Post-GWAS Analysis I

PEILIN JIA

BEIJING INSTITUTE OF GENOMICS

Outline

Genotype imputation

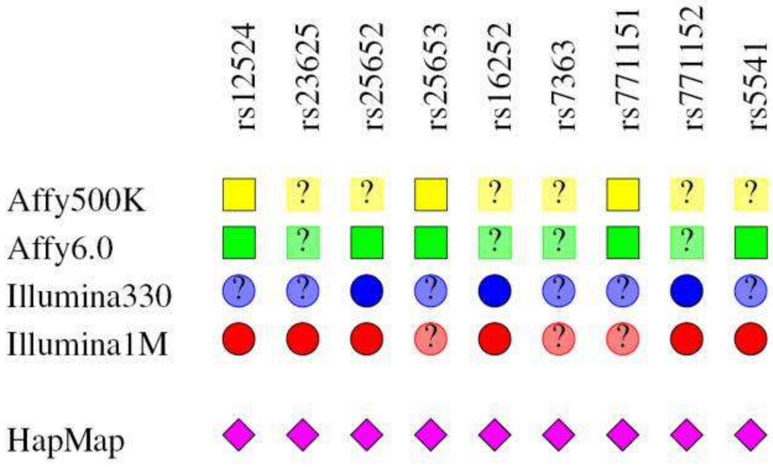
Quantitative Trait Locus (QTL)

Regulatory roles of genetic variants

Resources for secondary analyses

Genotype Imputation

Genotype imputation is **a process of estimating missing genotypes from the haplotype or genotype reference panel.**



The detection of more loci requires a larger sample size, larger sequencing depth for whole-genome sequencing, and a denser SNP array for microarray-based genotyping.

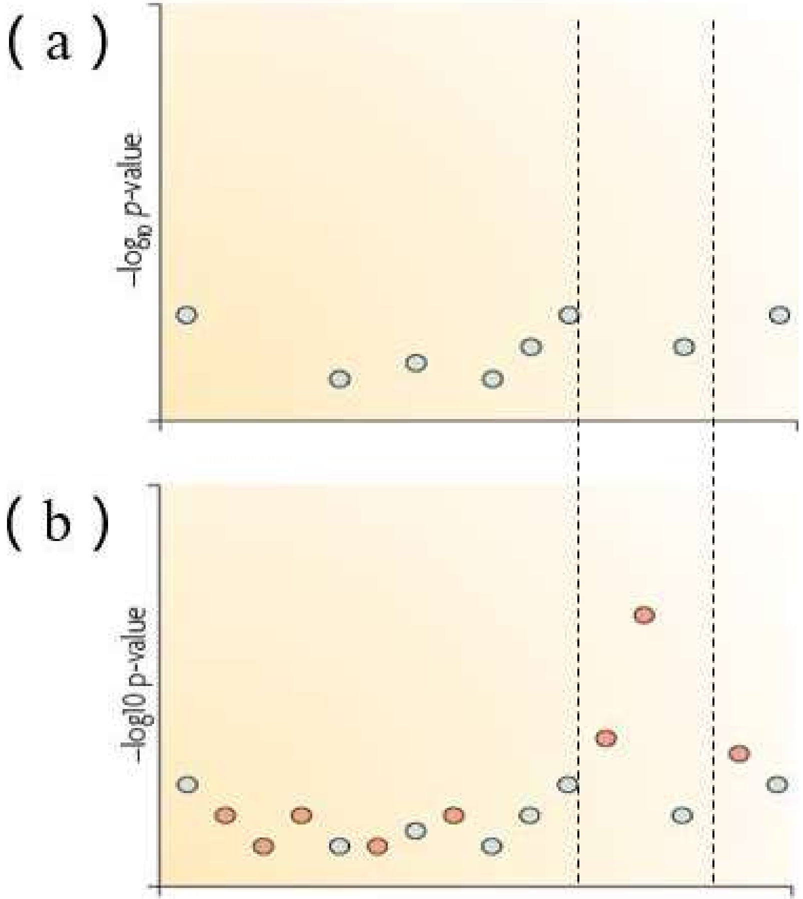
Genotype imputation can be used to solve this dilemma by predicting untyped genotypes from the haplotype reference panel.

Genotype imputation

Testing association at typed SNPs may not lead to a clear signal

Testing association at imputed SNPs may boost the signal

Imputation attempts to predict these missing genotypes

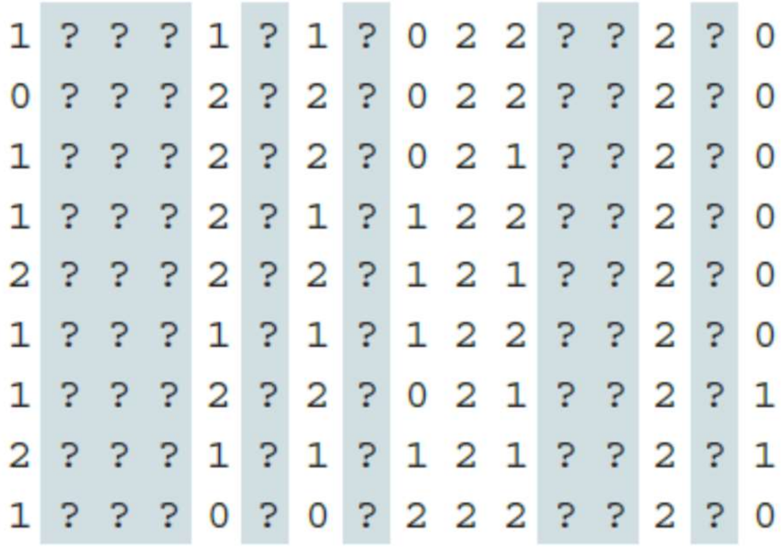


Genotype imputation

- ❑ Genotype imputation is the term used to describe the process of predicting or imputing genotypes that are not directly assayed in a sample of individuals.
- ❑ Common practice: a reference panel of haplotypes at a dense set of SNPs is used to impute into a study sample of individuals that have been genotyped at a subset of the SNPs.
- ❑ Genotype imputation can be carried out across the whole genome as part of a genome-wide association (GWA) study or in a more focused region as part of a fine-mapping study.

Intuitive example

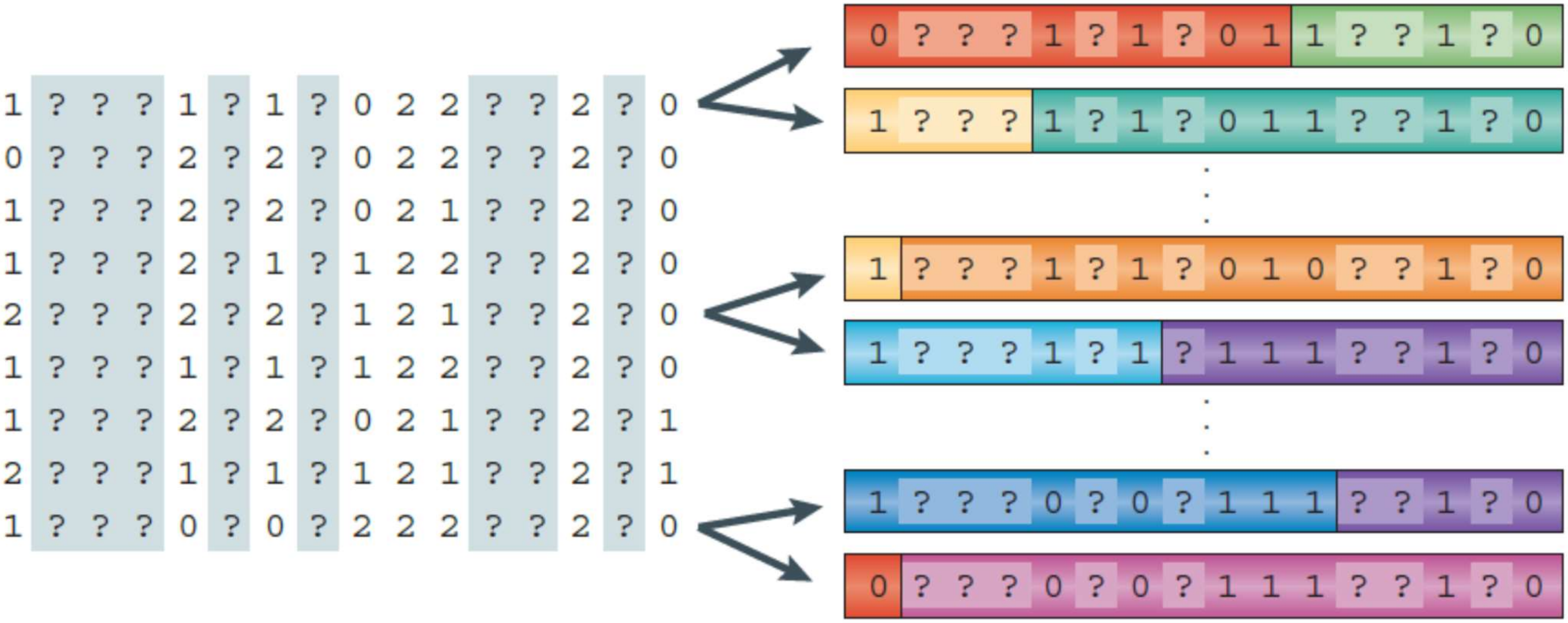
Step 1. Genotype data with missing data at untyped SNPs (grey question marks)



1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

Intuitive example

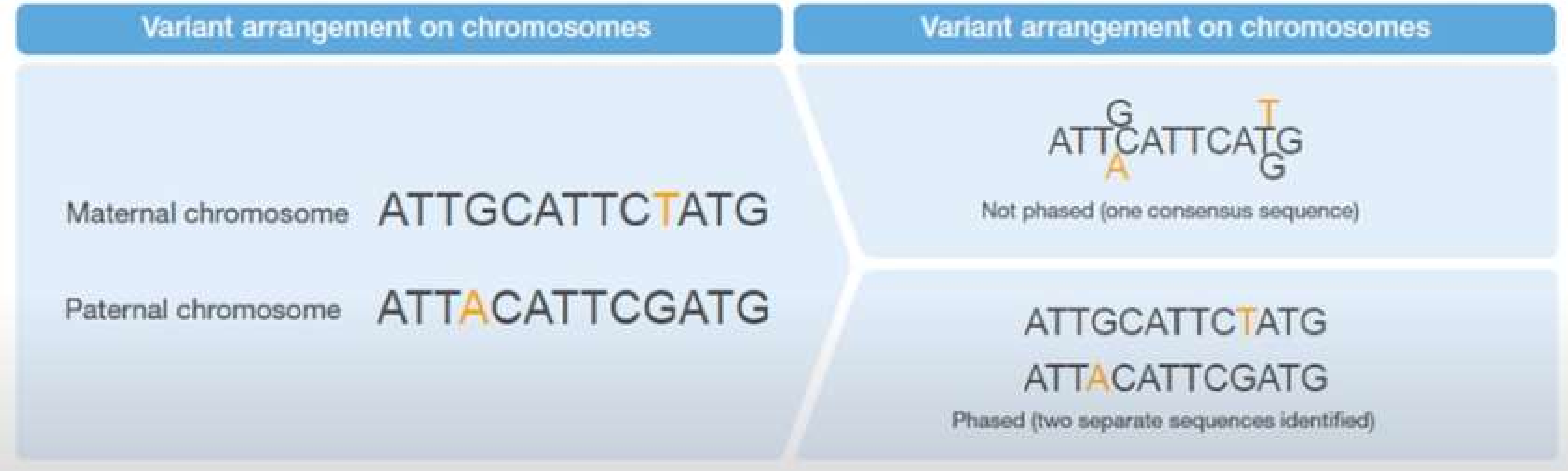
Step 2. Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel



```
1 ? ? ? 1 ? 1 ? 0 2 2 ? ? 2 ? 0
0 ? ? ? 2 ? 2 ? 0 2 2 ? ? 2 ? 0
1 ? ? ? 2 ? 2 ? 0 2 1 ? ? 2 ? 0
1 ? ? ? 2 ? 1 ? 1 2 2 ? ? 2 ? 0
2 ? ? ? 2 ? 2 ? 1 2 1 ? ? 2 ? 0
1 ? ? ? 1 ? 1 ? 1 2 2 ? ? 2 ? 0
1 ? ? ? 2 ? 2 ? 0 2 1 ? ? 2 ? 1
2 ? ? ? 1 ? 1 ? 1 2 1 ? ? 2 ? 1
1 ? ? ? 0 ? 0 ? 2 2 2 ? ? 2 ? 0
```

```
0 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
                ⋮
1 ? ? ? 1 ? 1 ? 0 1 0 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 1 1 1 ? ? 1 ? 0
                ⋮
1 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
0 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
```

What does “phasing” mean?

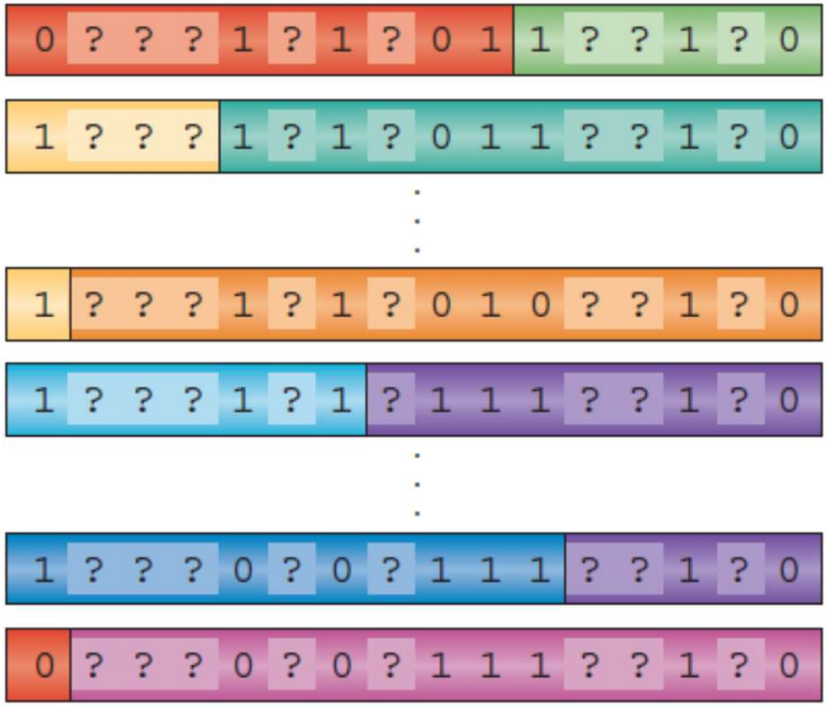

Variant arrangement on chromosomes
Maternal chromosome
ATTGCATTCTATG
Paternal chromosome
ATTACATTCGATG
Variant arrangement on chromosomes
G
T
ATTCATTCATG
A
G
Not phased (one consensus sequence)
ATTGCATTCTATG
ATTACATTCGATG
Phased (two separate sequences identified)

Phasing refers to the separation of a consensus sequence into individual sequence strands to identify which variants occur together or in phase.

Phasing separate the consensus strand into two separate identifiable sequences and we can see how the non-reference alleles in the two loci are organized.

Intuitive example

Step 3. These haplotypes are compared to the dense haplotypes in the reference panel



0 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0

1 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0

⋮

1 ? ? ? 1 ? 1 ? 0 1 0 ? ? 1 ? 0

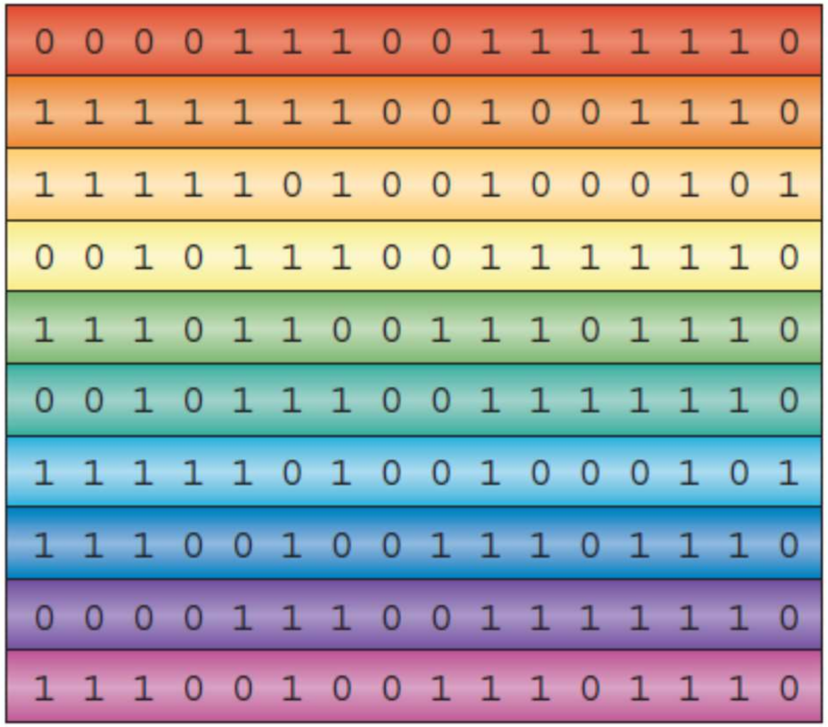
1 ? ? ? 1 ? 1 ? 1 1 1 ? ? 1 ? 0

⋮

1 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0

0 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0

Reference set of haplotypes, e.g., HapMap



0 0 0 0 1 1 1 0 0 1 1 1 1 1 1 0

1 1 1 1 1 1 1 0 0 1 0 0 1 1 1 0

1 1 1 1 1 0 1 0 0 1 0 0 0 1 0 1

0 0 1 0 1 1 1 0 0 1 1 1 1 1 1 0

1 1 1 0 1 1 0 0 1 1 1 0 1 1 1 0

0 0 1 0 1 1 1 0 0 1 1 1 1 1 1 0

1 1 1 1 1 0 1 0 0 1 0 0 0 1 0 1

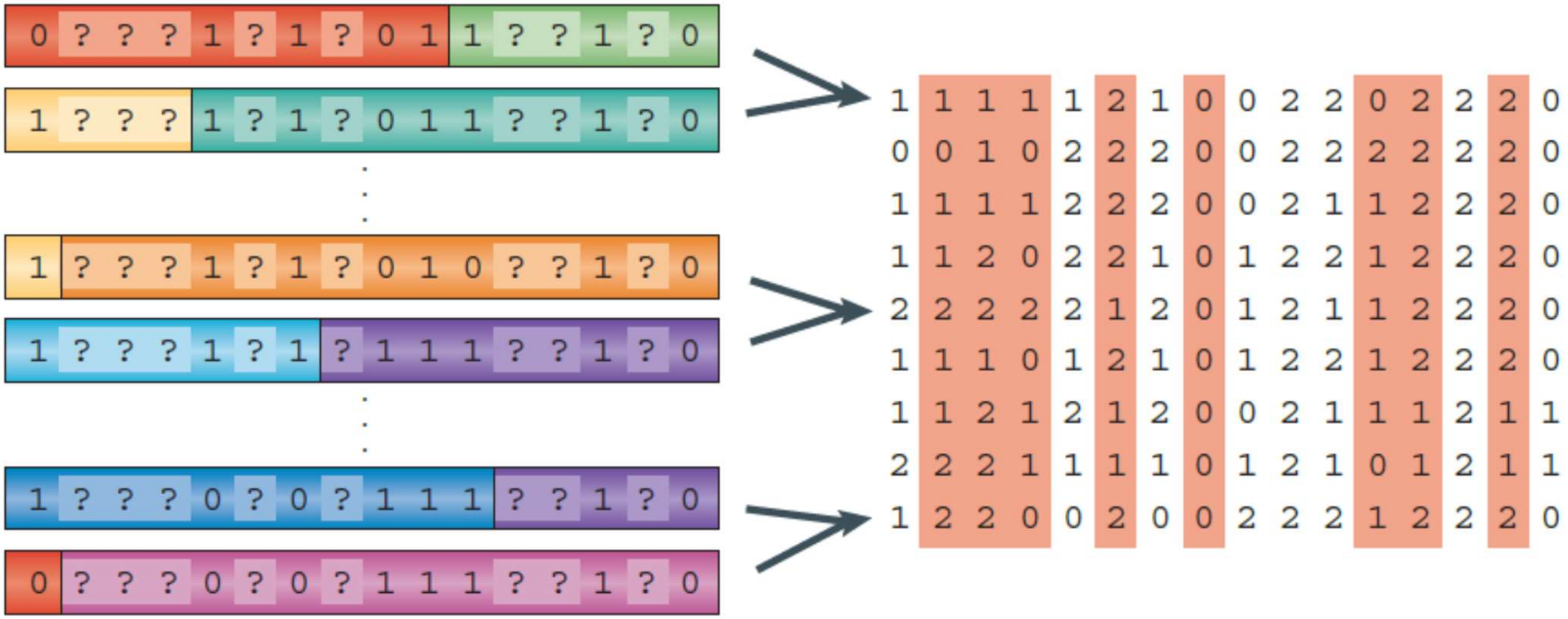
1 1 1 0 0 1 0 0 1 1 1 0 1 1 1 0

0 0 0 0 1 1 1 0 0 1 1 1 1 1 1 0

1 1 1 0 0 1 0 0 1 1 1 0 1 1 1 0

Intuitive example

Step 4. Missing genotypes in the study sample are then imputed using those matching haplotypes in the reference set



```
0 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 0 1 1 ? ? 1 ? 0
.
.
.
1 ? ? ? 1 ? 1 ? 0 1 0 ? ? 1 ? 0
1 ? ? ? 1 ? 1 ? 1 1 1 ? ? 1 ? 0
.
.
.
1 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
0 ? ? ? 0 ? 0 ? 1 1 1 ? ? 1 ? 0
```

1	1	1	1	1	2	1	0	0	2	2	0	2	2	2	0
0	0	1	0	2	2	2	0	0	2	2	2	2	2	2	0
1	1	1	1	2	2	2	0	0	2	1	1	2	2	2	0
1	1	2	0	2	2	1	0	1	2	2	1	2	2	2	0
2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	0
1	1	1	0	1	2	1	0	1	2	2	1	2	2	2	0
1	1	2	1	2	1	2	0	0	2	1	1	1	2	1	1
2	2	2	1	1	1	1	0	1	2	1	0	1	2	1	1
1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	0

In real examples, the genotypes are imputed with uncertainty and a probability distribution over all three possible genotypes is produced.

