Sequence similarity between a cis-QTL feature and another feature in trans (mis-mapping of reads)
  - More smaller and more cell type-specific effects than cis-QTLs
- Effect size estimation
  - Slope, SVE, allelic fold change (aFC) [携带替代eVariant等位基因的单倍型与携带参考等位基因的单倍型的表达量之间的对数比（log-ratio），单位为log2]
- Sample size
- Horizontal pleiotropy, haplotypic effect: variants in LD for cis-QTL

Software

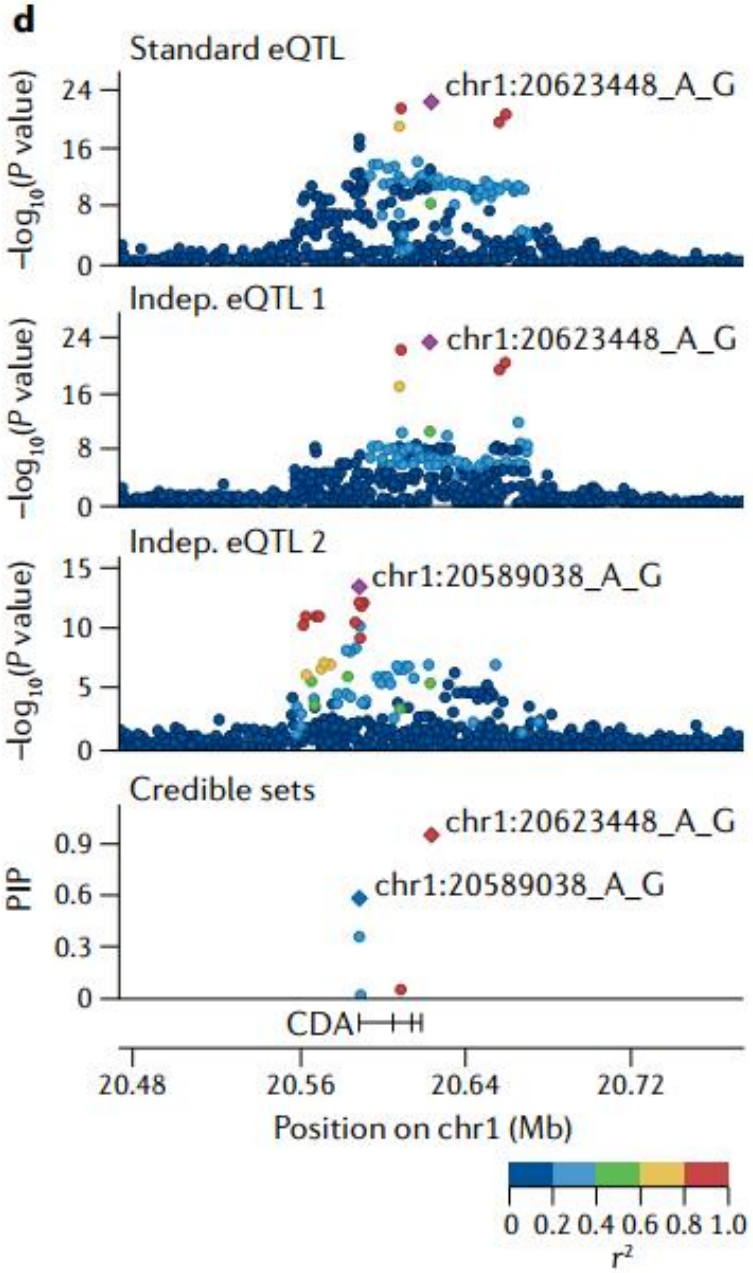
- MatrixeQTL
- FastQTL
- QTLtools
- TensorQTL

Visualization



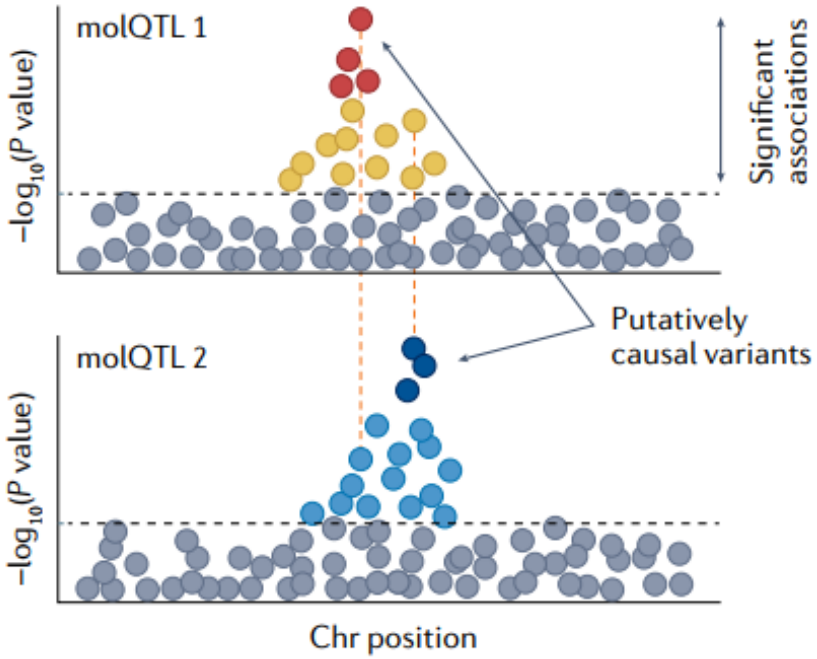
Fine-mapping

- Fine-mapping
  - Allelic heterogeneity: multiple LD-independent molQTLs in the same genomic region affect the same molecular feature.
  - Methods:
    - Iterative conditioning: forward selection
    - Posterior inclusion probabilities & credible sets of variants: CAVIAR, DAP-G, FINEMAP, SuSiE



