

Post-EWAS Platforms for Interrogating Results

Corina Lesseur

Assistant Professor

Dept. Environmental Medicine & Public Health



**Icahn School
of Medicine at
Mount
Sinai**

Outline

1. DNA methylation & EWAS

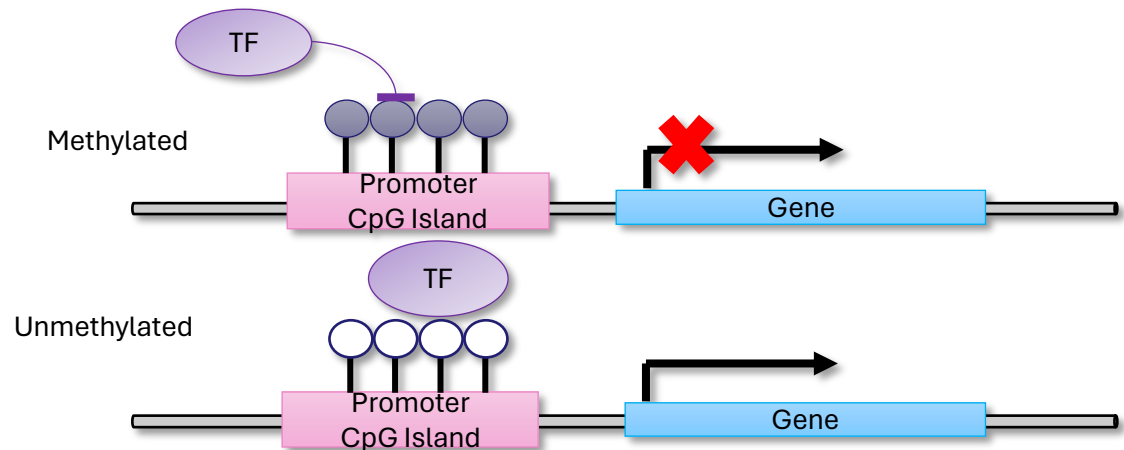
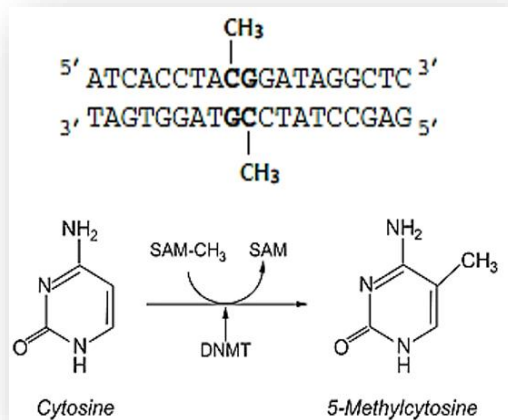
2. Post-EWAS analyses

- Overrepresentation & Gene Set Enrichment Analyses
- Other enrichment analyses
- Molecular quantitative trait loci
- Cell type proportions deconvolution & analyses
- Result annotation with genomics databases