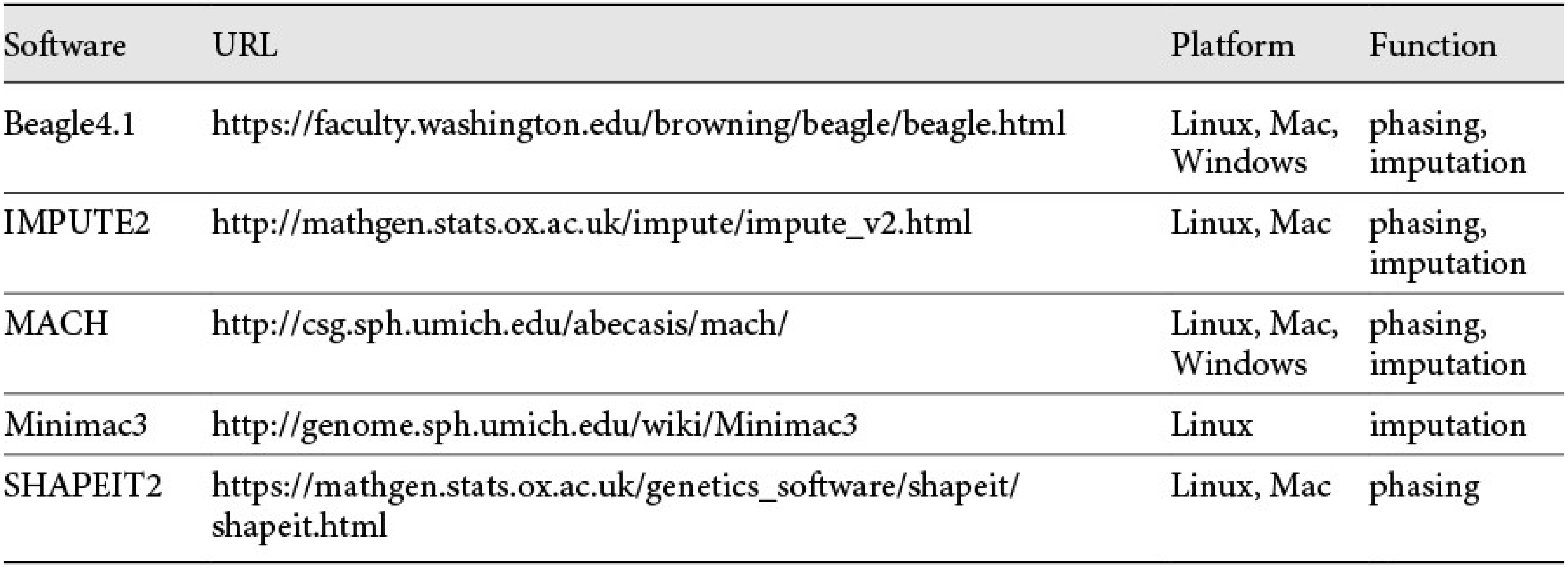
Factors affecting genotype imputation

The performance of genotype imputation is affected by many factors, such as software, reference selection, SNP density (see respective section in “Methods”), sample size, and sequencing coverage.

Steps

- **Quality control of genotypes**
- **Make sure to use the same version of reference genome**
- **Choose the reference panel**
- Quality control in post-imputation

Methods



Software	URL	Platform	Function
Beagle4.1	https://faculty.washington.edu/browning/beagle/beagle.html	Linux, Mac, Windows	phasing, imputation
IMPUTE2	http://mathgen.stats.ox.ac.uk/impute/impute_v2.html	Linux, Mac	phasing, imputation
MACH	http://csg.sph.umich.edu/abecasis/mach/	Linux, Mac, Windows	phasing, imputation
Minimac3	http://genome.sph.umich.edu/wiki/Minimac3	Linux	imputation
SHAPEIT2	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html	Linux, Mac	phasing

Uses of imputation

Boosting power

- Imputation can lead to a boost in power of up to 10% over testing only genotyped SNPs in GWAS.

Fine-mapping

- Imputation provides a high-resolution view of an associated region and increases the chance that a causal SNP can be directly identified.

Meta-analysis

- If different cohorts have used different genotyping chips, imputation can be used to equate the set of SNPs in each study.

Uses of imputation

Imputation of untyped variation

- Imputation of SNPs which have not been typed in the haplotype reference panel or the study sample is also possible.

Imputation of non-SNP variation

- The general idea of imputation is readily extended to other types of genetic variation such as copy number variants and classical human leukocyte antigen alleles

Sporadic missing data imputation and correction of genotyping errors

- Many of the widely used imputation programs allow imputation of sporadic missing genotypes that can occur when calling genotypes from genotyping chips

Outline

Genotype imputation

Quantitative Trait Locus (QTL)

- **QTL introduction**
- **Integration of GWAS Variants and xQTLs**

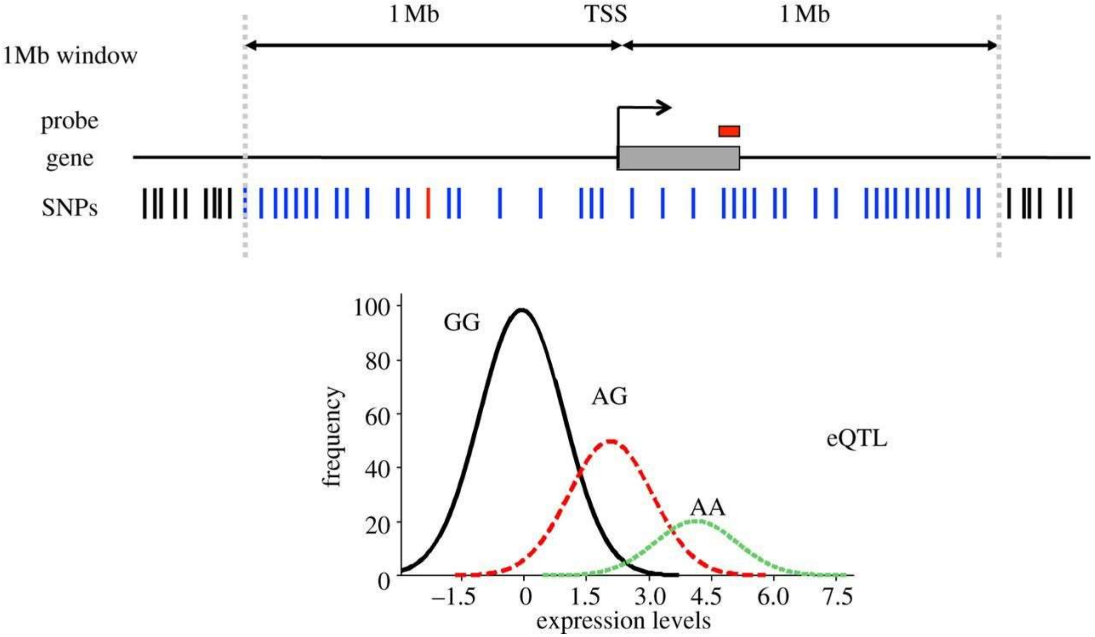
Regulatory roles of genetic variants

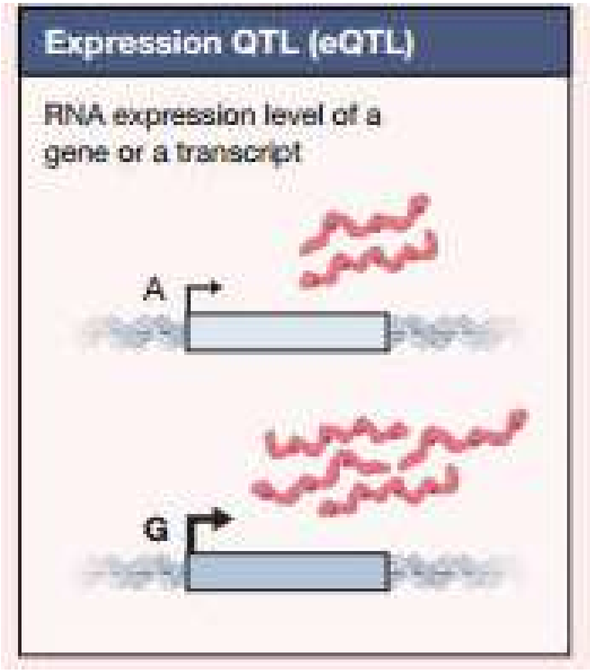
Resources for secondary analyses

Quantitative Trait Locus (QTL)

A quantitative trait locus (QTL) is a locus that correlates with variation of a quantitative trait of a population of organisms.

Expression QTL (eQTL) are QTL that modulate transcript abundance in pedigrees or crosses.





Regression models for QTL

Quantitative traits

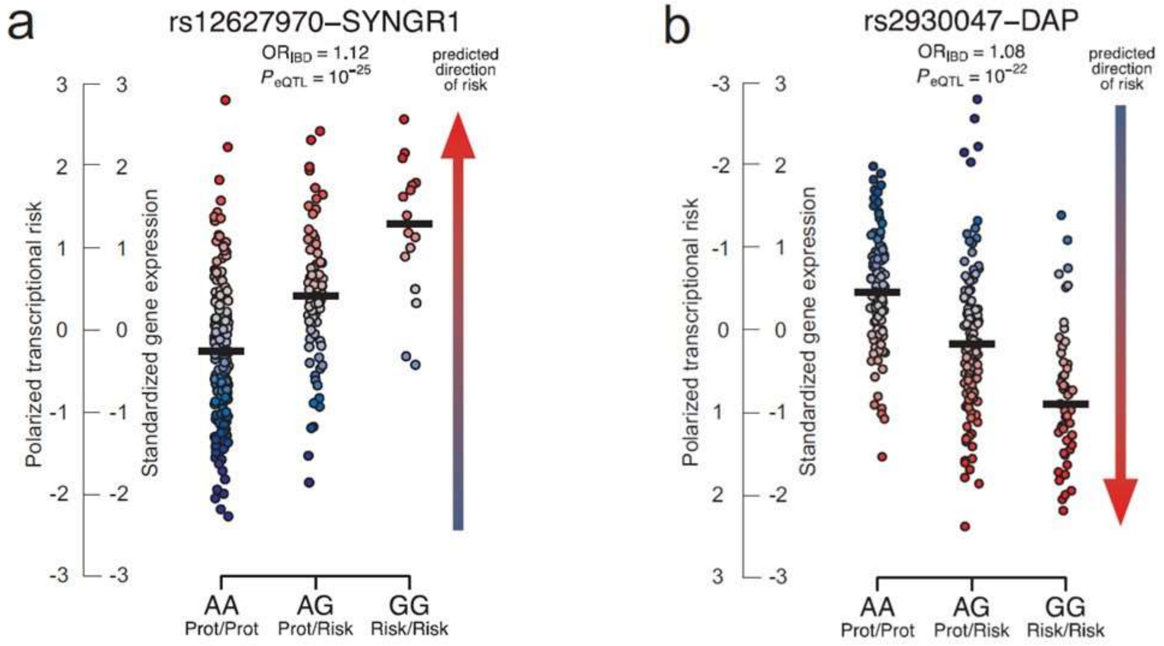
Simple linear regression

$y = \beta_0 + \beta_1 \times SNP$

Y can be any quantitative traits, e.g., gene expression, protein expression, and so on.

A couple of eSNPs

β_1: effect size



$\beta_1 > 0$

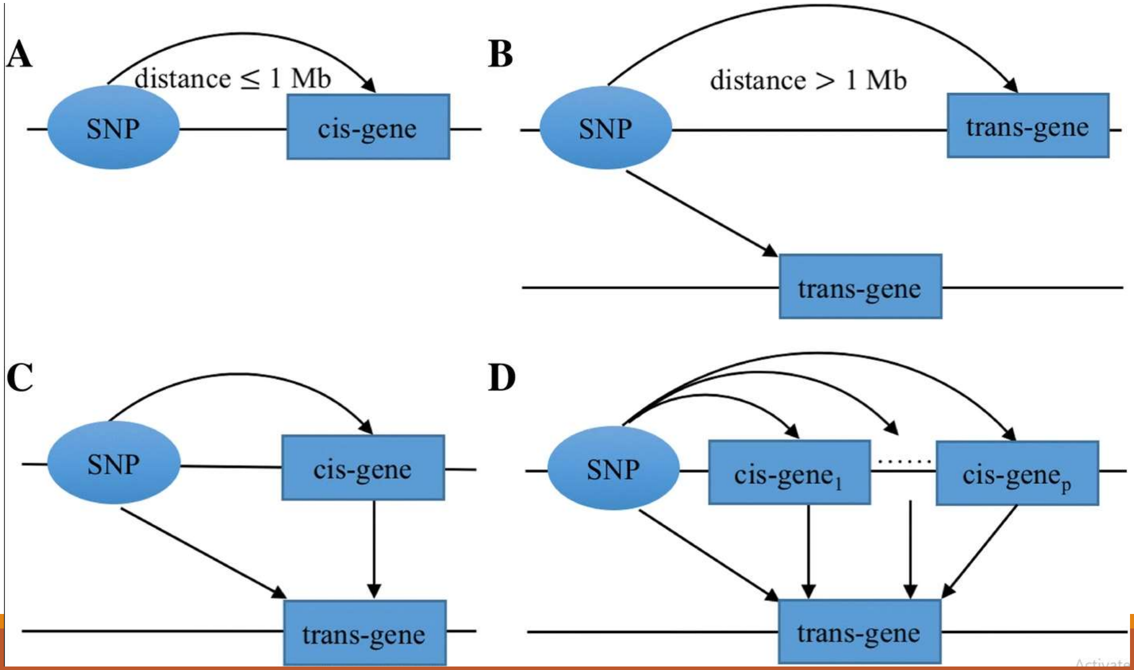
$\beta_1 < 0$

Expression QTL analysis

Expression SNP (eSNP) are SNPs that associate with transcript abundance in cohort studies. The target gene is called eGene

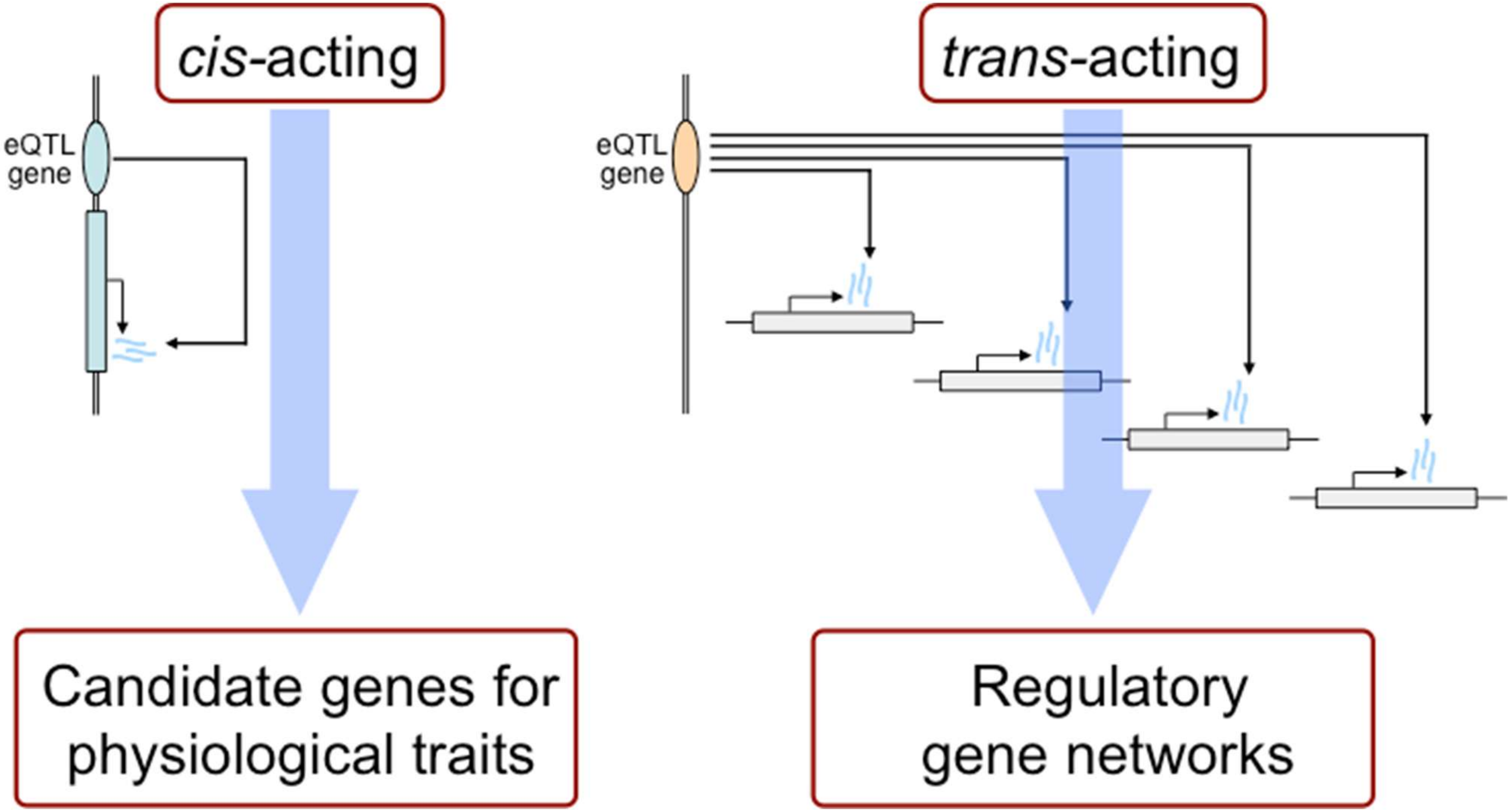
cis-eQTL: genetic variations act on local genes

trans-eQTL: genetic variations act on distant genes and genes residing on different chromosomes