Colocalization

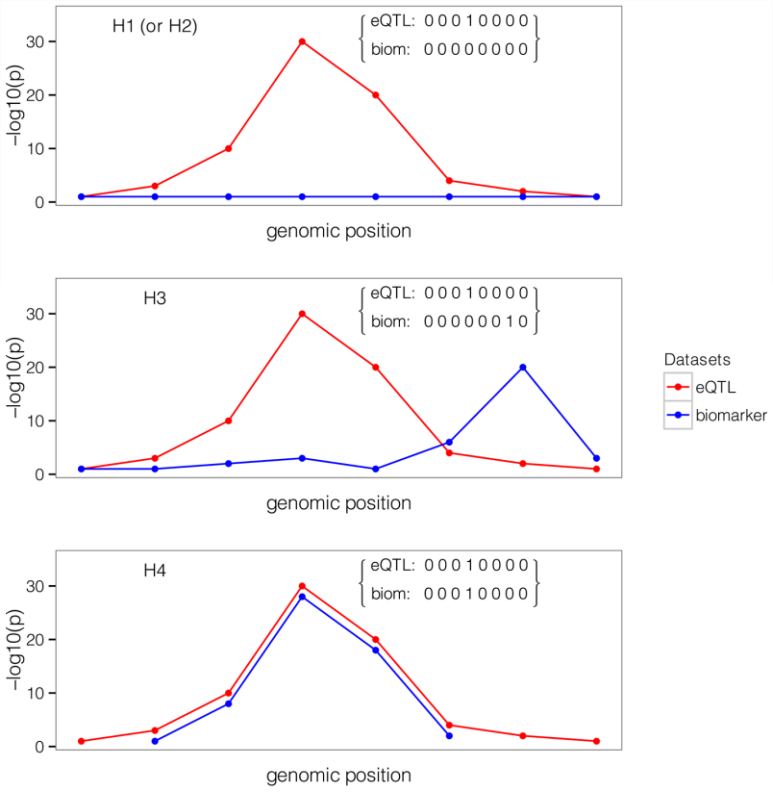
- Genetic colocalization: The phenomenon whereby genetic factors at a particular locus are shared between two or more traits
- Tests for genetic colocalization try to separate between two scenarios: (i) there is a causal variant for trait A that is distinct from the causal variant for trait B, whilst being at the same locus, and (ii) the causal variant for trait A and trait B are shared.



Colocalization

The idea behind the ABF analysis is that the association of each trait with SNPs in a region may be summarised by a vector of 0s and at most a single 1, with the 1 indicating the causal SNP (so, assuming a single causal SNP for each trait). The posterior probability of each possible configuration can be calculated and so, crucially, can the posterior probabilities that the traits share their configurations. This allows us to estimate the support for the following cases:

- H_0: neither trait has a genetic association in the region
- H_1: only trait 1 has a genetic association in the region
- H_2: only trait 2 has a genetic association in the region
- H_3: both traits are associated, but with different causal variants
- H_4: both traits are associated and share a single causal variant



GWAS_summary_statistics.csv

rsID	BETA	SE	N	s	MAF	type
rs1	0.4	0.02	10000	0.5	0.2	cc
rs2	0.1	0.01	10000	0.5	0.1	cc
...	...	...	...	...	...	...
rs1000	0.2	0.01	10000	0.5	0.1	cc

eQTL_summary_statistics.csv

rsID	BETA	SE	N	MAF	type
rs1	0.3	0.03	1000	0.05	quant
rs2	0.1	0.01	1000	0.05	quant
...	...	...	...	...	...
rs1000	0.4	0.02	1000	0.05	quant

```
Coloc.abf(GWAS_summary_statistics.csv,
          eQTL_summary_statistics.csv)
```

Output:

```
PP.H0:     3e-7
PP.H1:     8e-6
PP.H2:     3e-5
PP.H3:     7e-5
PP.H4:     9e-1
```

Colocalization

- 第一种设想 H0: 表型1（GWAS）和 表型2 （以eQTL为例）与某个基因组区域的所有SNP位点无显著相关；
- 第二种设想 H1/H2: 表型1（GWAS）或表型2（以eQTL为例）与某个基因组区域的SNP位点显著相关；
- 第三种设想 H3: 表型1（GWAS）和 表型2 （以eQTL为例）与某个基因组区域的SNP位点显著相关，但由不同的因果变异位点驱动；
- 第四种设想 H4: 表型1（GWAS）和 表型2 （以eQTL为例）与某个基因组区域的SNP位点显著相关，且由同一个因果变异位点驱动
- 共定位分析，本质上是在检验第四种的后验概率

Colocalization in coloc

- configuration: one possible combination of pairs of binary vectors indicating whether the variant is associated with the selected trait.
- We can group the configurations into five sets, S_0, S_1, S_2, S_3, S_4, containing assignments of all SNPs Q to the functional role corresponding to the five hypothesis H_0, H_1, H_2, H_3, H_4.

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

$$\frac{P(H_h|D)}{P(H_0|D)} = \sum_{S \in S_h} \frac{P(D|S)}{P(D|S_0)} \times \frac{P(S)}{P(S_0)}$$

Bayes factor (BF)

$$PP4$$

$$= P(H_4|D)$$

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

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

Approximate Bayes factor (ABF)

$$ABF = \sqrt{1-r} \times exp\left[\frac{Z^2}{2} \times r\right]$$

$$Z = \hat{\beta}/\sqrt{V}$$

$$r = W/(V+W)$$

prior

- p0 is the prior probability that a SNP is not associated with either trait, p1 is defined as the prior probability that the SNP is only associated with trait 1, p2 the prior probability that the SNP is associated only with trait 2, while p12 is the prior probability that the SNP is associated with both traits

- If $S \in S_0$, then $P(S) = p_0^Q$
- If $S \in S_1$, then $P(S) = p_0^{Q-1} \times p_1$
- If $S \in S_2$, then $P(S) = p_0^{Q-1} \times p_2$
- If $S \in S_3$, then $P(S) = p_0^{Q-2} \times p_1 \times p_2$
- If $S \in S_4$, then $P(S) = p_0^{Q-1} \times p_{12}$