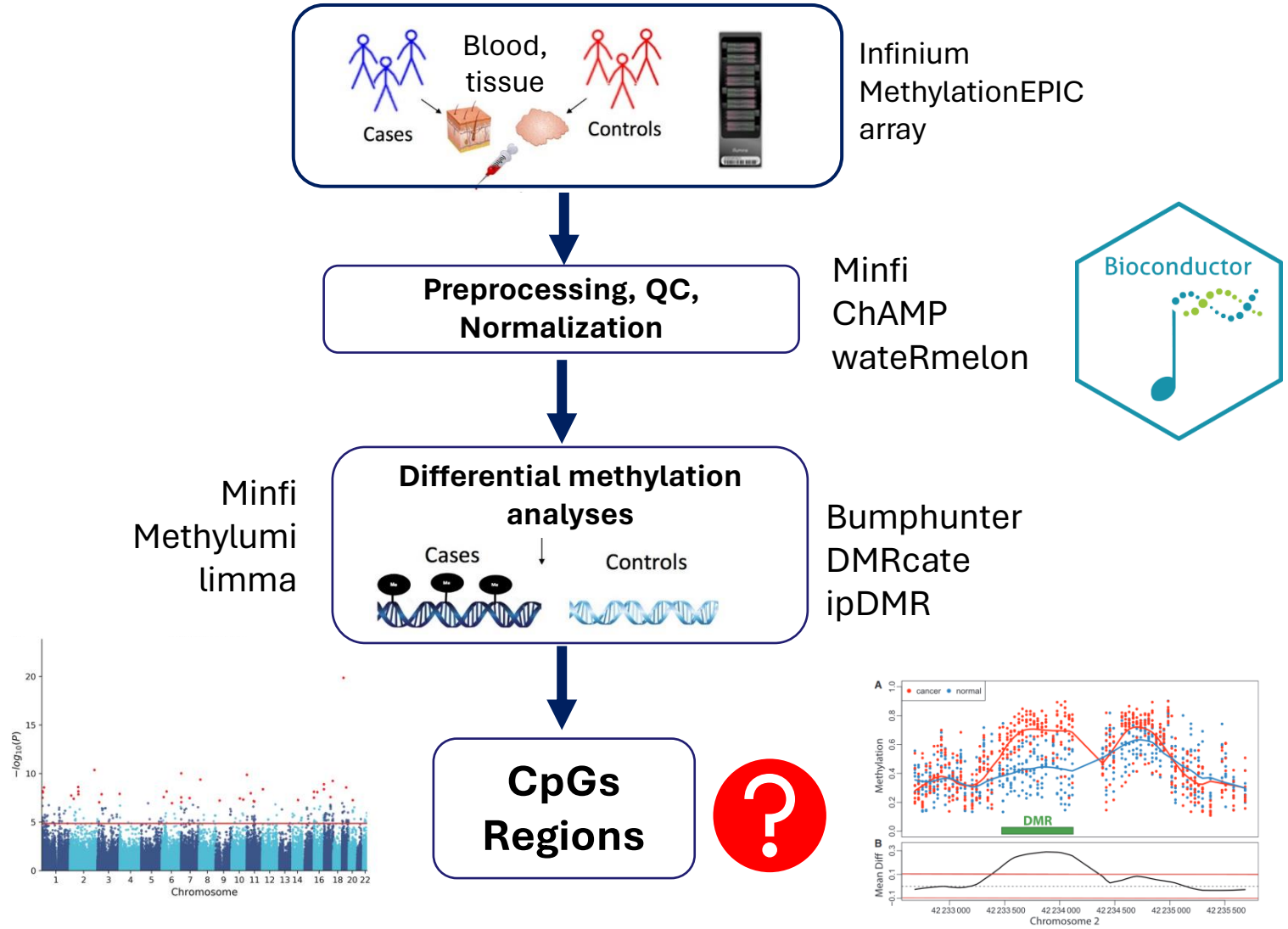


DNA methylation

- Epigenetic marks that do not involve a change sequence and regulate gene expression.
- Addition of a CH_3 to cytosines (5mC) within CpG.
- Essential for cell type differentiation & development.
- Highly *cell-type-specific*.
- Influenced by genotype & environmental exposures (i.e., smoking).
- DNAm changes with age.

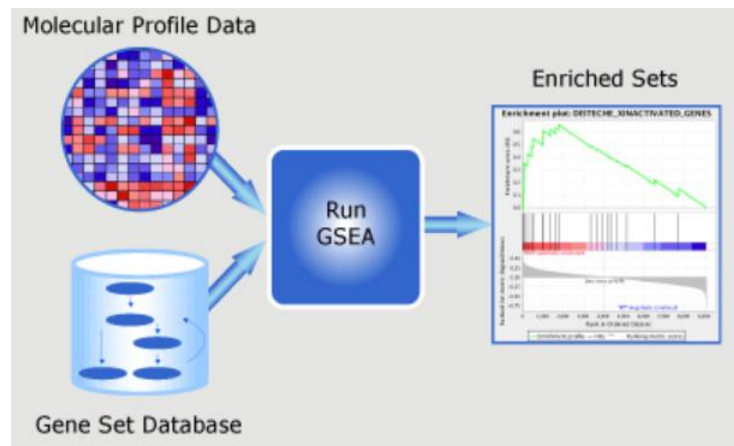


Epigenome-wide Association Studies (EWAS)



Overrepresentation & Gene Set Enrichment Analyses

- Gain a systems-level understanding of the changes in methylation/expression by examining the pathways or gene sets involved.
- Gene sets and pathways are pre-defined in exiting databases (GO, Kegg, Molecular Signatures Database (MSigDB)).
- General approaches: Over-Representation Analysis & Functional Class Scoring.



<https://www.gsea-msigdb.org/gsea/index.jsp/>



Overrepresentation & Gene Set Enrichment Analyses



[About](#) [Ontology](#) [Annotations](#) [Downloads](#) [Help](#)



Current release 2024-01-17: 42,442 GO terms | 7,655,937 annotations
1,537,348 gene products | 5,387 species ([see statistics](#))

THE GENE ONTOLOGY RESOURCE

GO Enrichment Analysis

The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.



Powered by PANTHER

Your gene IDs here...

<https://geneontology.org>

<https://bioconductor.org/packages/release/data/annotation/html/GO.db.html>

<https://www.genome.jp/kegg/>

<https://www.bioconductor.org/packages//2.12/data/annotation/html/KEGG.db.html>

The Molecular Signatures Database (MSigDB)

34,550 gene sets

9 collections

Human & mouse

Human Collections

H

hallmark gene sets are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C5

ontology gene sets consist of genes annotated by the same ontology term.

C1

positional gene sets corresponding to human chromosome cytogenetic bands.

C6

oncogenic signature gene sets defined directly from microarray gene expression data from cancer gene perturbations.