The prefixes “cis” and “trans” are from Latin: *cis*: “this side of”, and *trans*: “the other side of”

cis- and *trans*-acting eQTLs



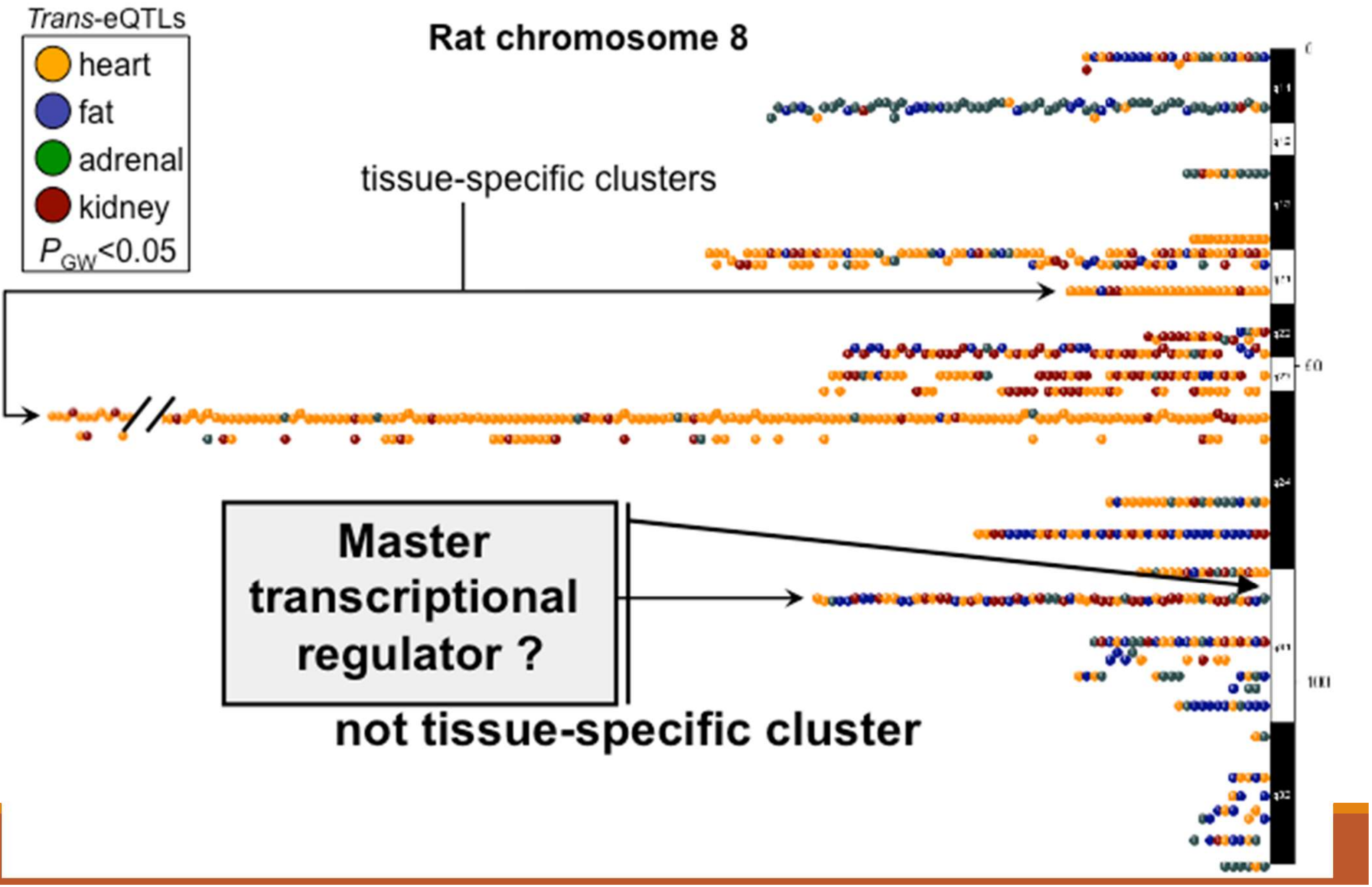
cis-acting

trans-acting

Candidate genes for physiological traits

Regulatory gene networks

trans-eQTLs hot-spots



tissue-specific clusters

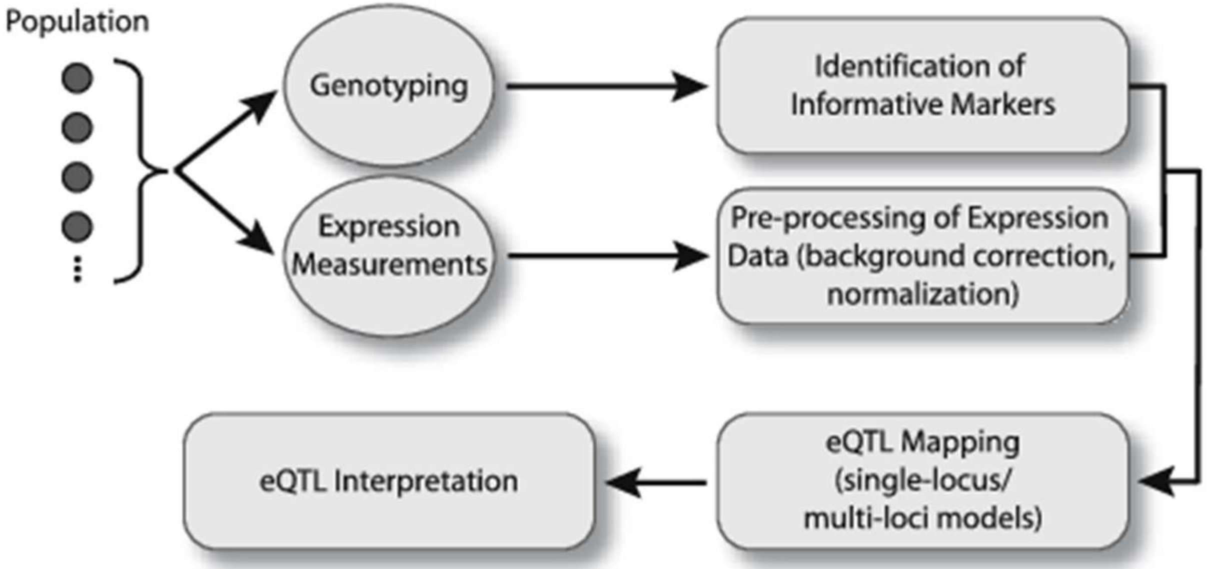
Master transcriptional regulator ?

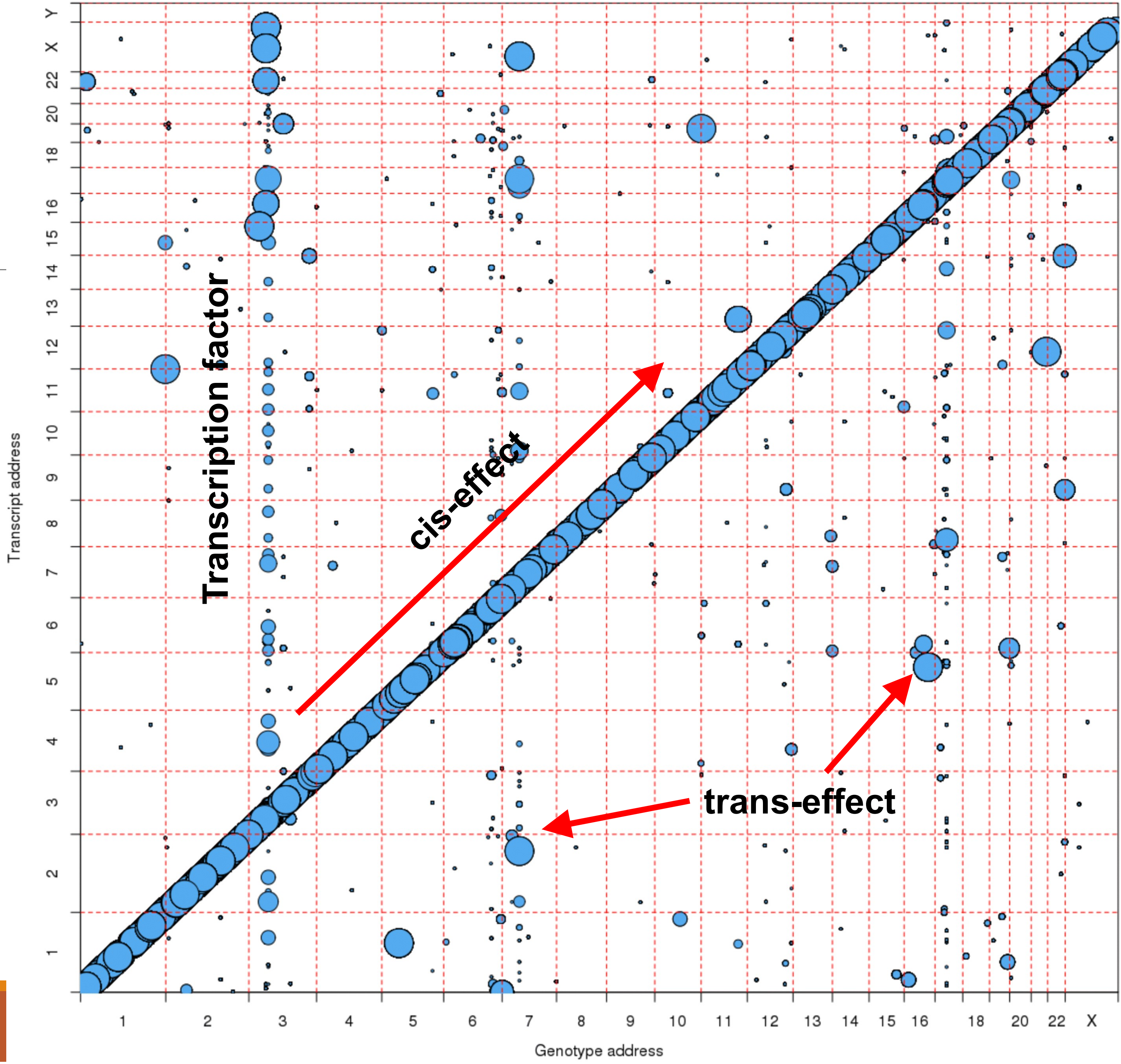
not tissue-specific cluster

Pipeline for eQTL analyses

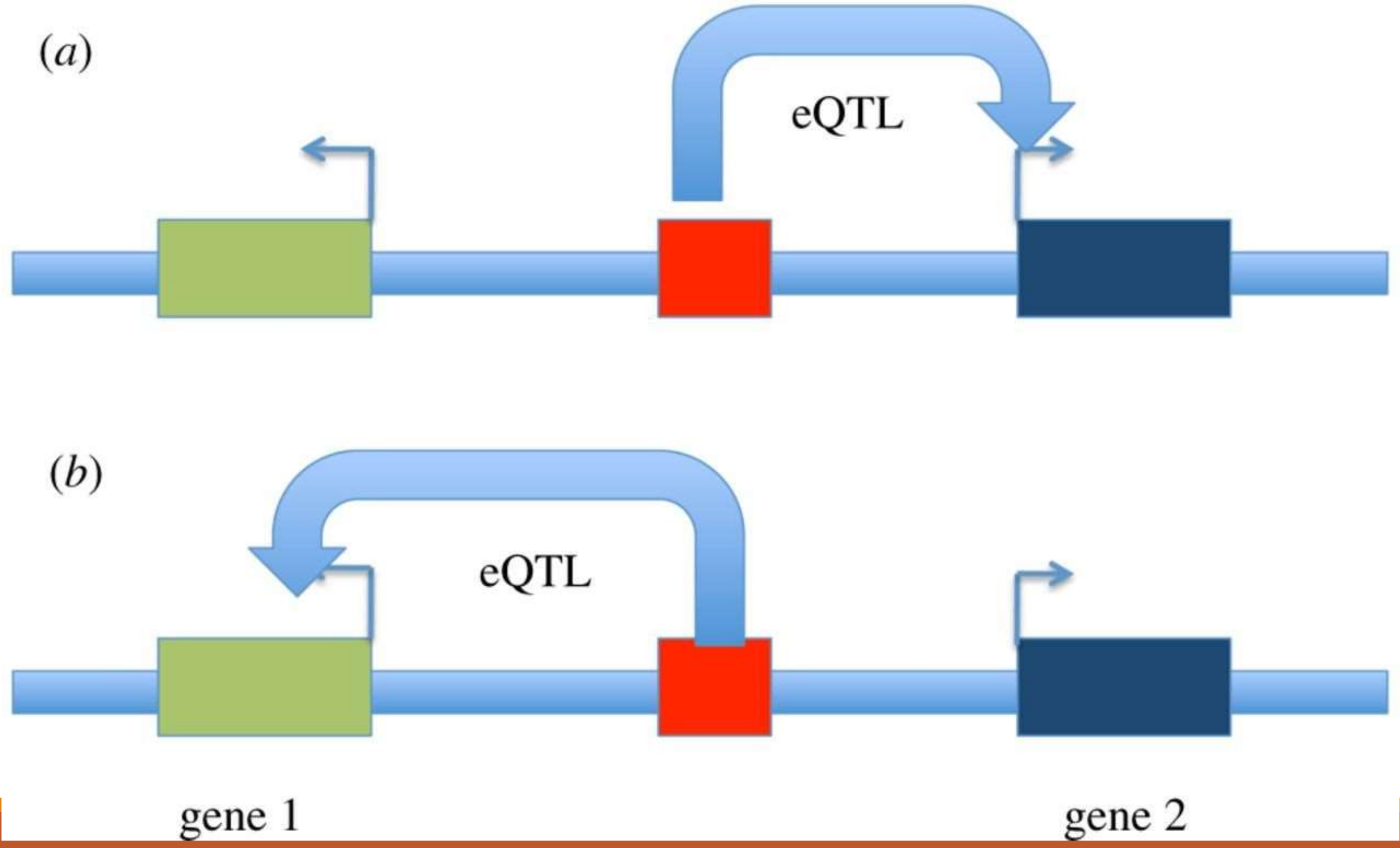
Data: genotyping data and (tissue) expression data

Method: linear regression models


Population
Genotyping
Expression Measurements
Identification of Informative Markers
Pre-processing of Expression Data (background correction, normalization)
eQTL Mapping (single-locus/ multi-loci models)
eQTL Interpretation

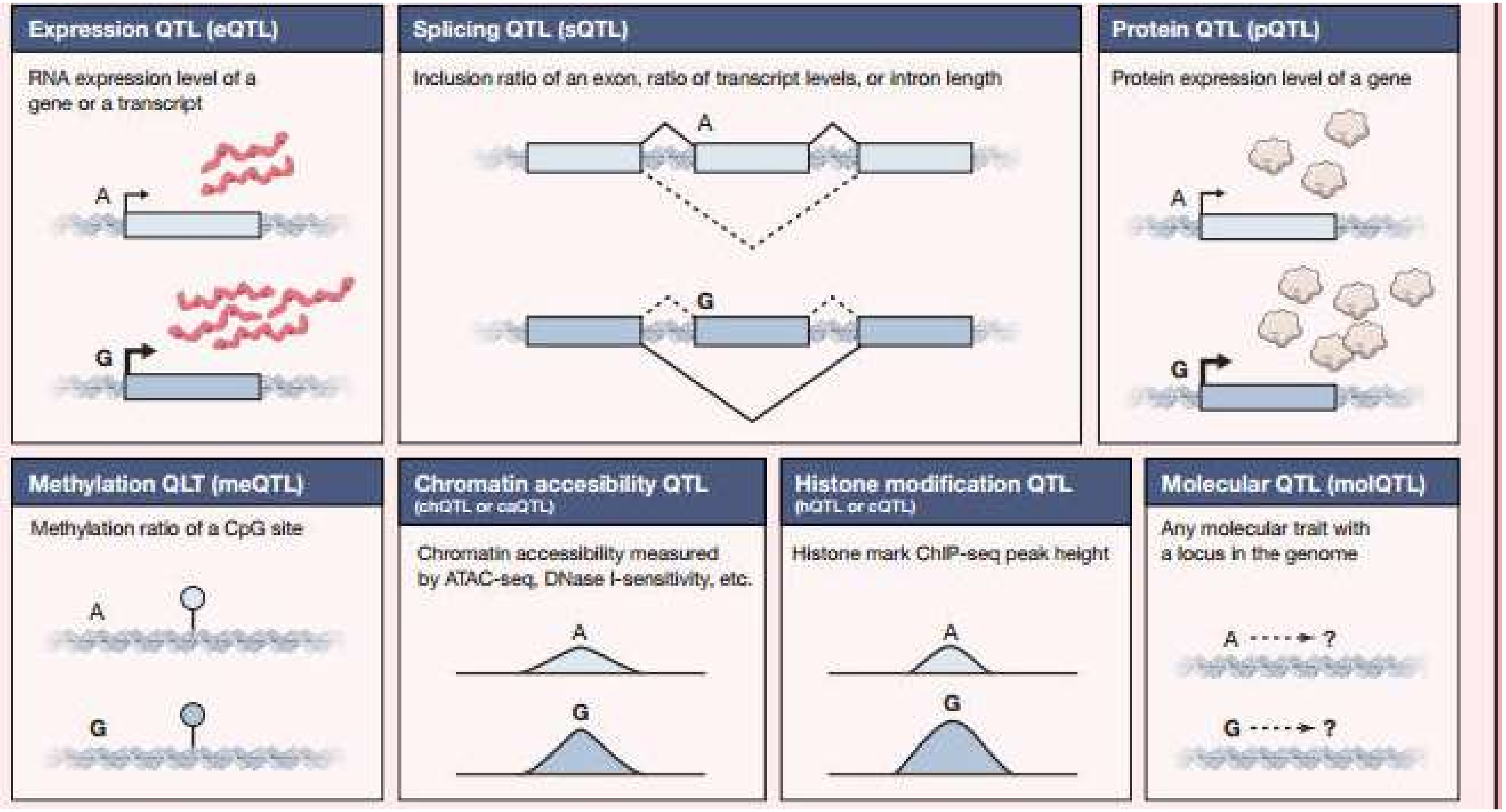

Transcription factor
cis-effect
trans-effect
Genotype address
Transcript address
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
20
22
X
Y

Tissue eQTL



Forms of QTLs

- Quantitative traits
- $y = \beta_0 + \beta_1 \times SNP$
- Y can be any quantitative traits, e.g., gene expression, protein expression, and so on.



Expression QTL (eQTL)
RNA expression level of a gene or a transcript
A
G
Splicing QTL (sQTL)
Inclusion ratio of an exon, ratio of transcript levels, or intron length
A
G
Protein QTL (pQTL)
Protein expression level of a gene
A
G
Methylation QLT (meQTL)
Methylation ratio of a CpG site
A
G
Chromatin accesibility QTL (chQTL or caQTL)
Chromatin accessibility measured by ATAC-seq, DNase I-sensitivity, etc.
A
G
Histone modification QTL (hQTL or cQTL)
Histone mark ChIP-seq peak height
A
G
Molecular QTL (molQTL)
Any molecular trait with a locus in the genome
A ----→ ?
G ----→ ?

Software tools for QTL

PLINK: The basic tool for GWAS
http://pngu.mgh.harvard.edu/~purcell/plink/tutorial.shtml

Matrix eQTL: Ultra-fast eQTL analysis,
http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/

GEMMA: Genome-wide Efficient Mixed Model Association (GEMMA),
http://stephenslab.uchicago.edu/software.html#gemma

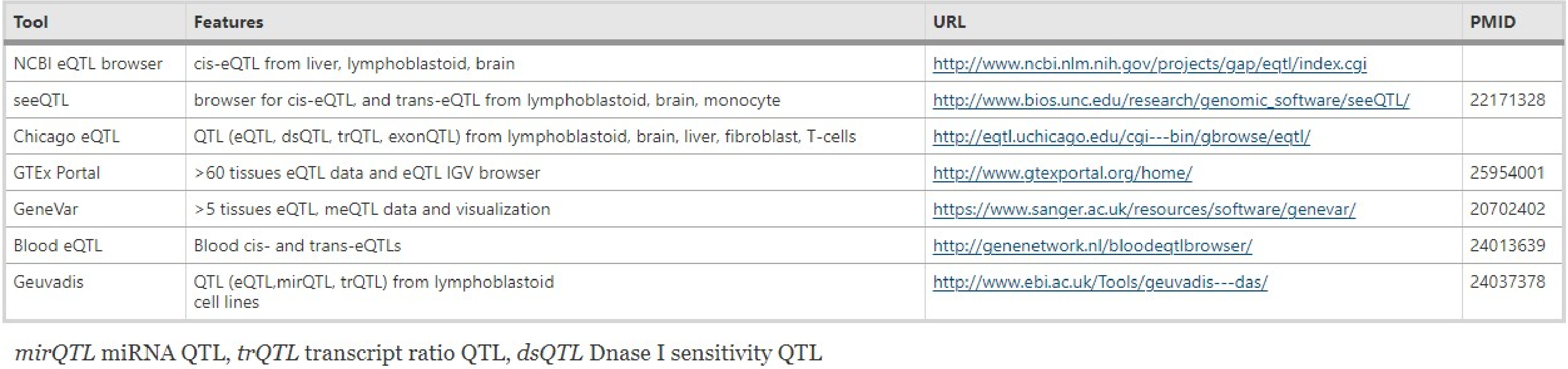
FMeQTL: Bayesian Joint mapping, https://github.com/xqwen/fmeqtl

DAP: Deterministic Approcimation of Posteriors (Fast Bayesian),
https://github.com/xqwen/dap