- If $S \in S_0$, then $\frac{P(S)}{P(S_0)} = \frac{p_0^Q}{p_0^Q} = 1$
- If $S \in S_1$, then $\frac{P(S)}{P(S_0)} = \frac{p_0^{Q-1}}{p_0^Q} \times p_1 = \frac{p_1}{p_0} \approx p_1$
- If $S \in S_2$, then $\frac{P(S)}{P(S_0)} = \frac{p_0^{Q-1}}{p_0^Q} \times p_2 = \frac{p_2}{p_0} \approx p_2$
- If $S \in S_3$, then $\frac{P(S)}{P(S_0)} = \frac{p_0^{Q-2}}{p_0^Q} \times p_1 \times p_2 = \frac{p_1}{p_0} \times \frac{p_2}{p_0} \approx p_1 \times p_2$
- If $S \in S_4$, then $\frac{P(S)}{P(S_0)} = \frac{p_0^{Q-1}}{p_0^Q} \times p_{12} = \frac{p_{12}}{p_0} \approx p_{12}$

- $\frac{P(H_0|D)}{P(H_0|D)} = 1$
- $\frac{P(H_1|D)}{P(H_0|D)} = p_1 \times \sum_{j=1}^{Q} ABF_j^1$
- $\frac{P(H_2|D)}{P(H_0|D)} = p_2 \times \sum_{j=1}^{Q} ABF_j^2$
- $\frac{P(H_3|D)}{P(H_0|D)} = p_1 \times p_2 \times \sum_{j,k,j\neq k} ABF_j^1 ABF_k^2$
- $\frac{P(H_4|D)}{P(H_0|D)} = p_{12} \times \sum_{j=1}^{Q} ABF_j^1 \times ABF_j^2$

$$\frac{P(H_3 \mid D)}{P(H_0 \mid D)} = p_1 \times p_2 \times \sum_{j=1}^{Q} ABF_j^1 \sum_{j=1}^{Q} ABF_j^2 - \left[\frac{p_1 \times p_2}{p_{12}} \times \frac{P(H_4 \mid D)}{P(H_0 \mid D)}\right]$$

ABF

- phenotype-genotype $Y = \mu + \beta X$

under the null we assume that the effect size β = 0. Under the alternative, β is normally distributed with mean 0 and variance W

- asymptotically

$$\hat{\beta} \to N(\beta, V). \qquad \begin{bmatrix} \hat{\mu} \\ \hat{\beta} \end{bmatrix} \sim N_{p+1}\left(\begin{bmatrix} \mu \\ \beta \end{bmatrix} \begin{bmatrix} \mathbb{I}_{\mu\mu} & \mathbb{I}_{\mu\beta} \\ \mathbb{I}^T_{\mu\beta} & \mathbb{I}_{\beta\beta} \end{bmatrix}^{-1} \right)$$

$$\gamma = \mu + \frac{\mathbb{I}_{\mu\beta}}{\mathbb{I}_{\mu\mu}}\beta \qquad \begin{bmatrix} \hat{\gamma} \\ \hat{\beta} \end{bmatrix} \sim N_{p+1}\left(\begin{bmatrix} \gamma \\ \beta \end{bmatrix} \begin{bmatrix} \mathbb{I}^*_{\mu\mu} & 0 \\ 0^T & \mathbb{I}_{\beta\beta} \end{bmatrix}^{-1} \right)$$

- BF

$$\frac{P(H_h|D)}{P(H_0|D)} = \sum_{S \in S_h} \frac{P(D|S)}{P(D|S_0)} \times \frac{P(S)}{P(S_0)} \qquad BF = \int \frac{f(\beta)}{f(0)} \pi(\beta) d\beta \qquad ABF = \sqrt{1-r} \times exp\left[\frac{Z^2}{2} \times r\right]$$