CAVIAR: Bayesian Fine Mapping, http://genetics.cs.ucla.edu/caviar/

Ventham et al (2016) *Nature Communications* **7**: 13507

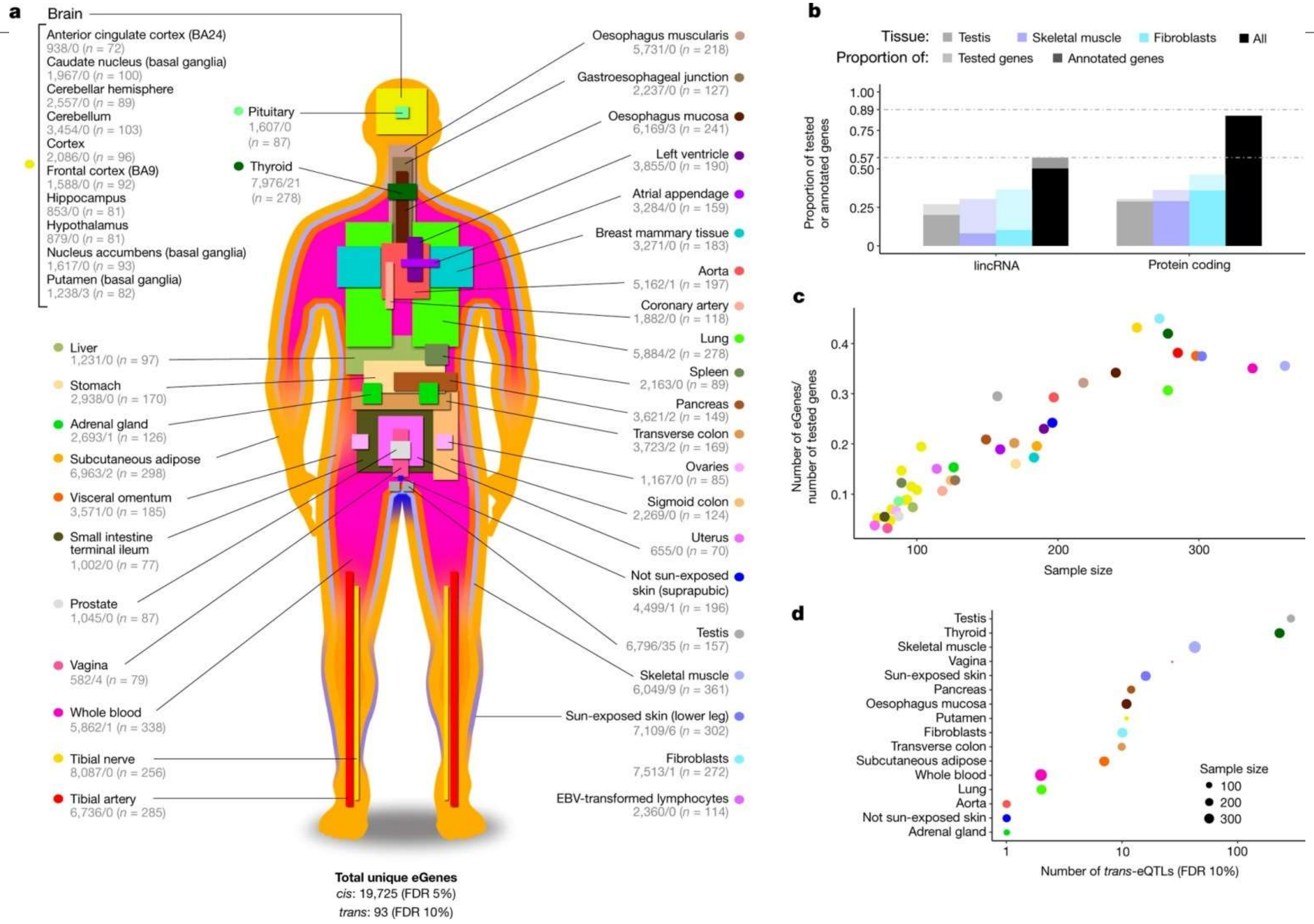
Sources of eQTL databases



Tool	Features	URL	PMID
NCBI eQTL browser	cis-eQTL from liver, lymphoblastoid, brain	http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi	
seeQTL	browser for cis-eQTL, and trans-eQTL from lymphoblastoid, brain, monocyte	http://www.bios.unc.edu/research/genomic_software/seeQTL/	22171328
Chicago eQTL	QTL (eQTL, dsQTL, trQTL, exonQTL) from lymphoblastoid, brain, liver, fibroblast, T-cells	http://eqtl.uchicago.edu/cgi---bin/gbrowse/eqtl/	
GTEx Portal	>60 tissues eQTL data and eQTL IGV browser	http://www.gtexportal.org/home/	25954001
GeneVar	>5 tissues eQTL, meQTL data and visualization	https://www.sanger.ac.uk/resources/software/genevar/	20702402
Blood eQTL	Blood cis- and trans-eQTLs	http://genenetwork.nl/bloodeqtlbrowser/	24013639
Geuvadis	QTL (eQTL,mirQTL, trQTL) from lymphoblastoid cell lines	http://www.ebi.ac.uk/Tools/geuvadis---das/	24037378

mirQTL miRNA QTL, *trQTL* transcript ratio QTL, *dsQTL* Dnase I sensitivity QTL

Sample size and eGene discovery in the GTEx v6p study



A Battle *et al.* *Nature* **550,** 204–213 (2017) doi:10.1038/nature24277

Outline

Genotype imputation

QTL

- QTL introduction
- **Integration of GWAS Variants and eQTLs**

Regulatory roles of genetic variants

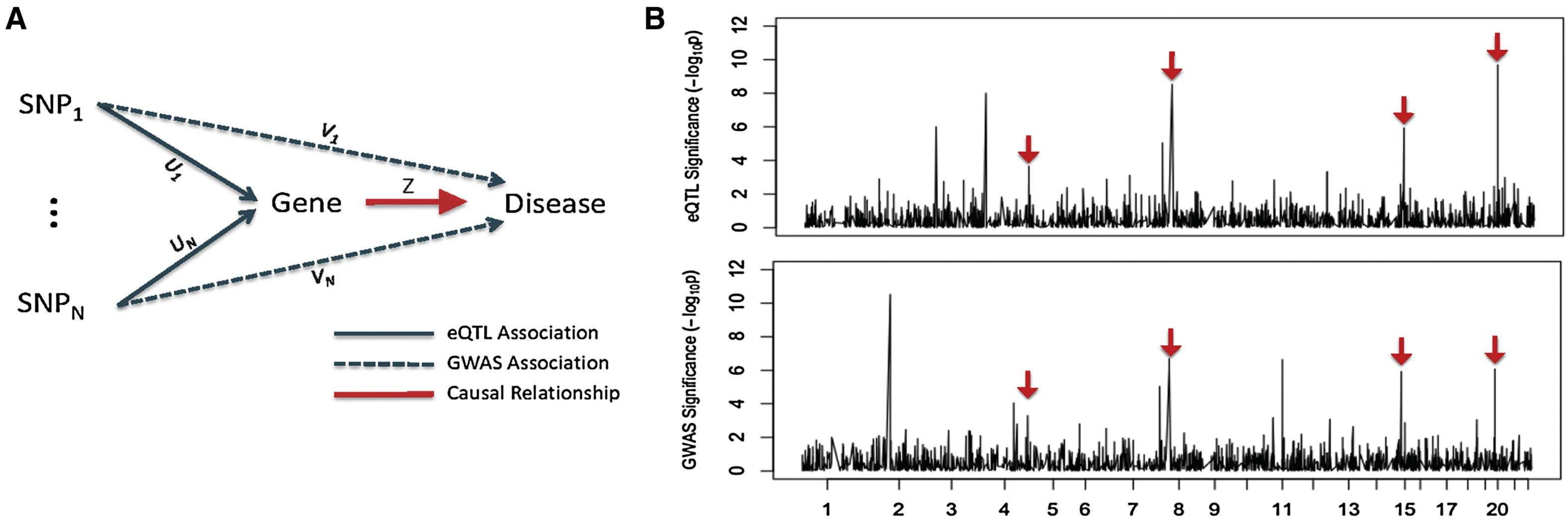
Resources for secondary analyses

Integration of GWAS variants and eQTLs

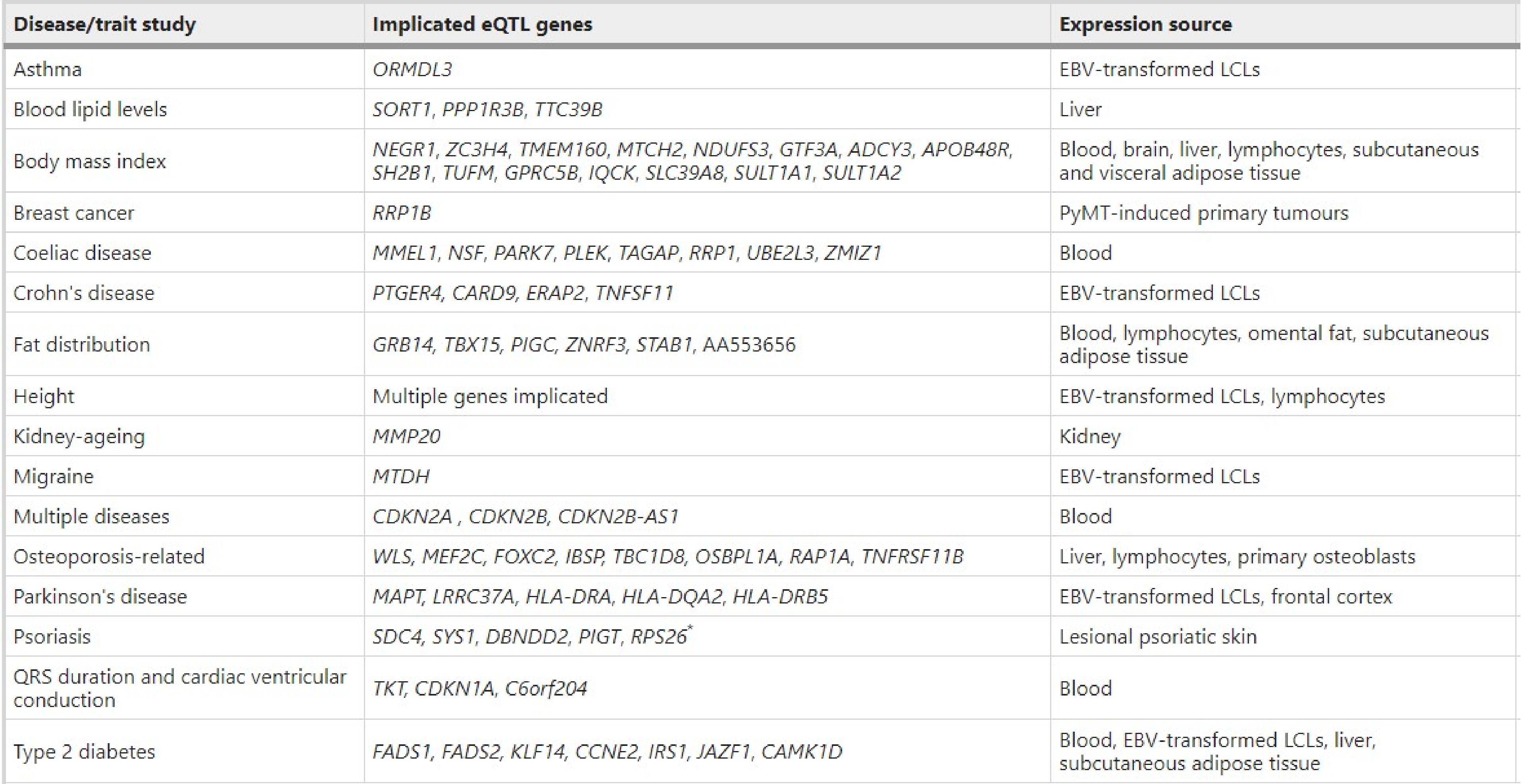
Level 1: overlap

Level 2: enrichment types of analyses

Level 3: colocalization

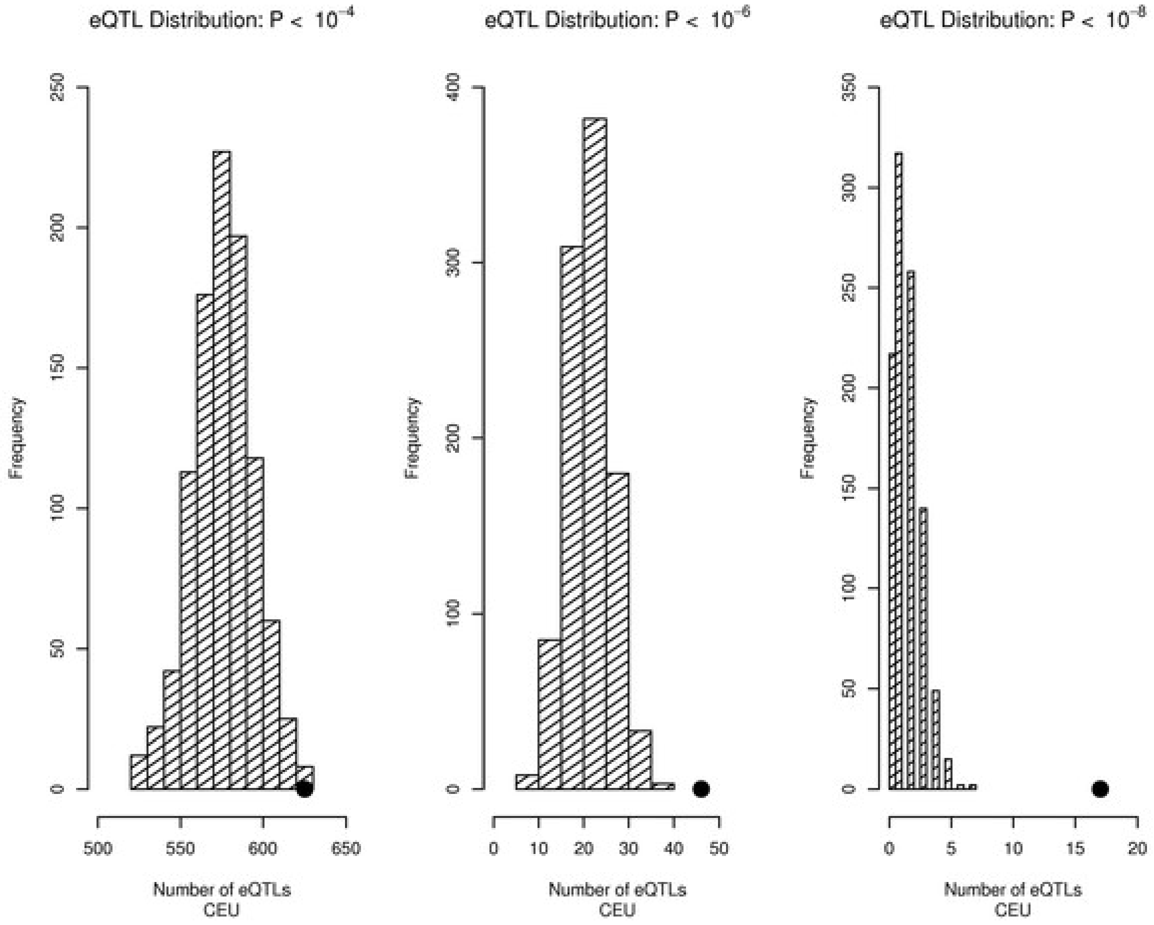


eQTL can aid in identifying candidate genes for GWAS variants



Disease/trait study	Implicated eQTL genes	Expression source
Asthma	*ORMDL3*	EBV-transformed LCLs
Blood lipid levels	*SORT1, PPP1R3B, TTC39B*	Liver
Body mass index	*NEGR1, ZC3H4, TMEM160, MTCH2, NDUFS3, GTF3A, ADCY3, APOB48R, SH2B1, TUFM, GPRC5B, IQCK, SLC39A8, SULT1A1, SULT1A2*	Blood, brain, liver, lymphocytes, subcutaneous and visceral adipose tissue
Breast cancer	*RRP1B*	PyMT-induced primary tumours
Coeliac disease	*MMEL1, NSF, PARK7, PLEK, TAGAP, RRP1, UBE2L3, ZMIZ1*	Blood
Crohn's disease	*PTGER4, CARD9, ERAP2, TNFSF11*	EBV-transformed LCLs
Fat distribution	*GRB14, TBX15, PIGC, ZNRF3, STAB1*, AA553656	Blood, lymphocytes, omental fat, subcutaneous adipose tissue
Height	Multiple genes implicated	EBV-transformed LCLs, lymphocytes
Kidney-ageing	*MMP20*	Kidney
Migraine	*MTDH*	EBV-transformed LCLs
Multiple diseases	*CDKN2A , CDKN2B, CDKN2B-AS1*	Blood
Osteoporosis-related	*WLS, MEF2C, FOXC2, IBSP, TBC1D8, OSBPL1A, RAP1A, TNFRSF11B*	Liver, lymphocytes, primary osteoblasts
Parkinson's disease	*MAPT, LRRC37A, HLA-DRA, HLA-DQA2, HLA-DRB5*	EBV-transformed LCLs, frontal cortex
Psoriasis	*SDC4, SYS1, DBNDD2, PIGT, RPS26**	Lesional psoriatic skin
QRS duration and cardiac ventricular conduction	*TKT, CDKN1A, C6orf204*	Blood
Type 2 diabetes	*FADS1, FADS2, KLF14, CCNE2, IRS1, JAZF1, CAMK1D*	Blood, EBV-transformed LCLs, liver, subcutaneous adipose tissue

Trait-associated SNPs are more likely to be eQTLs



Nicolae DL, et al. (2010) Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLOS Genetics 6(4): e1000888.

Integration of GWAS variants and eQTLs

Level 3: colocalization of pairs of association signals

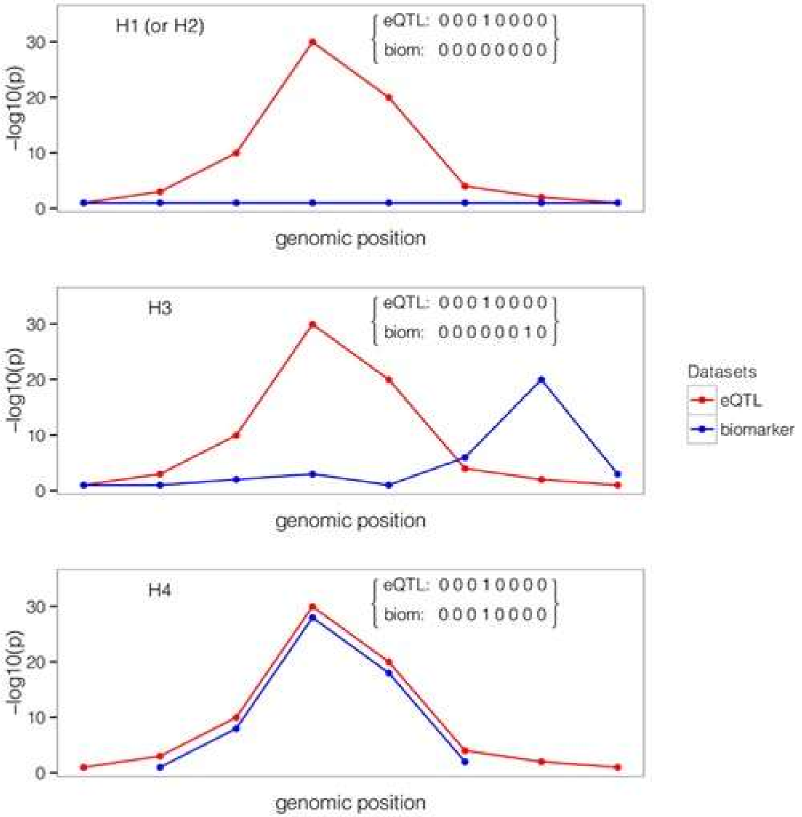
H1 is the hypothesis that there is only an eQTL signal at a locus

H2 is the hypothesis that there is only a GWAS signal at a locus.

H3 is the hypothesis that there are two independent eQTL and GWAS signals in linkage.

H4 is the strong hypothesis that the same SNP (not just the locus) is responsible for both the GWAS and eQTL.

Bayesian analysis evaluate each H relative to the other four and generates a confidence level for the most likely one.



Giambartolomei et al (2014) *PLoS Genetics* **10(5)**: e1004383

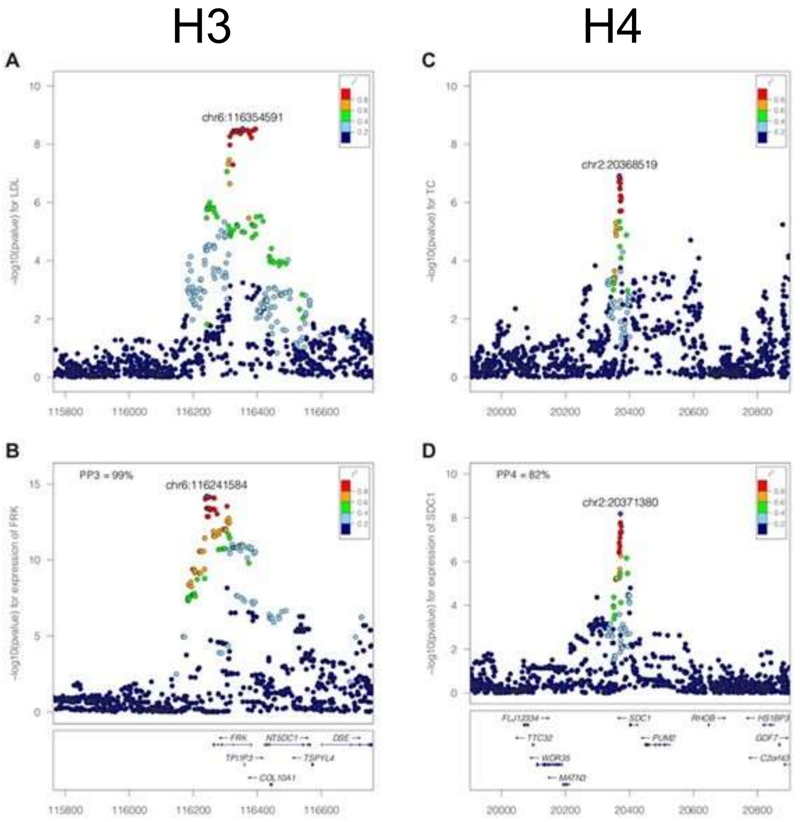
Examples of H3 and H4

On the left, the profile of association at the FRK locus with LDL (top) is very different from that with *FRK* expression.

H3 is the supported hypothesis.

On the right, even though there are two different peak SNPs, they are in the same strong LD region and the profiles are almost the same for Total Cholesterol and *Soc1* expression.

H4 is the supported hypothesis.