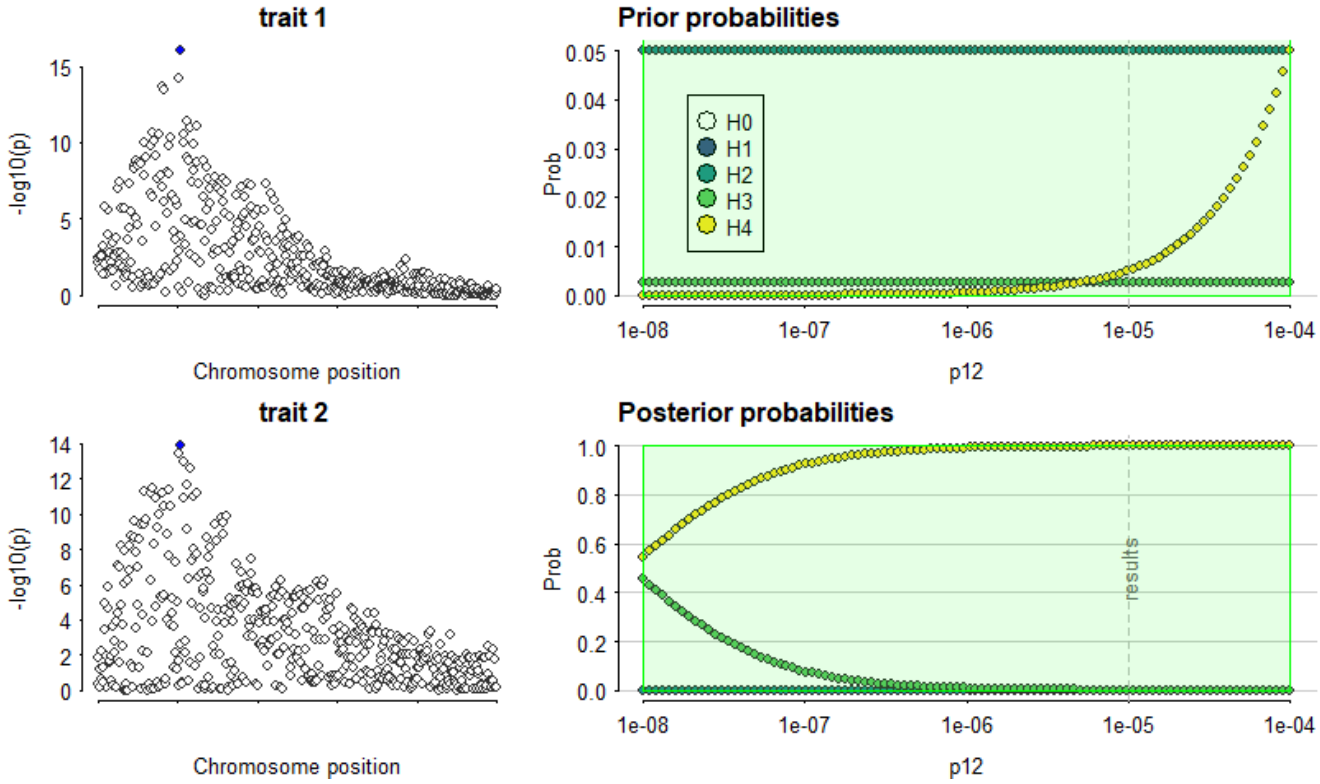
$$Z = \hat{\beta}/\sqrt{V}$$

$$r = W/(V+W)$$

Bayes factor (BF)

the relative contributions of the prior (W) and likelihood (V) to the inference.

sensitivity analysis



Specifying prior values for coloc.abf() is important, as results can be dependent on these values. Defaults of $p_1 = p_2 = 10^{-4}$ seem justified in a wide range of scenarios, because these broadly correspond to a 99% belief that there is true association when we see $p < 5 \times 10^{-8}$ in a GWAS. However, choice of p_{12} is more difficult. We hope the coloc explorer app will be helpful in exploring what various choices mean, at a per-SNP and per-hypothesis level. However, having conducted an enumeration-based coloc analysis, it is still helpful to check that any inference about colocalisation is robust to variations in prior values specified.

A sensitivity analysis can be used, post-hoc, to determine the range of prior probabilities for which a conclusion is still supported. The sensitivity() function shows this for variable p_{12} in the bottom right plot, along with the prior probabilities of each hypothesis, which may help decide whether a particular range of p_{12} is valid. The green region shows the region - the set of values of p_{12} - for which $H_4 > 0.5$ - the rule that was specified. In this case, the conclusion of colocalisation looks quite robust. On the left (optionally) the input data are also presented, with shading to indicate the posterior probabilities that a SNP is causal if H_4 were true. This can be useful to indicate serious discrepancies also.

Colocalization: SuSiE

- coloc has adopted the SuSiE framework for fine mapping in the presence of multiple causal variants. This framework requires the LD matrix is known, so first check our datasets hold an LD matrix of the right format. =check_dataset= should return NULL if there are no problems, or print informative error messages if there are.
- fine-mapping: variable selection

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e},$$

$$\mathrm{PIP}_j := \Pr(b_j \neq 0 \mid \boldsymbol{X}, \boldsymbol{y}).$$

posterior inclusion probability (PIP)

- Sum of Single Effects (SuSiE) model

$$\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l,$$

each vector bl is a "single effect" vector; that is, a vector with exactly one non-zero element.

L=1: "single effect regression" (SER) model

L>1: Iterative Bayesian Stepwise Selection (IBSS)

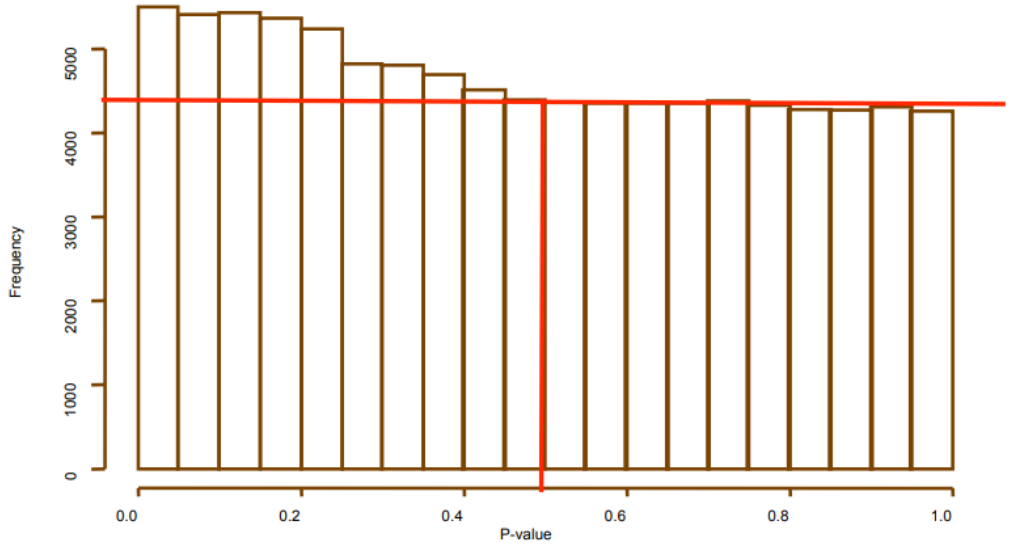
Replication

- Simple overlap of significant variant-feature pairs between two sets: suboptimal
  - Difference in power
  - LD contamination
- π_1 statistic
  - quantifies the proportion of true positives based on the distribution of P values

$$1 - \hat{\pi}_0$$

- colocalization
- using allelic data

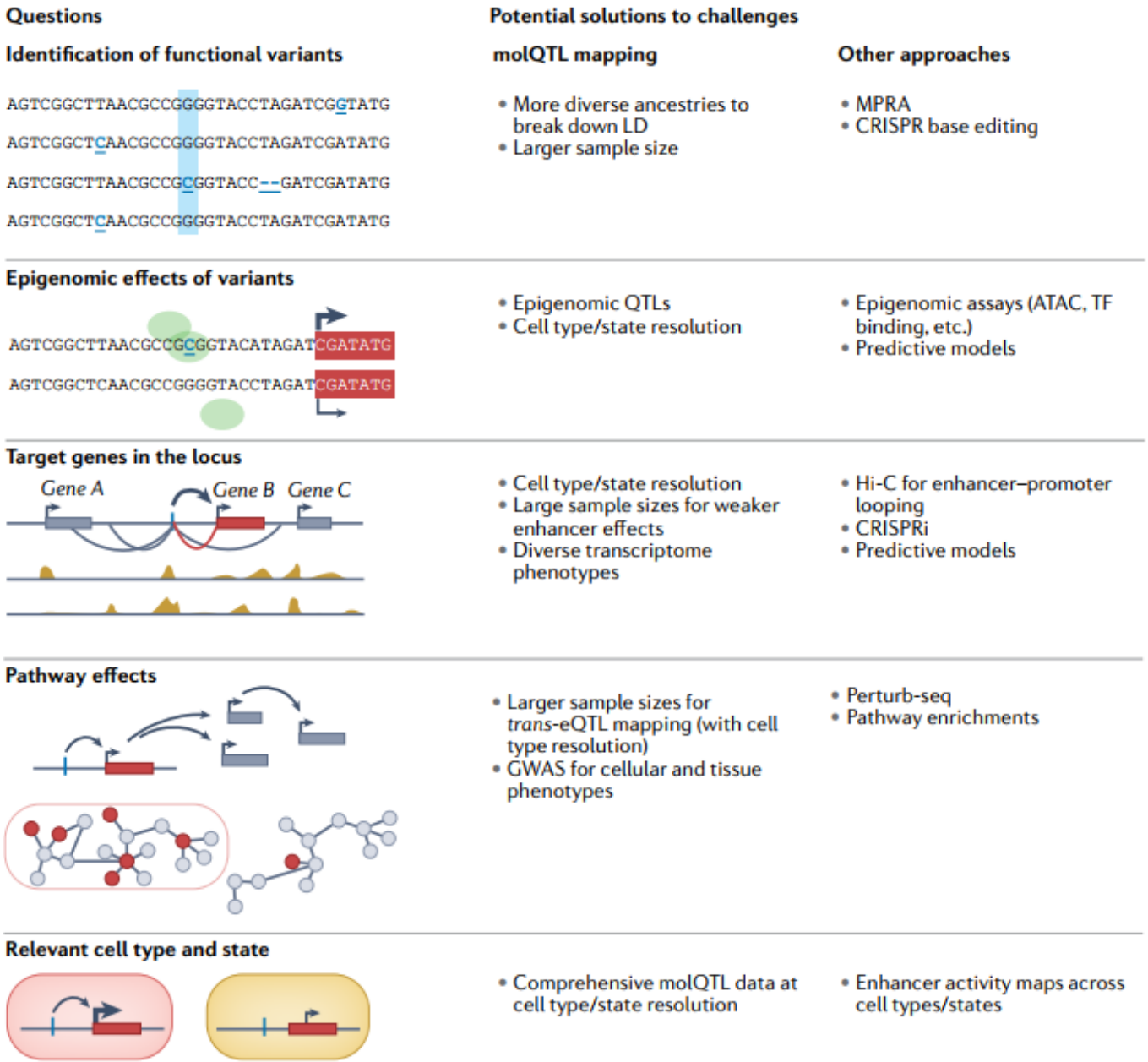
$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1,2,\ldots,m\}}{m(1-\lambda)}$$