Giambartolomei et al (2014) *PLoS Genetics* **10(5)**: e1004383

Outline

Genotype imputation

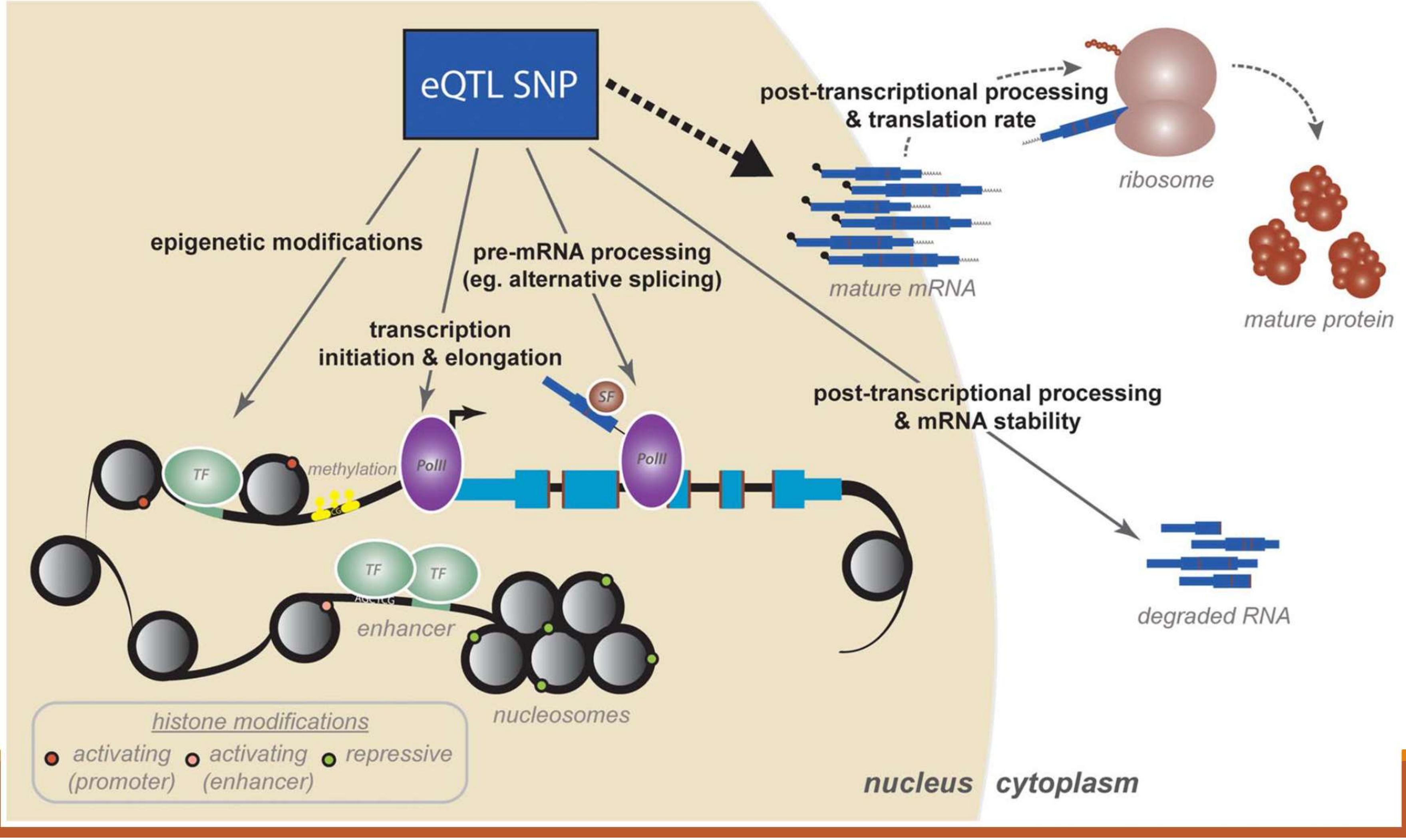
Quantitative Trait Locus (QTL)

Regulatory roles of genetic variants

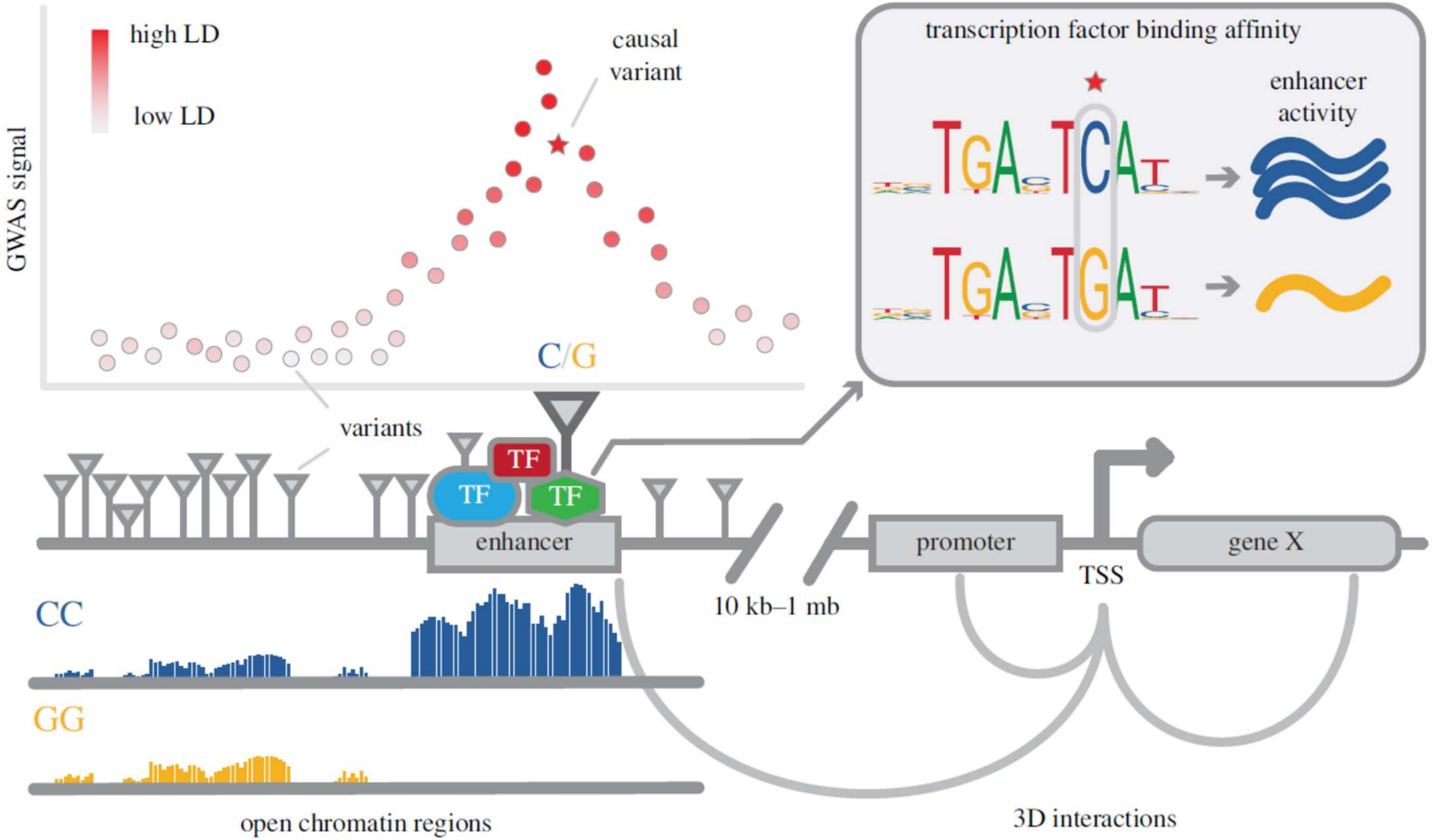
- **Types of regulatory roles**
- **Technologies to detect regulatory regions**
- **Regulation is context-specific**

Resources for secondary analyses

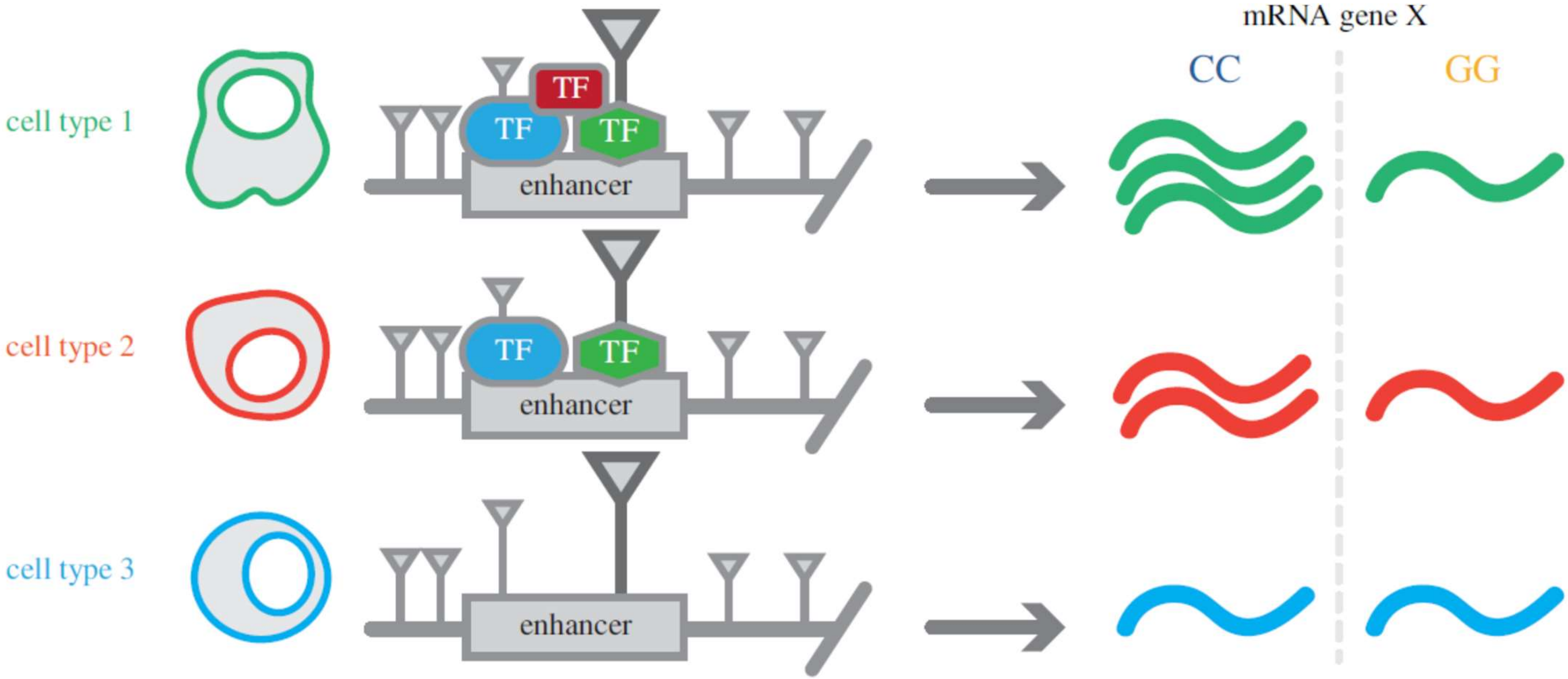
Regulatory roles of genetic variants

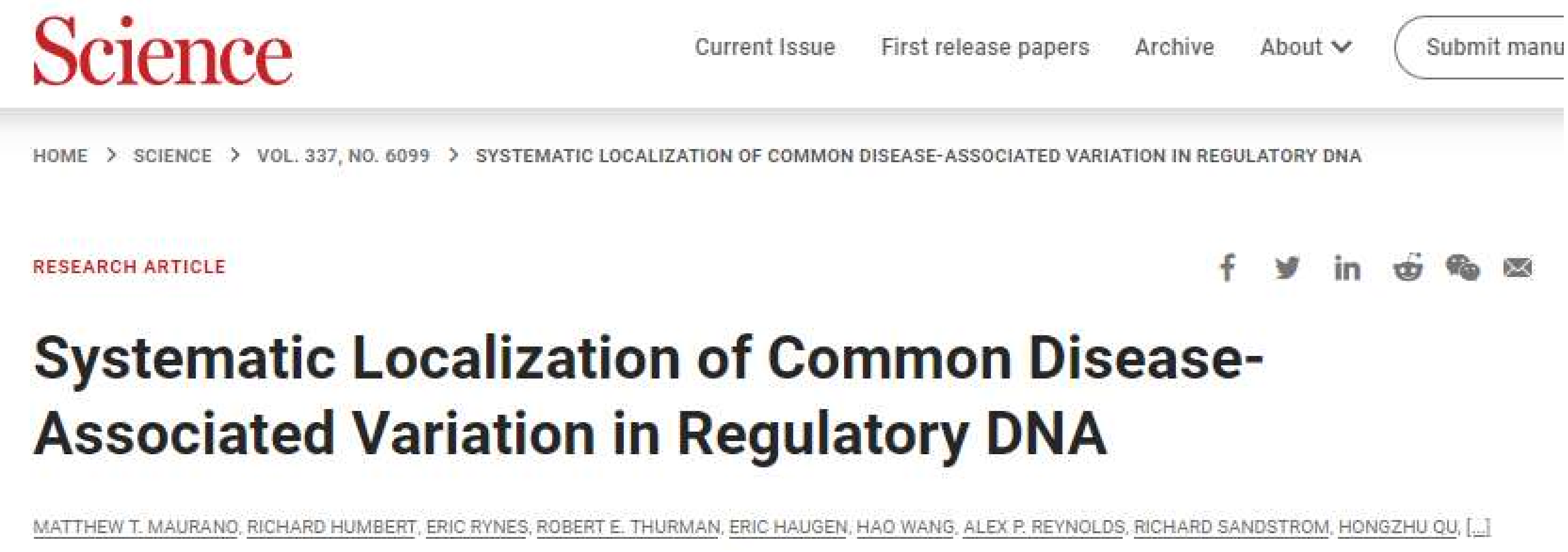


Mechanisms by which SNPs can influence enhancer activity



Cell-type-specific gene-expression differences





Science

HOME > SCIENCE > VOL. 337, NO. 6099 > SYSTEMATIC LOCALIZATION OF COMMON DISEASE-ASSOCIATED VARIATION IN REGULATORY DNA

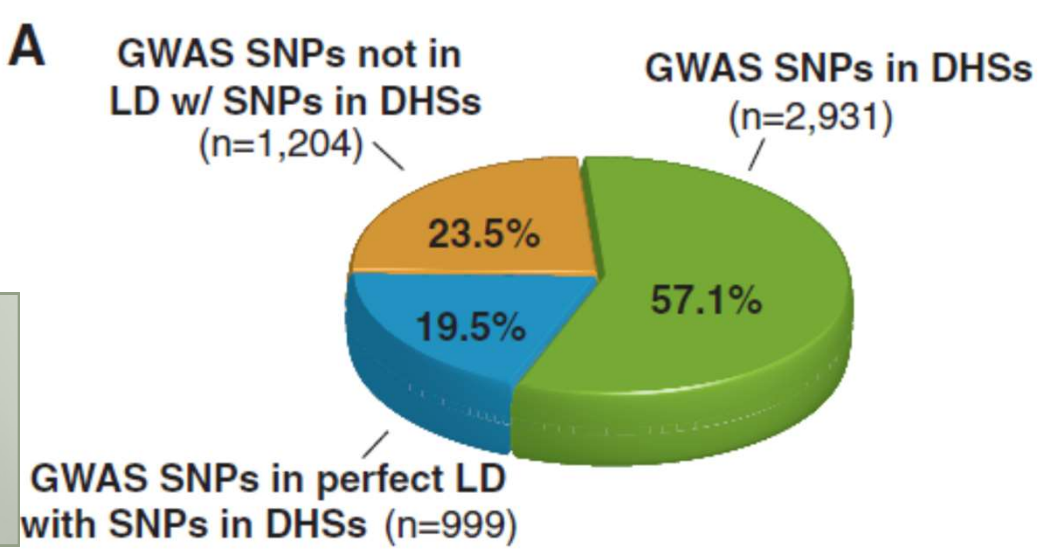
RESEARCH ARTICLE

Systematic Localization of Common Disease-Associated Variation in Regulatory DNA

MATTHEW T. MAURANO, RICHARD HUMBERT, ERIC RYNES, ROBERT E. THURMAN, ERIC HAUGEN, HAO WANG, ALEX P. REYNOLDS, RICHARD SANDSTROM, HONGZHU QU, [...] JOHN A. STAMATOYANNOPOULOS +23 authors Authors Info &

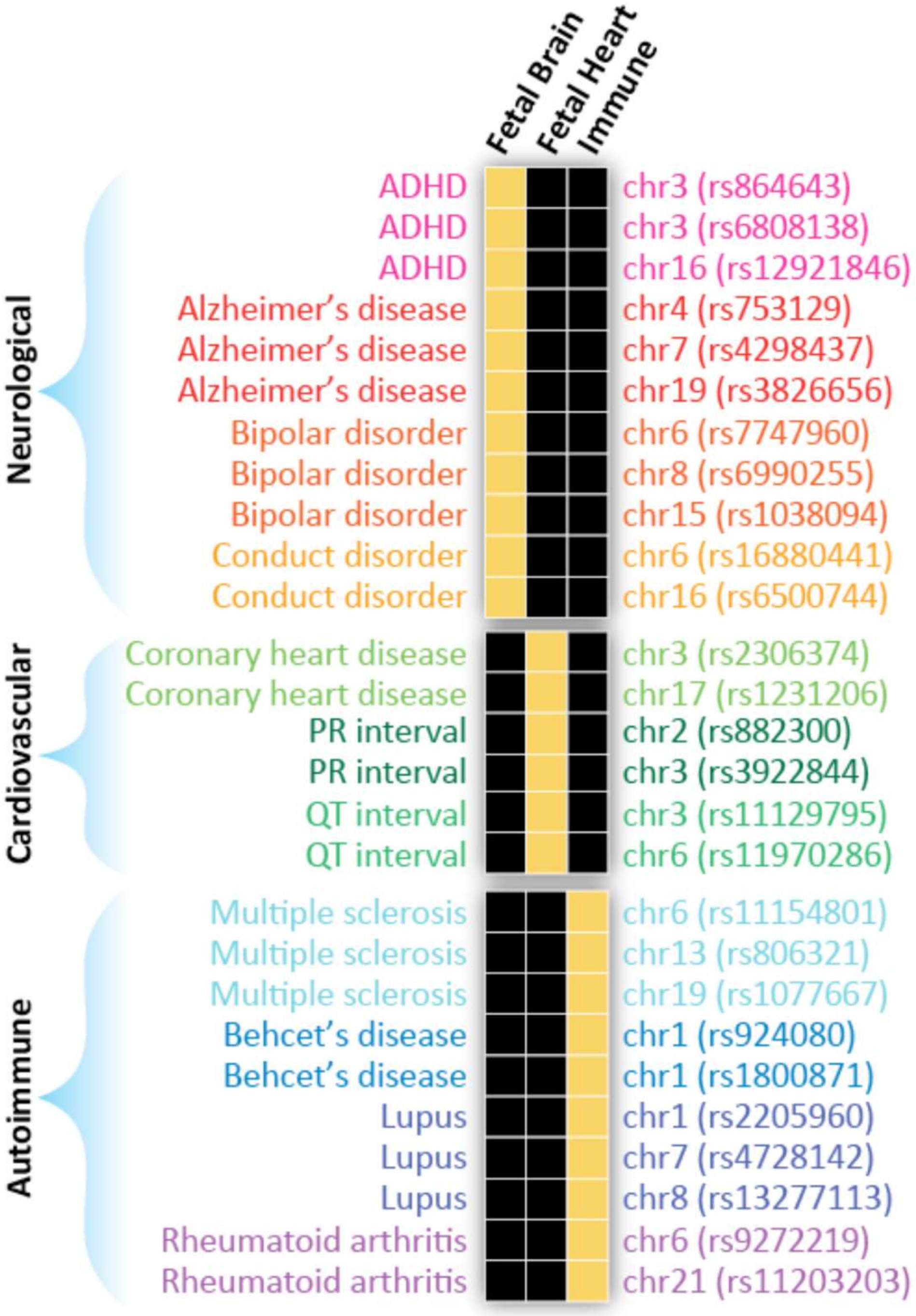
SCIENCE • 7 Sep 2012 • Vol 337, Issue 6099 • pp. 1190-1195 • DOI: 10.

301 1,918



Enriched in regulatory sequences (promoters and enhancers) that are identified through histone mark ChIP-seq or DNase-seq

Multiple distinct genomic disease associations repeatedly localize within **relevant cell-selective DHSs**.