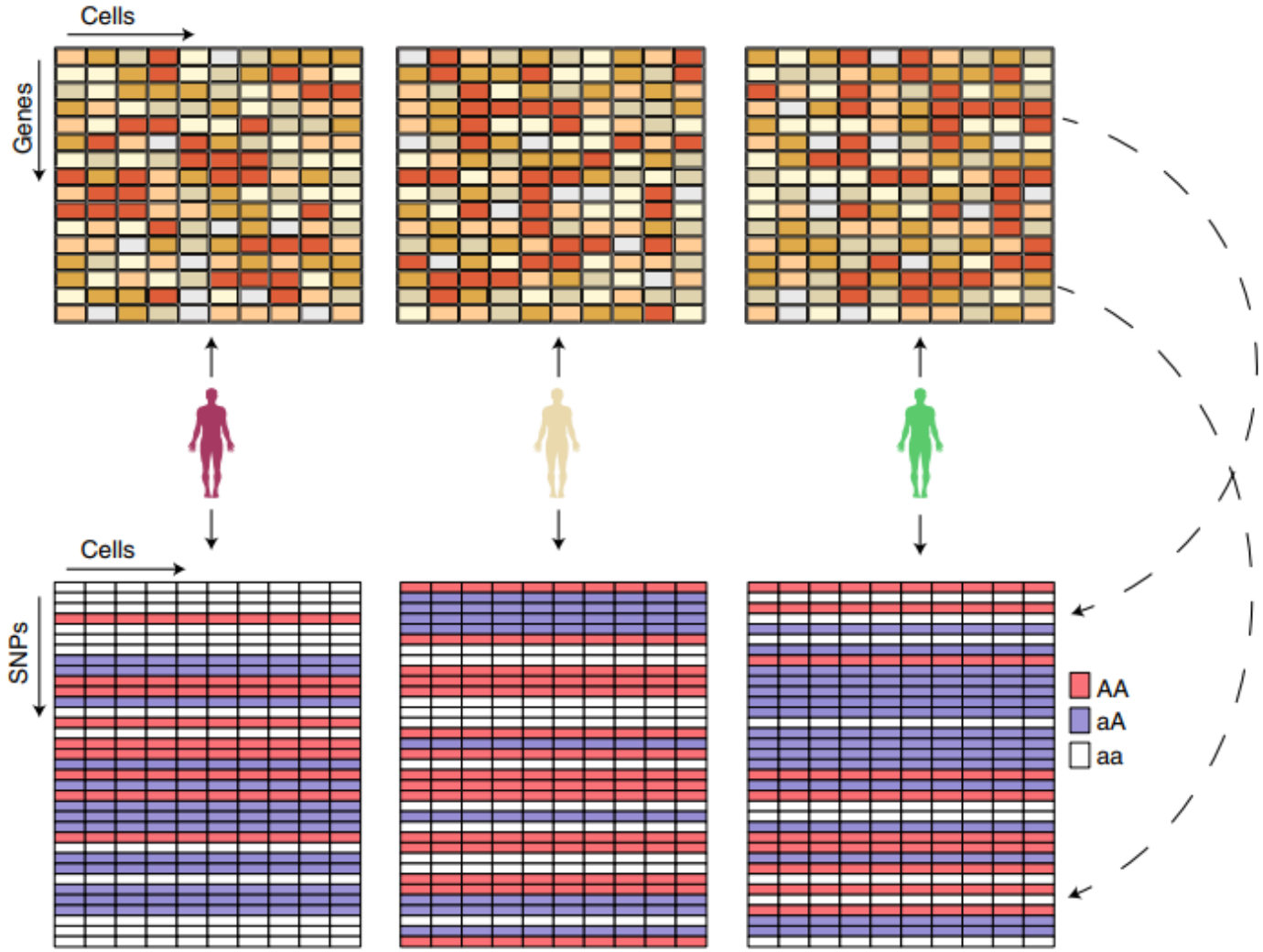
More concern

- Tissue and cell context specificity
- Population differences
- Multi-omic QTL integration
  - combining fine-mapping and co-localization methods to identify shared signals, followed by mediation analysis to identify causal relationships.
- GWAS integration
  - GWAS co-localization analyses: assess whether the association signal for two traits in a given genomic region is driven by shared or distinct causal variants.
  - Transcriptome-wide association studies (TWAS)
- Dynamic, spatial patterns

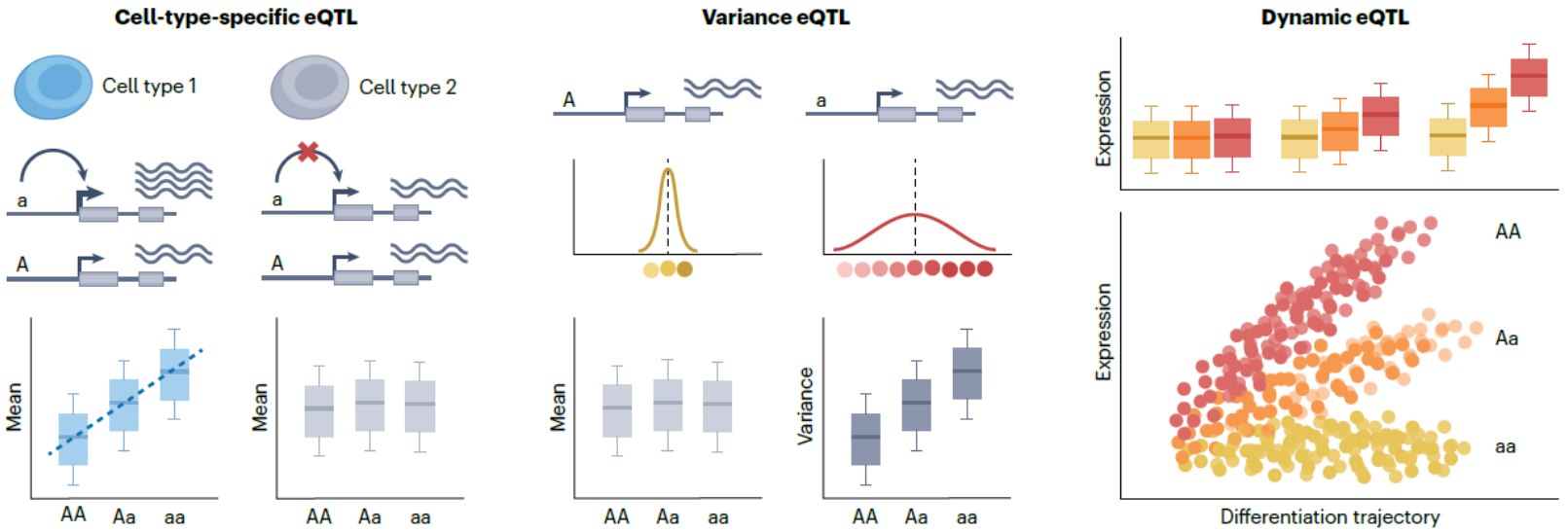
More analysis



eQTL with single-cell data

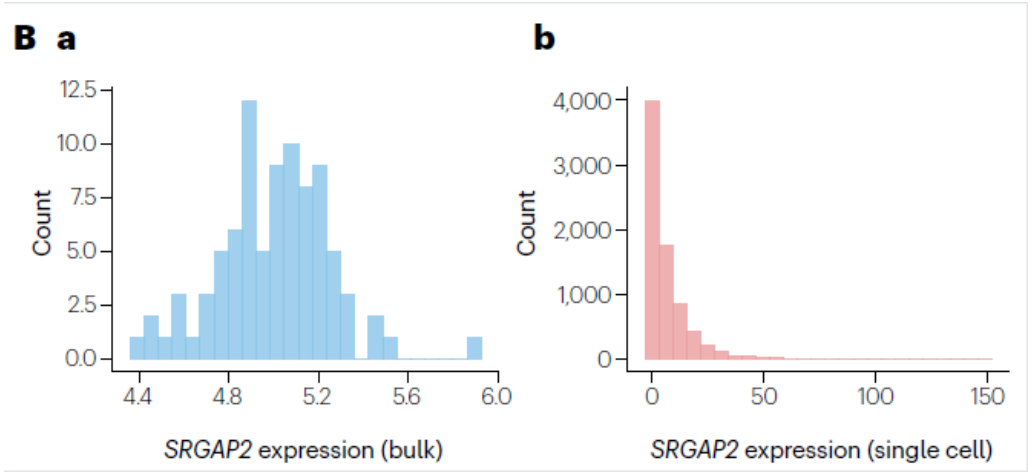


eQTL with single-cell data



QTL with single-cell genomics

- pseudo-bulk data: aggregated expression of each gene in each cluster or cell type with genotypes of individuals at nearby variants
- single-cell resolution: cell type & cell state
  - sparsity & non-normality: Poisson, negative binomial, multinomial
  - scalability: quality of cells, consistency of outputs, memory and computation (parallelizing, sparse matrices)