		Fetal Brain	Fetal Heart	Immune	
Neurological	ADHD	🟨	⬛	⬛	chr3 (rs864643)
	ADHD	🟨	⬛	⬛	chr3 (rs6808138)
	ADHD	🟨	⬛	⬛	chr16 (rs12921846)
	Alzheimer's disease	🟨	⬛	⬛	chr4 (rs753129)
	Alzheimer's disease	🟨	⬛	⬛	chr7 (rs4298437)
	Alzheimer's disease	🟨	⬛	⬛	chr19 (rs3826656)
	Bipolar disorder	🟨	⬛	⬛	chr6 (rs7747960)
	Bipolar disorder	🟨	⬛	⬛	chr8 (rs6990255)
	Bipolar disorder	🟨	⬛	⬛	chr15 (rs1038094)
	Conduct disorder	🟨	⬛	⬛	chr6 (rs16880441)
	Conduct disorder	🟨	⬛	⬛	chr16 (rs6500744)
Cardiovascular	Coronary heart disease	⬛	🟨	⬛	chr3 (rs2306374)
	Coronary heart disease	⬛	🟨	⬛	chr17 (rs1231206)
	PR interval	⬛	🟨	⬛	chr2 (rs882300)
	PR interval	⬛	🟨	⬛	chr3 (rs3922844)
	QT interval	⬛	🟨	⬛	chr3 (rs11129795)
	QT interval	⬛	🟨	⬛	chr6 (rs11970286)
Autoimmune	Multiple sclerosis	⬛	⬛	🟨	chr6 (rs11154801)
	Multiple sclerosis	⬛	⬛	🟨	chr13 (rs806321)
	Multiple sclerosis	⬛	⬛	🟨	chr19 (rs1077667)
	Behcet's disease	⬛	⬛	🟨	chr1 (rs924080)
	Behcet's disease	⬛	⬛	🟨	chr1 (rs1800871)
	Lupus	⬛	⬛	🟨	chr1 (rs2205960)
	Lupus	⬛	⬛	🟨	chr7 (rs4728142)
	Lupus	⬛	⬛	🟨	chr8 (rs13277113)
	Rheumatoid arthritis	⬛	⬛	🟨	chr6 (rs9272219)
	Rheumatoid arthritis	⬛	⬛	🟨	chr21 (rs11203203)

Technologies to detect regulatory regions

NGS Technologies for Epigenome Regulators

DNA methylation

- Whole genome bisulfite sequencing

DNA-protein interaction

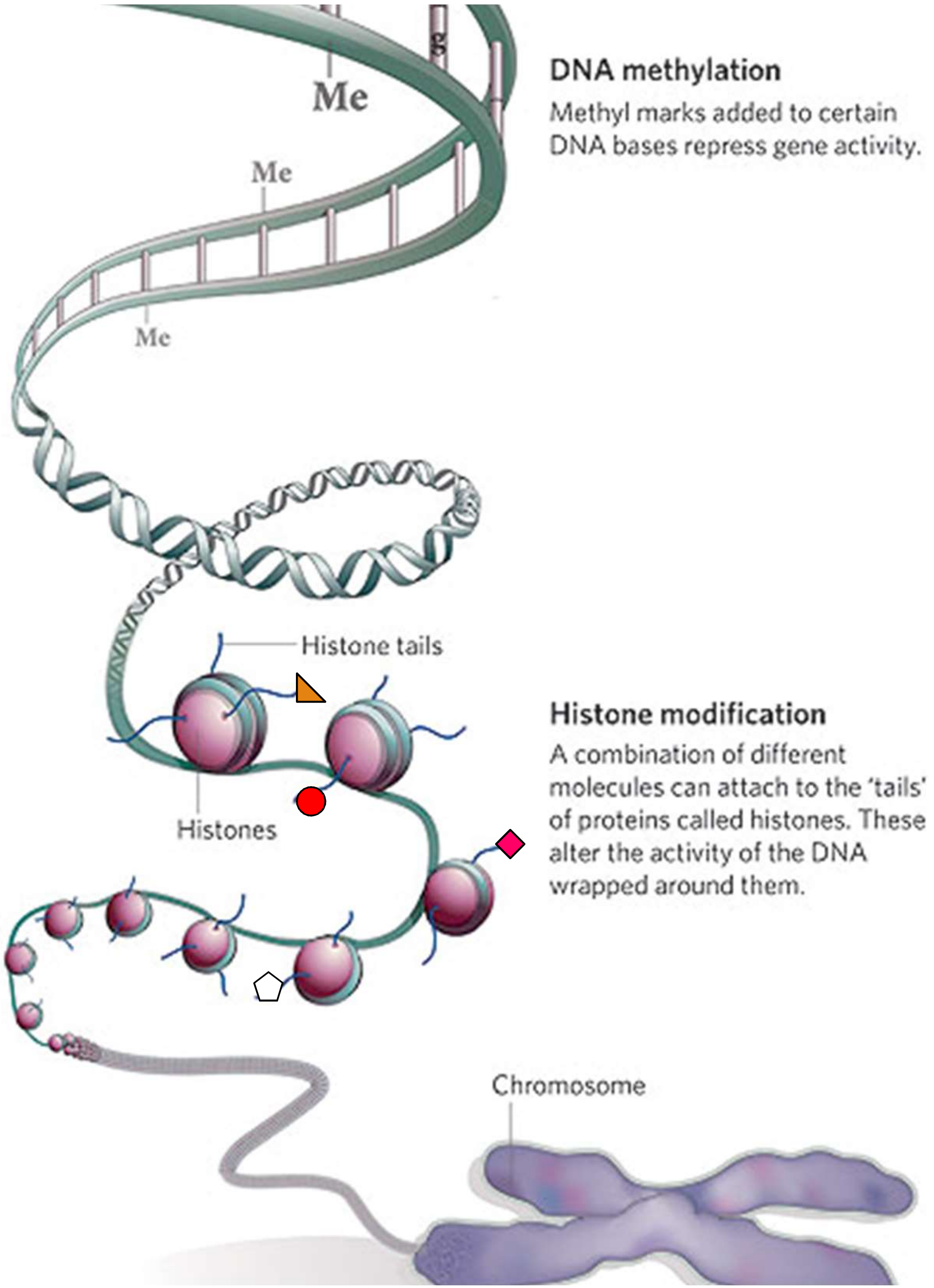
- ChIP-seq TF
- ChIP-seq histone marks

Chromatin accessibility

- ATAC-seq
- DNase-seq
- FAIRE-seq
- MNase-seq

Chromosomal interaction

- Hi-C
- ChIA-PET



DNA methylation

Methyl marks added to certain DNA bases repress gene activity.

Histone modification

A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.

ChIP-seq

Regular TF ChIP-seq: sonication, antibody against TF

Histone mark ChIP-seq: sonication or MNase, antibody against the histone modification

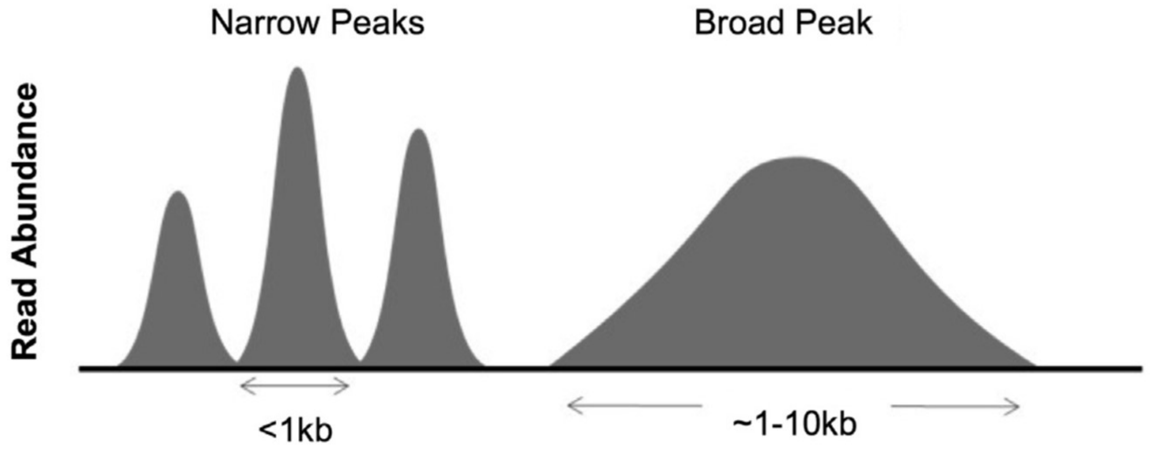


Illustration of two types of peaks in the ChIP-seq datasets. Narrow peaks are generally associated with TF binding, and broad peaks indicate regions with histone modification marks.

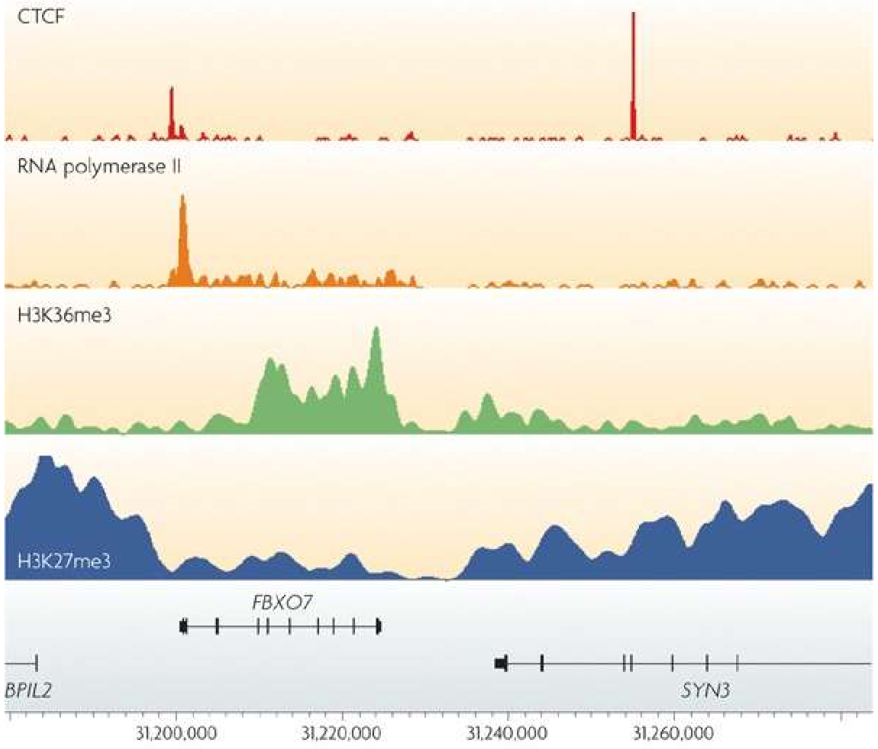
Example of peaks

Insulator binding protein CTCF: sharp binding sites

A mixture of shapes, such as RNA polymerase II (orange), which has a sharp peak followed by a broad region of enrichment;

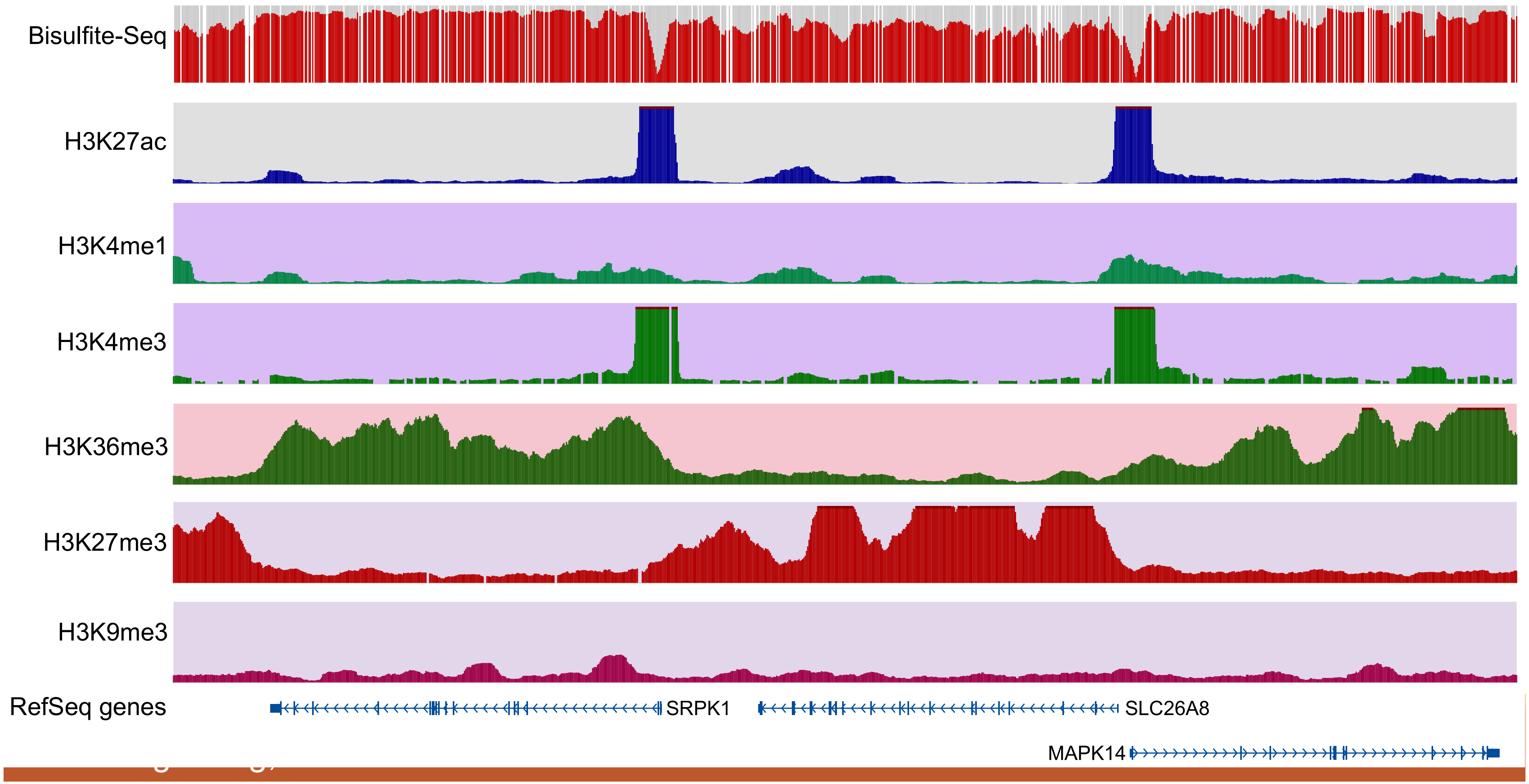
Medium size broad peaks, such as histone H3 trimethylated at lysine 36 (H3K36me3; green), which is associated with transcription elongation over the gene;

Large domains, as shown for histone H3 trimethylated at lysine 27 (H3K27me3; blue), which is a repressive mark that is indicative of Polycomb-mediated silencing



Nature Reviews | Genetics

Histone Modifications in Relation to Gene Transcription



Histone Modifications

Gene body mark: H3K36me3, H3K79me3

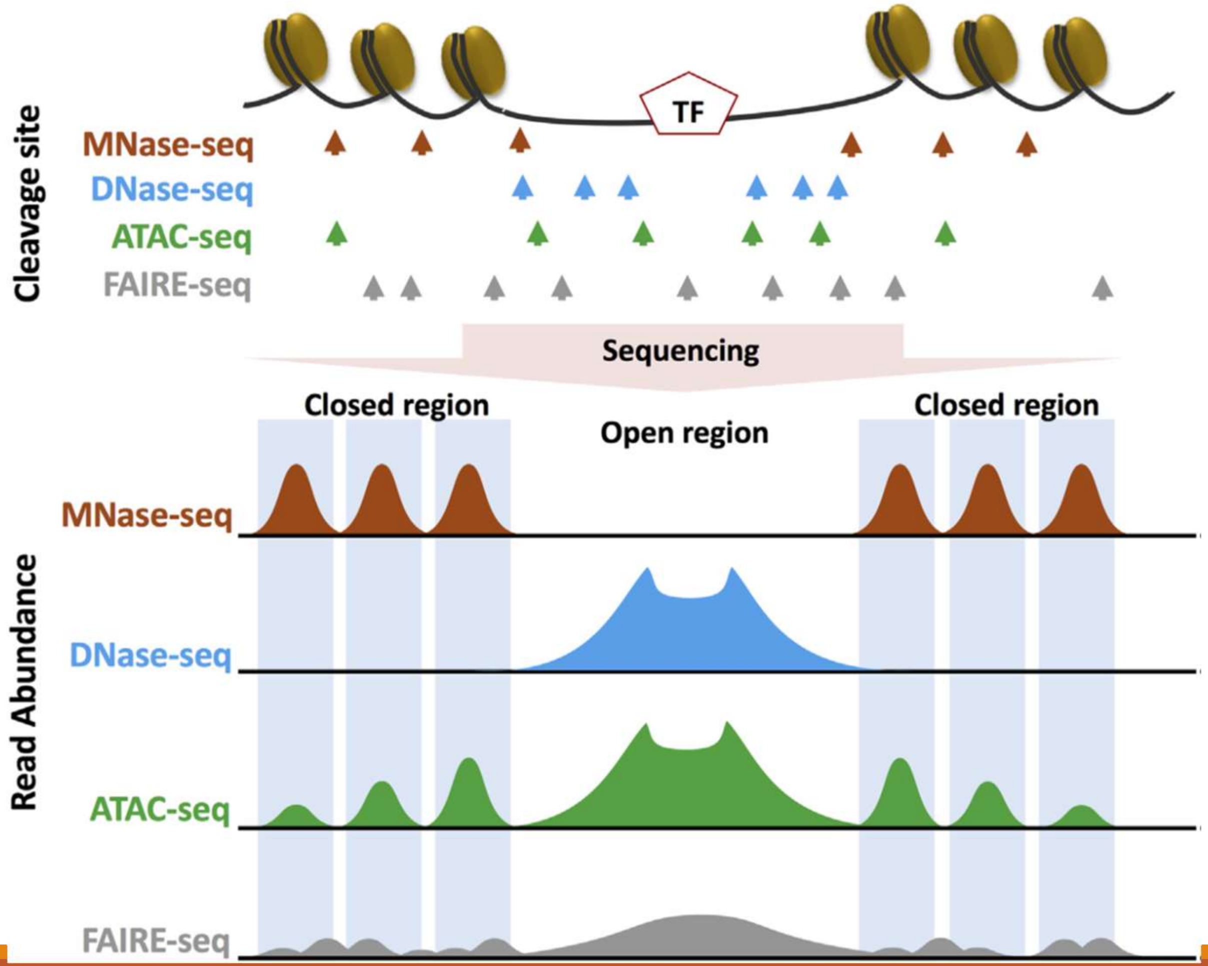
Active promoter (TSS) mark: H3K4me3

Active enhancer (TF binding) mark: H3K4me1, H3K27ac

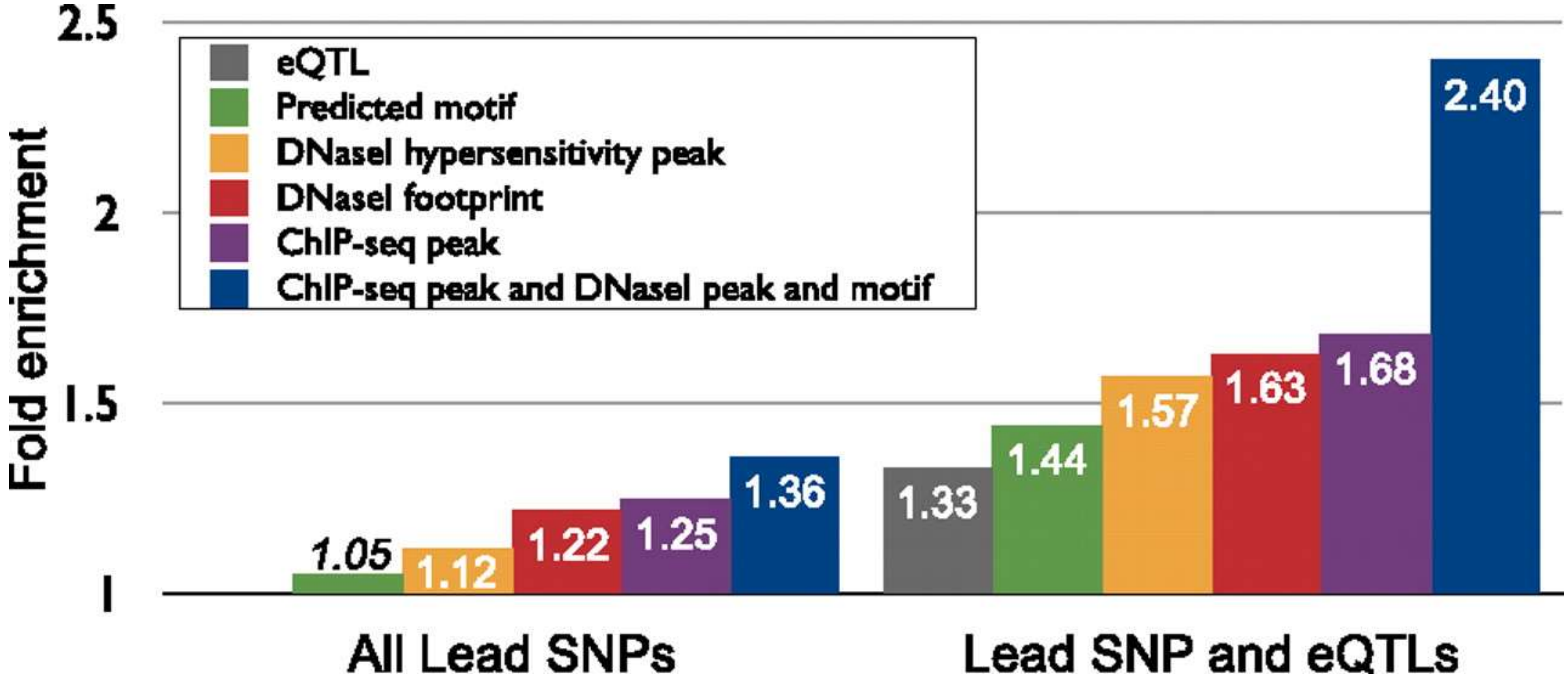
Both enhancers and promoters: H3K4me2, H3/H4ac, H2AZ

Repressive mark: H3K27me3, H3K9me3

Assessing Chromatin Accessibility With Different NGS Techniques



Overview of Enrichment for Different Combinations of Assays



Enrichments are reported for all lead SNPs associated with a phenotype and separately for lead SNPs that are also eQTLs or in strong linkage disequilibrium with an eQTL. The enrichment for predicted motifs alone (italics) is not significant. These results show that combining multiple types of experimental evidence increases the observed enrichment.

Marc A. Schaub et al. Genome Res. 2012;22:1748-1759

Outline

Genotype imputation

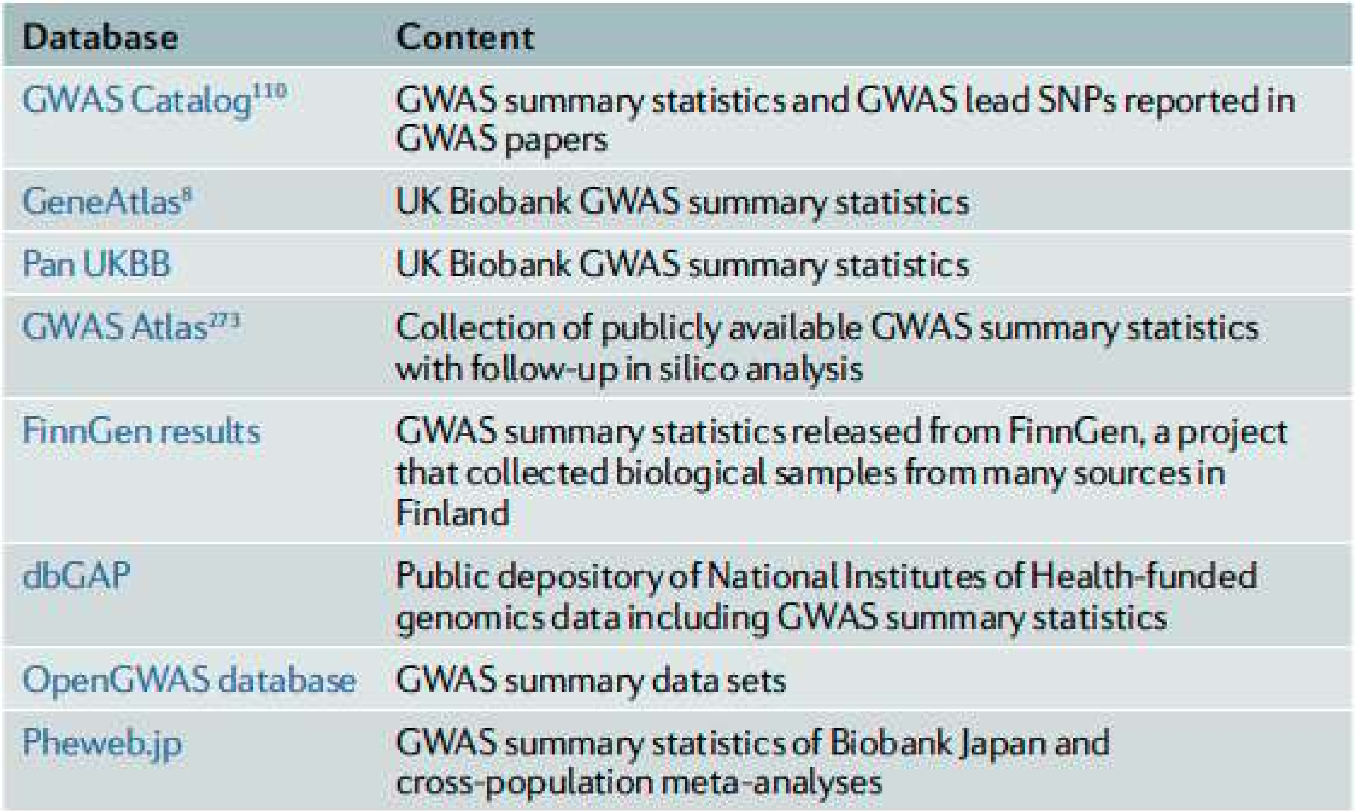
QTL

Regulatory roles of genetic variants

Resources for secondary analyses

- **GTEx: tissue transcriptomes**
- **Roadmap and ENCODE**
- **Biobank**

Databases of GWAS summary statistics



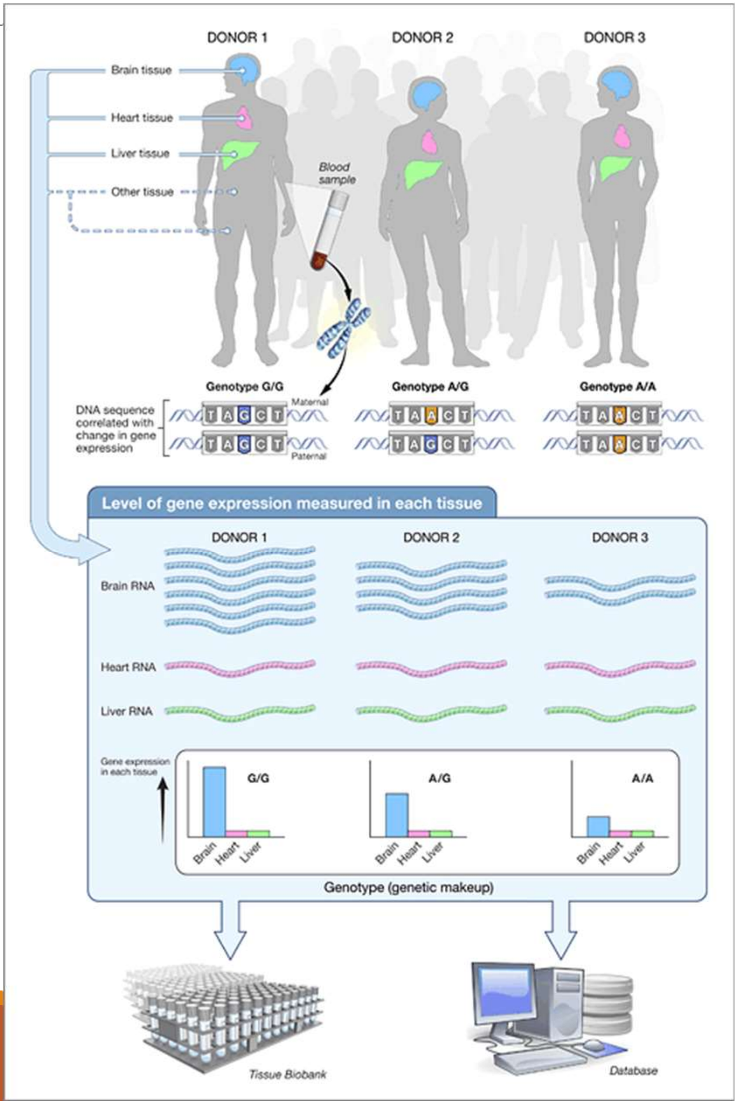
Database	Content
GWAS Catalog[110]	GWAS summary statistics and GWAS lead SNPs reported in GWAS papers
GeneAtlas[8]	UK Biobank GWAS summary statistics
Pan UKBB	UK Biobank GWAS summary statistics
GWAS Atlas[273]	Collection of publicly available GWAS summary statistics with follow-up in silico analysis
FinnGen results	GWAS summary statistics released from FinnGen, a project that collected biological samples from many sources in Finland
dbGAP	Public depository of National Institutes of Health-funded genomics data including GWAS summary statistics
OpenGWAS database	GWAS summary data sets
Pheweb.jp	GWAS summary statistics of Biobank Japan and cross-population meta-analyses

GTEx

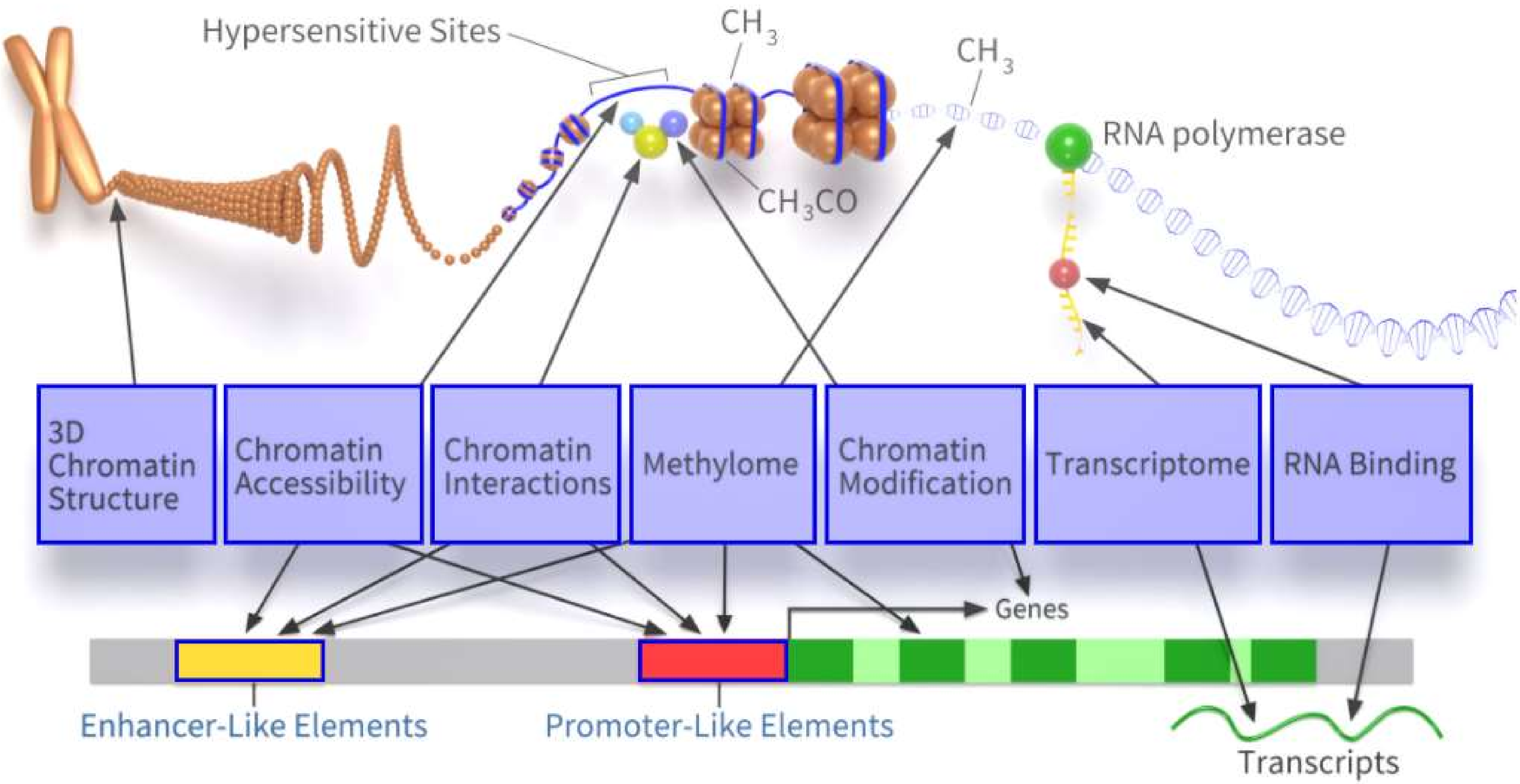
Correlations between genotype and tissue-specific gene expression levels

Tissue expression and eQTLs

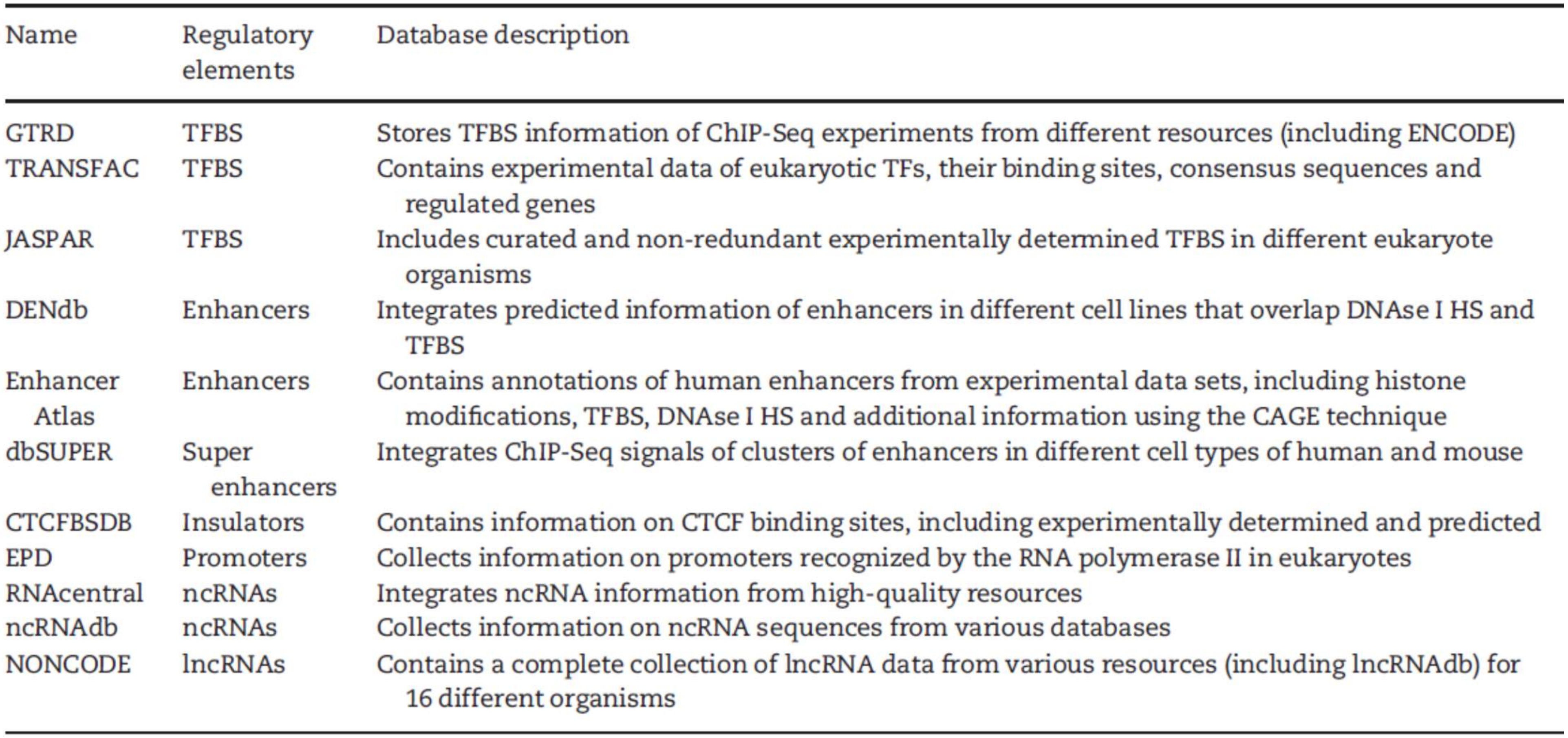
GTEx Single Cell Data (ongoing)



Roadmap/ENCODE

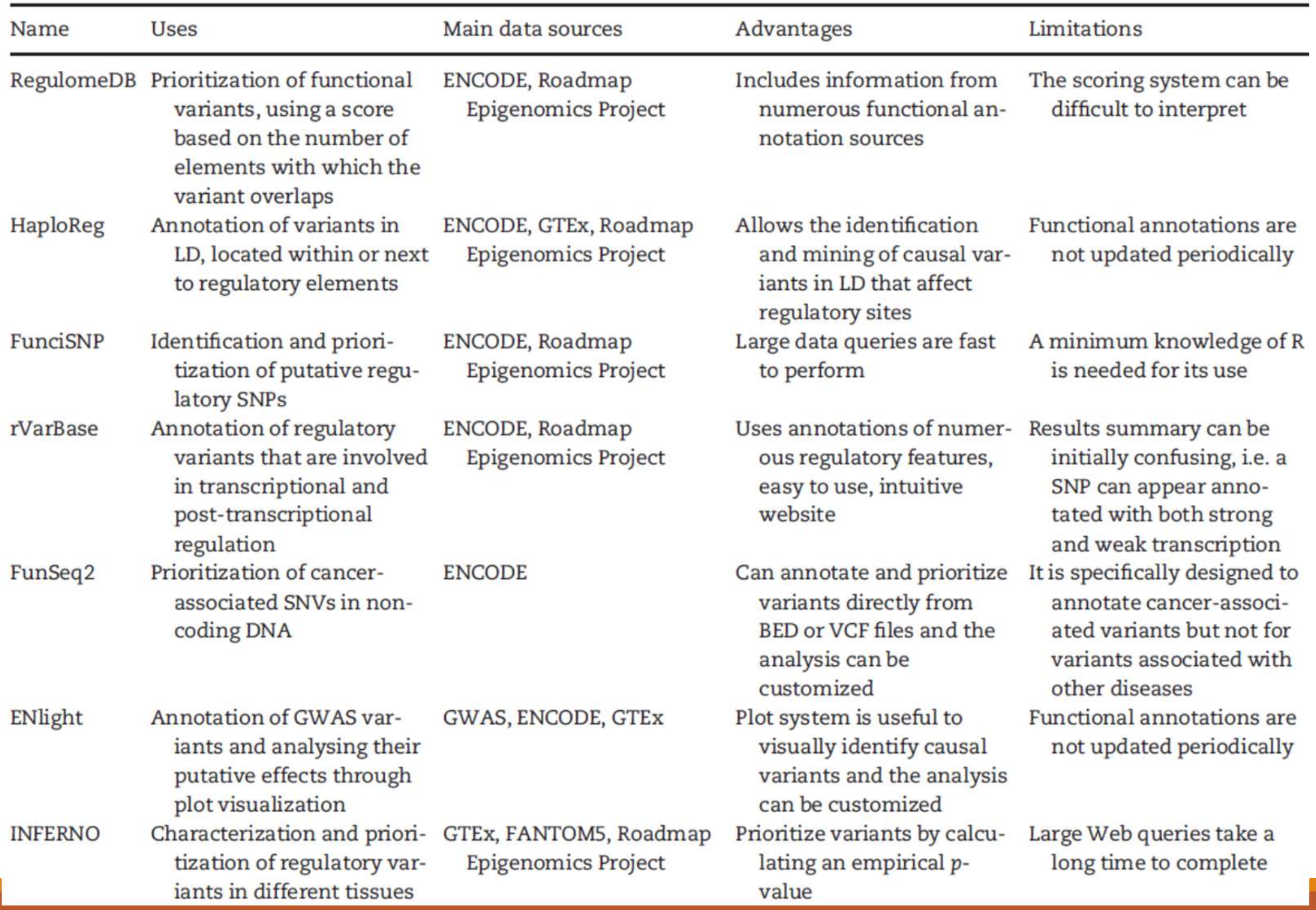


Databases for annotations of regulatory elements



Name	Regulatory elements	Database description
GTRD	TFBS	Stores TFBS information of ChIP-Seq experiments from different resources (including ENCODE)
TRANSFAC	TFBS	Contains experimental data of eukaryotic TFs, their binding sites, consensus sequences and regulated genes
JASPAR	TFBS	Includes curated and non-redundant experimentally determined TFBS in different eukaryote organisms
DENdb	Enhancers	Integrates predicted information of enhancers in different cell lines that overlap DNAse I HS and TFBS
Enhancer Atlas	Enhancers	Contains annotations of human enhancers from experimental data sets, including histone modifications, TFBS, DNAse I HS and additional information using the CAGE technique
dbSUPER	Super enhancers	Integrates ChIP-Seq signals of clusters of enhancers in different cell types of human and mouse
CTCFBSDB	Insulators	Contains information on CTCF binding sites, including experimentally determined and predicted
EPD	Promoters	Collects information on promoters recognized by the RNA polymerase II in eukaryotes
RNAcentral	ncRNAs	Integrates ncRNA information from high-quality resources
ncRNAdb	ncRNAs	Collects information on ncRNA sequences from various databases
NONCODE	lncRNAs	Contains a complete collection of lncRNA data from various resources (including lncRNAdb) for 16 different organisms

Annotation tools for non-coding DNA regions



Name	Uses	Main data sources	Advantages	Limitations
RegulomeDB	Prioritization of functional variants, using a score based on the number of elements with which the variant overlaps	ENCODE, Roadmap Epigenomics Project	Includes information from numerous functional annotation sources	The scoring system can be difficult to interpret
HaploReg	Annotation of variants in LD, located within or next to regulatory elements	ENCODE, GTEx, Roadmap Epigenomics Project	Allows the identification and mining of causal variants in LD that affect regulatory sites	Functional annotations are not updated periodically
FunciSNP	Identification and prioritization of putative regulatory SNPs	ENCODE, Roadmap Epigenomics Project	Large data queries are fast to perform	A minimum knowledge of R is needed for its use
rVarBase	Annotation of regulatory variants that are involved in transcriptional and post-transcriptional regulation	ENCODE, Roadmap Epigenomics Project	Uses annotations of numerous regulatory features, easy to use, intuitive website	Results summary can be initially confusing, i.e. a SNP can appear annotated with both strong and weak transcription
FunSeq2	Prioritization of cancer-associated SNVs in non-coding DNA	ENCODE	Can annotate and prioritize variants directly from BED or VCF files and the analysis can be customized	It is specifically designed to annotate cancer-associated variants but not for variants associated with other diseases
ENlight	Annotation of GWAS variants and analysing their putative effects through plot visualization	GWAS, ENCODE, GTEx	Plot system is useful to visually identify causal variants and the analysis can be customized	Functional annotations are not updated periodically
INFERNO	Characterization and prioritization of regulatory variants in different tissues	GTEx, FANTOM5, Roadmap Epigenomics Project	Prioritize variants by calculating an empirical *p*-value	Large Web queries take a long time to complete

Biobanks

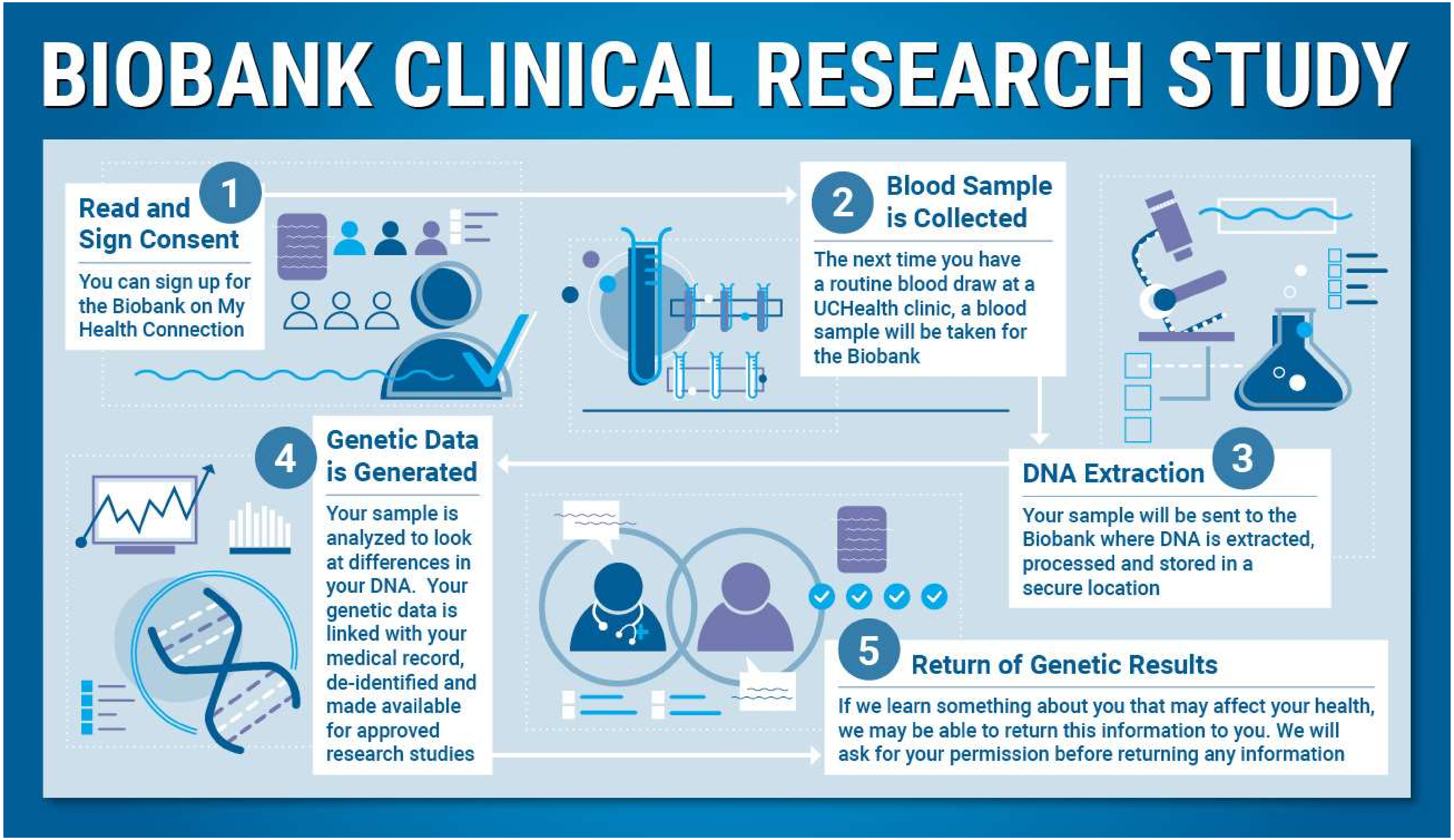
A biobank is a type of biorepository that stores biological samples (usually human) for use in research.