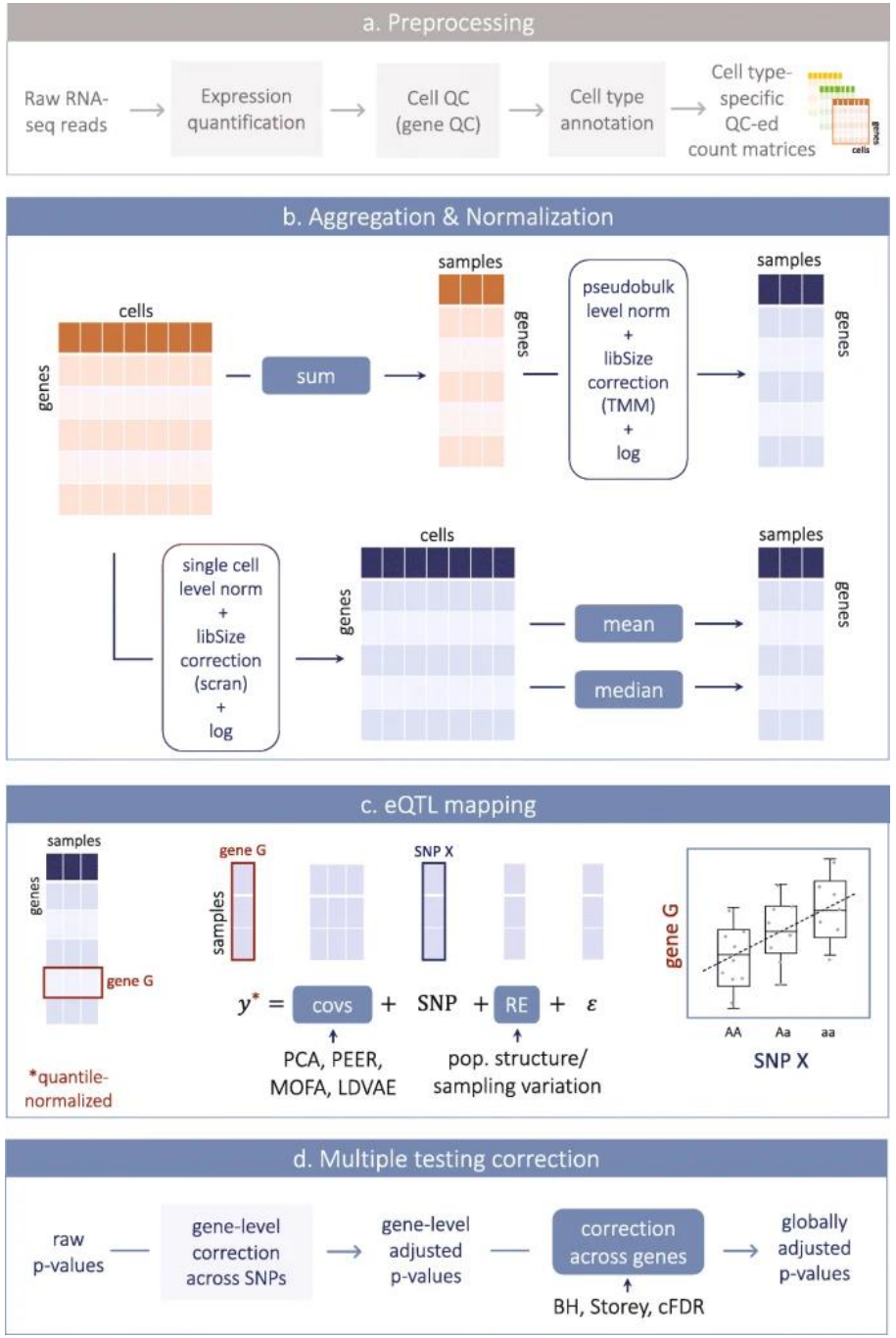
Cell-type specific eQTL

- Bulk data:
  - 把一个bulk里面能够反映各种细胞类型的比例或者代理指标作为协变量：交互效应

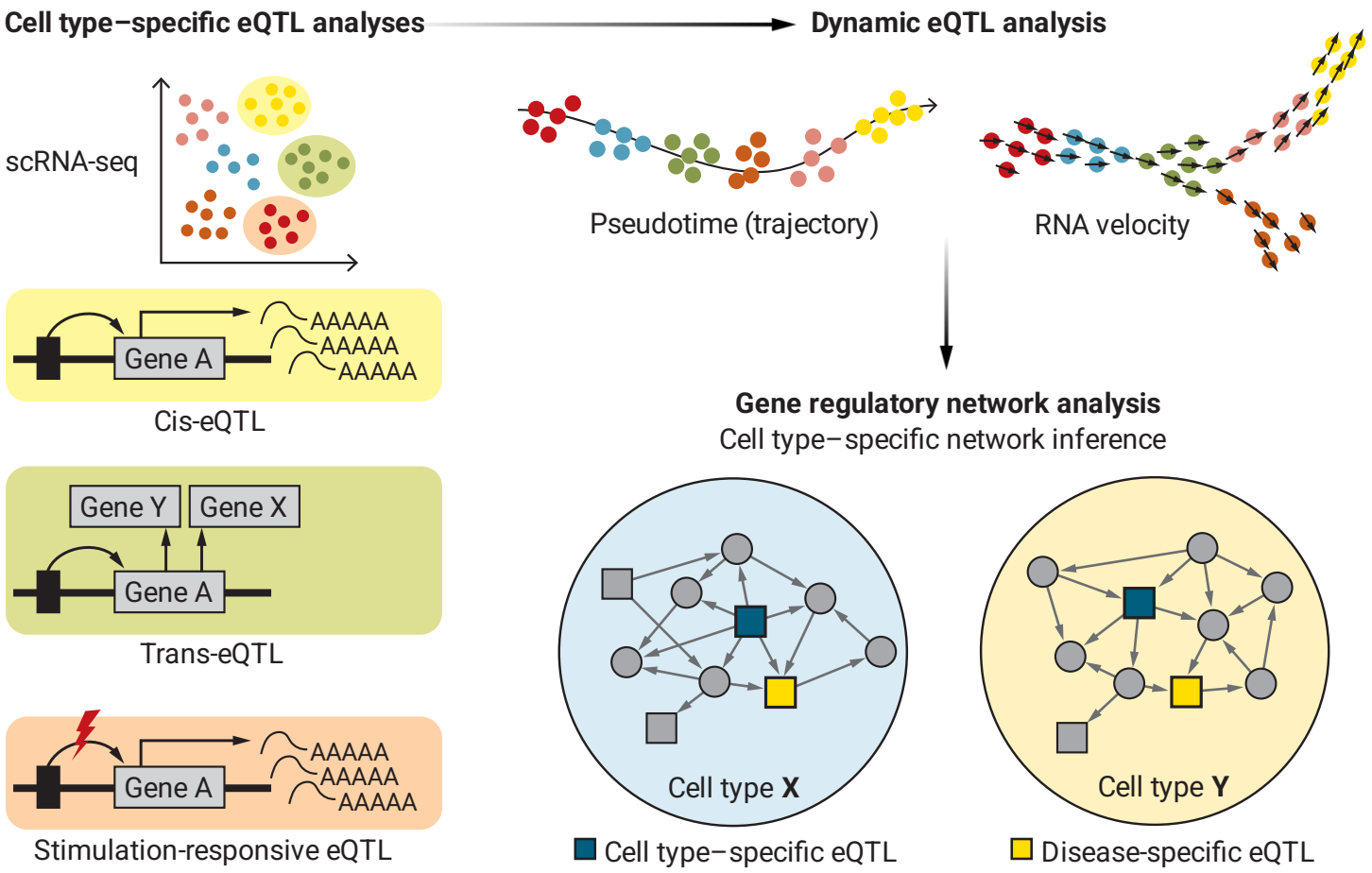
$$y_{ng} = \beta_g + \beta_{gs} x_{ns} + \beta_{gm} m_n + \beta_{g,sm} (x_{ns} \times m_n) + \varepsilon_{ngs},$$

- Single-cell data: Cell-type annotation
  - pseudo-bulk: 根据不同细胞类型把总样本分开；再将每种细胞类型的各个细胞按照个体整合成一个新的“个体”（先整合后bulk标准化/先单细胞标准化再整合）； 再用标准eQTL方法
  - Treat a single-cell as an individual sample: 需要考虑单细胞的特点： non-Gaussian, dropout, batch effect, etc.
    - cellRegMap

$$y_{ngi} = \beta_g + \beta_{gs} x_{ns} + \beta_{g,si} x_{ns} + u_{ng} + c_{ngi} + \varepsilon_{ngi},$$



From cell type-specific to dynamic eQTL



From single-cell eQTL to more molQTLs

Table 1 | **Representative human tissue profiling resources**

Database	No. of tissues/ samples/donors	Profiling method	Clinical state	Refs
mRNA				
GTEx	51 (+2 CL)/~11,500/~700	RNA-seq	Normal	14
HPA	44/122/122	RNA-seq	Normal	18
Protein				
HPA	32/122/122	IHC	Normal and diseased	18
Human Proteome Map	30/85/3	Mass spectrometry	Normal	155
Non-coding RNAs				
TissueAtlas (microRNAs)	61/61/2	Microarray	Normal	156
DASHR	86 (+51 PC, 48 CL)/>800/NR	Data integration	Normal	157
FANTOM5 (microRNAs, lncRNAs, promoters, enhancers)	400/150 (+570 PC, 250 CL)/3	Various	Normal	33,158
Regulatory elements				
GTEx (eQTLs)	48/~10,000/~600	eQTLs	Normal	14
ENCODE and Roadmap Epigenomics	>120/NR/hundreds	ChIP-seq, DNase-seq, ATAC-seq, FAIRE-seq	Normal	22,32
3D Genome	109/113/NR	Hi-C, ChIA-PET, Capture Hi-C, PLAC-seq	Normal	122
TiGER	30 tissues	Data integration	Normal	159
Single cells expression profiles				
Human Cell Atlas	3/17/17	scRNA-seq	Normal and diseased	16
Single Cell Portal	68 studies	scRNA-seq	Normal and diseased	

Limitation

- Donor selection and sample size
  - “environment”: lifestyle, demographics, other biomedical traits, etc.
  - ancestry
  - disease-relevant
- Biospecimen selection and resolution
- Molecular read-outs
  - many classes of mRNA and non-coding RNA lacking a polyA tail or subject to rapid degradation
- Rare variants
- LD contamination and pleiotropy (co-expression)