Biobanks have become an important resource in medical research, supporting many types of contemporary research like genomics and personalized medicine.

UK Biobank

Japan Biobank

Biobanks



BIOBANK CLINICAL RESEARCH STUDY

1 **Read and Sign Consent**

You can sign up for the Biobank on My Health Connection

2 **Blood Sample is Collected**

The next time you have a routine blood draw at a UCHealth clinic, a blood sample will be taken for the Biobank

3 **DNA Extraction**

Your sample will be sent to the Biobank where DNA is extracted, processed and stored in a secure location

4 **Genetic Data is Generated**

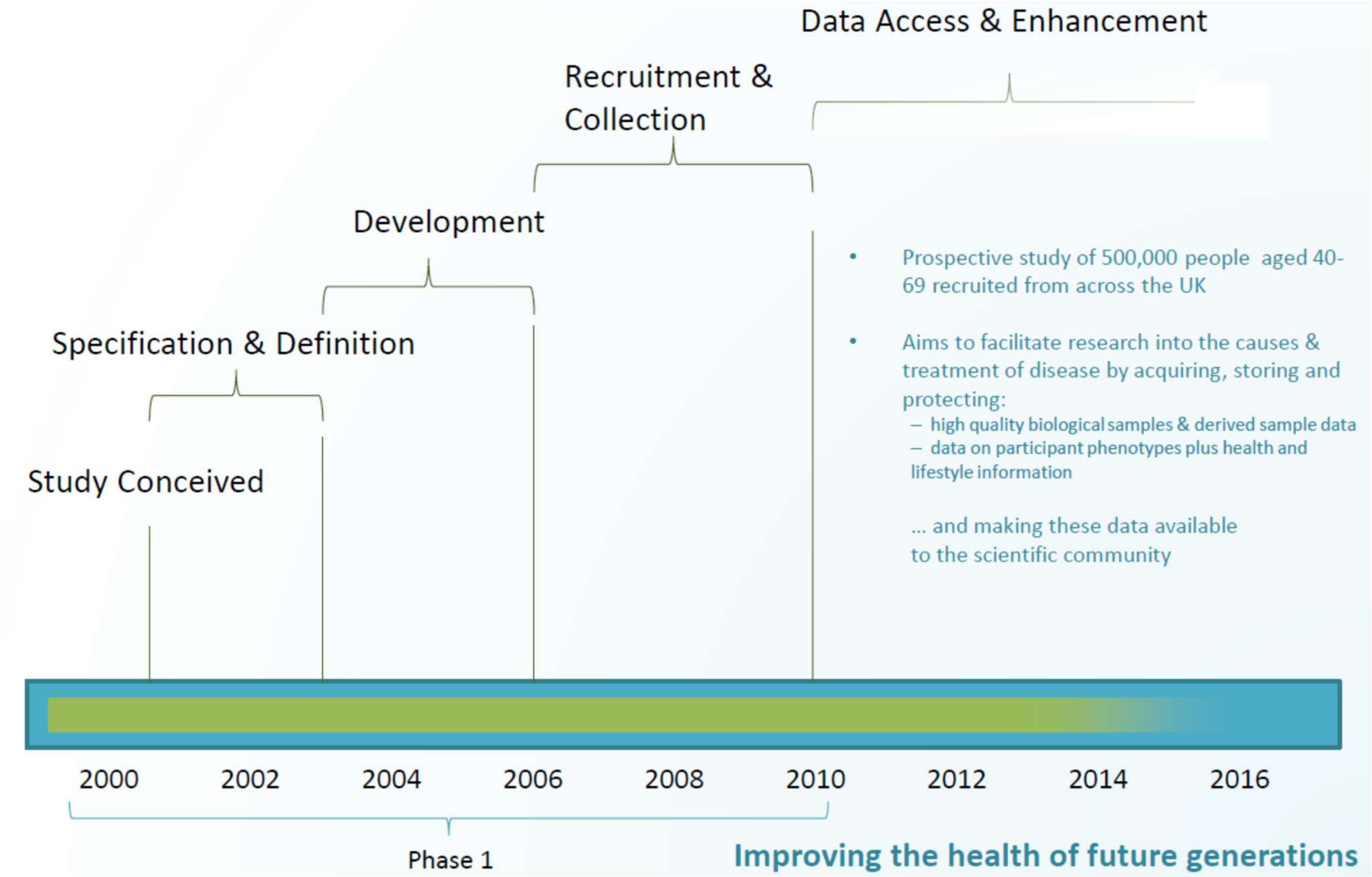
Your sample is analyzed to look at differences in your DNA. Your genetic data is linked with your medical record, de-identified and made available for approved research studies

5 **Return of Genetic Results**

If we learn something about you that may affect your health, we may be able to return this information to you. We will ask for your permission before returning any information

Biobank	Affiliation	Focus	Type	Location
All of Us		Population	non-profit	United States
BioBank Graz	Medical University of Graz		non-profit	Austria
BioBank Japan	RIKEN, University of Tokyo	Population, personalized medicine	non-profit	Japan
Canadian Biosample Repository	University of Alberta		non-profit	Canada
CARTaGENE biobank	Centre hospitalier universitaire Sainte-Justine		non-profit	Quebec
FINBB		Population	non-profit	Finland
FinnGen		Population, disease focused	public-private	Finland
Generation Scotland	NHS Scotland		government	Scotland
HUNT Biobank	Norwegian University of Science and Technology		non-profit	Norway
Plasma Services Group		Autoimmune, Infectious, Coagulation, Diagnostics	commercial	United States
The Malaysian Cohort	National University of Malaysia		non-profit	Malaysia
UK Biobank			non-profit	United Kingdom
Sapien Biosciences	Apollo Hospitals & Saarum Innovations	Population, with special focus on tailoring treatment for Cancer	private	India (headquartered in Hyderabad)
Lifelines	University of Groningen & University Medical Centre Groningen	Healthy aging	non-profit	Groningen, The Netherlands

UK Biobank: a prospective cohort epidemiology study



- Prospective study of 500,000 people aged 40-69 recruited from across the UK
- Aims to facilitate research into the causes & treatment of disease by acquiring, storing and protecting:
  - high quality biological samples & derived sample data
  - data on participant phenotypes plus health and lifestyle information

... and making these data available to the scientific community

Improving the health of future generations

UK Biobank

To understand the interplay of genes, lifestyle and the environment in health and disease

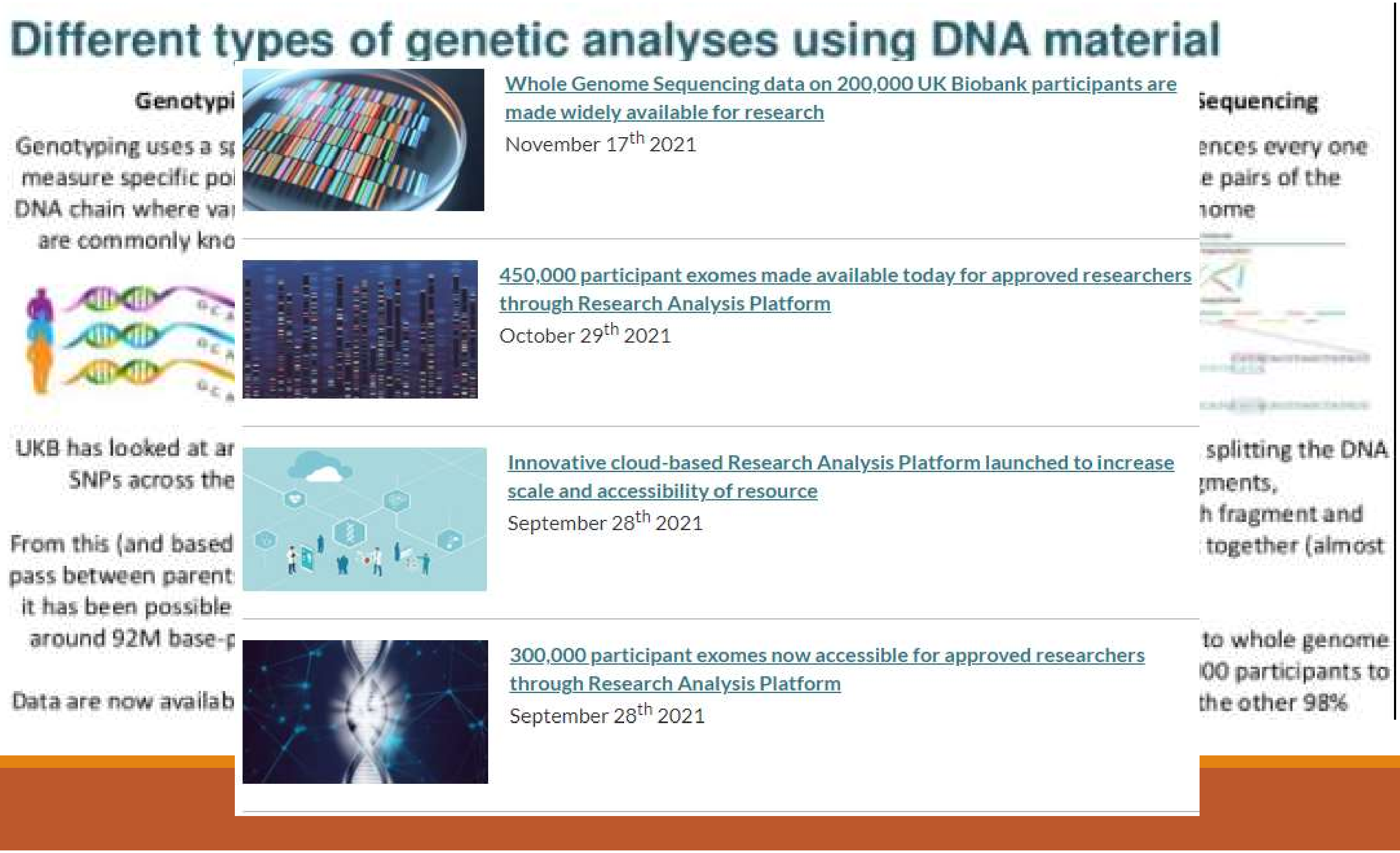
500,000 UK men and women aged 40-69 years when recruited and assessed during 2006-2010

General consent for all types of health research; no feedback of individual results to participants

Extensive baseline questions and measurements, with biological samples stored for future assays

Follow-up of health outcomes through linkage to health records and direct contact with participants

Genetic data in UK Biobank



Different types of genetic analyses using DNA material

Genotypi

Genotyping uses a s
measure specific poi
DNA chain where va
are commonly kno

UKB has looked at ar
SNPs across the

From this (and based
pass between parent
it has been possible
around 92M base-p

Data are now availab

Whole Genome Sequencing data on 200,000 UK Biobank participants are made widely available for research

November 17th 2021

450,000 participant exomes made available today for approved researchers through Research Analysis Platform

October 29th 2021

Innovative cloud-based Research Analysis Platform launched to increase scale and accessibility of resource

September 28th 2021

300,000 participant exomes now accessible for approved researchers through Research Analysis Platform

September 28th 2021

Sequencing

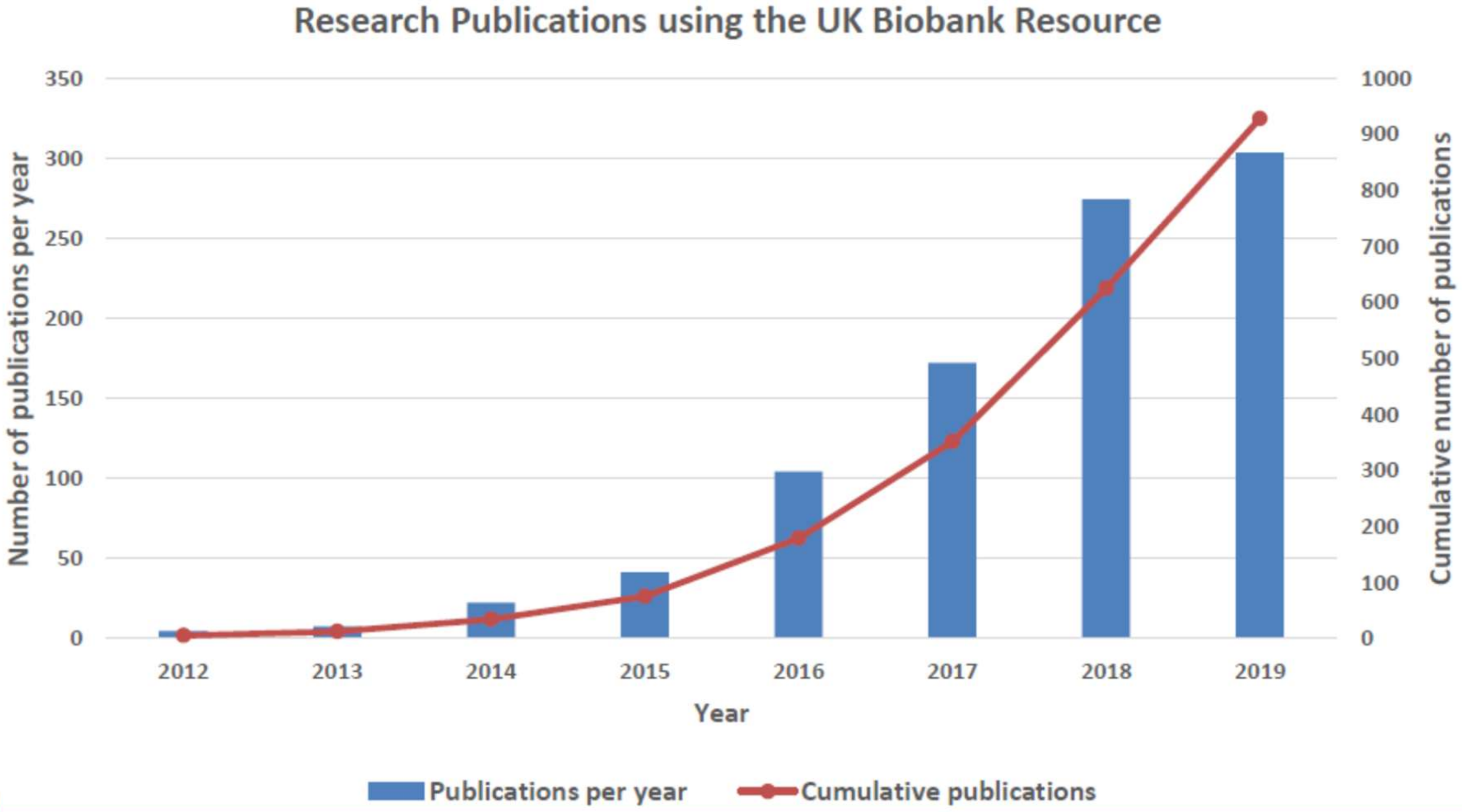
ences every one
e pairs of the
nome

splitting the DNA
gments,
h fragment and
together (almost

to whole genome
00 participants to
the other 98%

What impact is UK Biobank making

There are now >930 published research papers using the UK Biobank Resource



UK Biobank discoveries

One-off genetic test to detect heart attack risk.

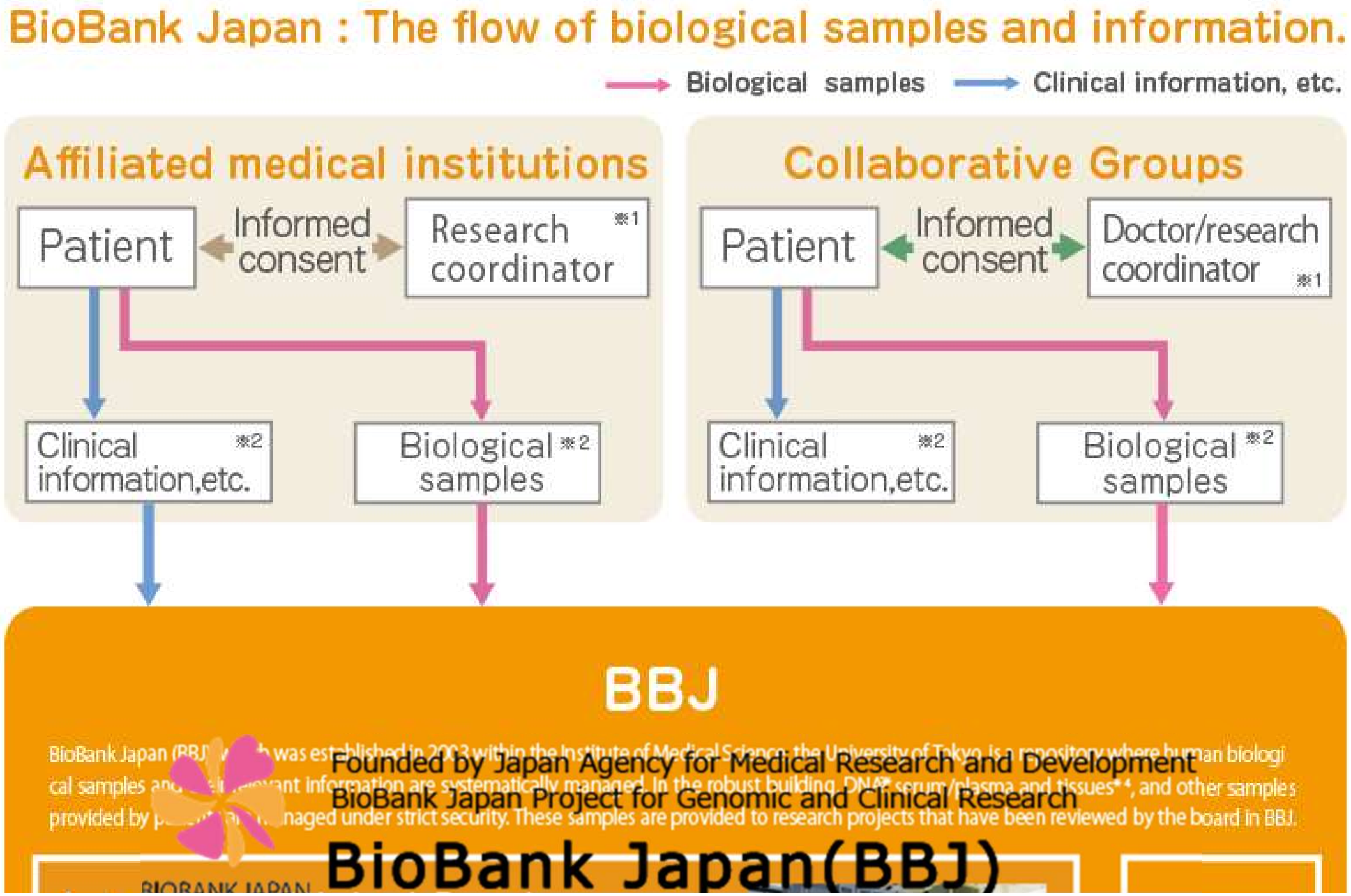
Authors devised the Genomic Risk Score (GRS) to predict risk of coronary heart disease (CHD) and explain why people with apparently no conventional risk factors, such as high cholesterol, can still go on to have a heart attack.

Participants with a GRS in the top 20% were more than four times more likely to develop coronary heart disease than those with scores in the bottom 20%.

Inouye et al., J Am Coll Cardiol . 2018 Oct 16;72(16):1883-1893

BioBank Japan (BBJ)

260,000 patients representing 440,000 cases of 51 primarily multifactorial (common) diseases



BioBank Japan : The flow of biological samples and information.

BioBank Japan (BBJ) w[illegible]h was established in 2003 within the Institute of Medical Science, the University of Tokyo, is a repository where human biological samples and [illegible] relevant information are systematically managed. In the robust building, DNA*, serum/plasma and tissues*4, and other samples provided by p[illegible] managed under strict security. These samples are provided to research projects that have been reviewed by the board in BBJ.

Founded by Japan Agency for Medical Research and Development
BioBank Japan Project for Genomic and Clinical Research

BioBank Japan(BBJ)