C2

curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.

C7

immunologic signature gene sets represent cell states and perturbations within the immune system.

C3

regulatory target gene sets based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C8

cell type signature gene sets curated from cluster markers identified in single-cell sequencing studies of human tissue.

C4

computational gene sets defined by mining large collections of cancer-oriented expression data.

<https://www.gsea-msigdb.org/gsea/msigdb>

<https://bioconductor.org/packages/release/data/experiment/vignettes/msigdb/inst/doc/msigdb.html>



Gene set enrichment analysis for DNA methylation data

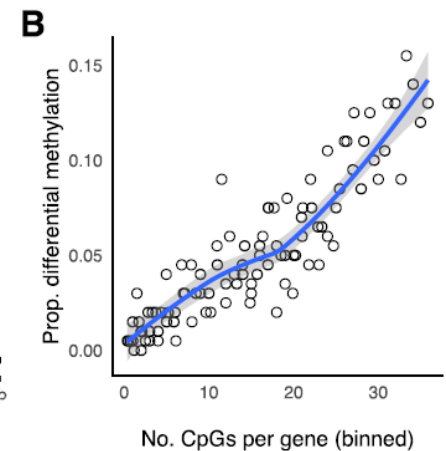
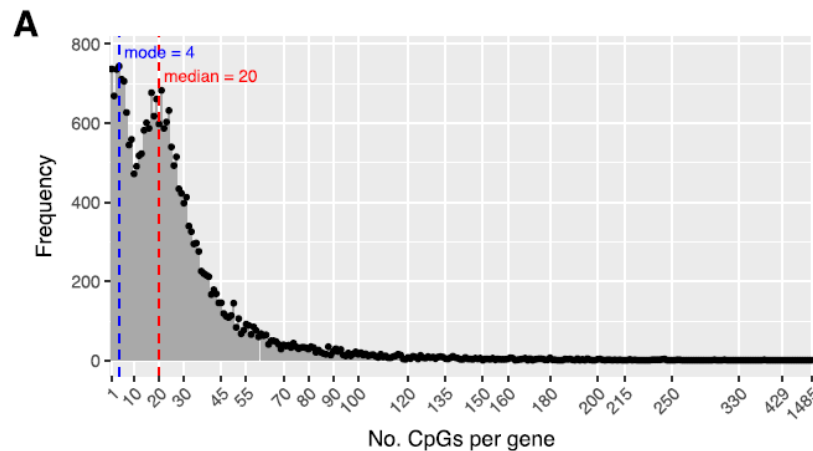
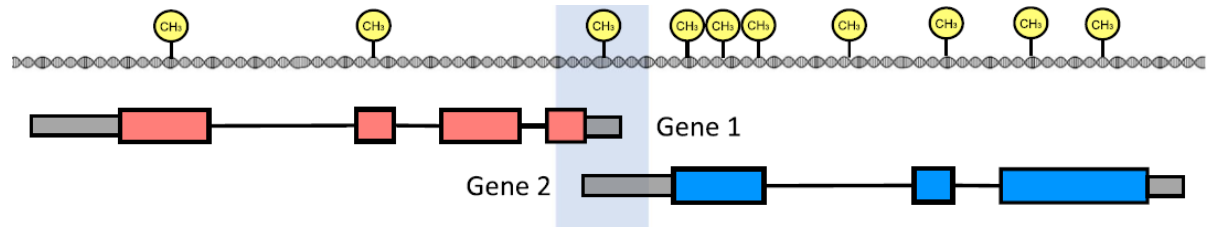
- Methylation occurs anywhere on the genome and is not as directly related to expression.
- How to assign CpGs to genes?
- Annotate CpGs to genes → differentially methylated → gene list for enrichment.

Probe number bias:

genes \neq numbers of CpGs.

Multiple annotation bias:

CpG annotated to >1 gene.



R Bioconductor
MissMethyl::gseameth
ChAMP::ebGSEA
MethylGSA

missMethyl



- CpG list enrichment based on Wallenius' noncentral hypergeometric test.
- *Accounts for probe number and multi-gene biases using weights.*
- Can subset the CpG list to specific genomic regions (i.e. promoters).
- **gometh**: List CpGs with GO.db and Kegg .db annotation packages
- **gsameth**: List CpGs with *other gene sets*.
- **goregion**: List DMRs with GO.db and Kegg .db annotation packages
- **gsaregion**: List DMRs with *other gene sets*.

```
> gometh(sigcpgs,all.cpg=NULL,collection = c("GO", "KEGG"),  
array.type="450K",plot.bias = FALSE,prior.prob = TRUE,  
anno = NULL,equiv.cpg = TRUE,fract.counts = TRUE,  
genomic.features = c("ALL", "TSS200", "TSS1500", "Body",  
"1stExon", "3'UTR", "5'UTR","ExonBnd"),sig.genes = FALSE)
```

<https://bioconductor.org/packages/release/bioc/html/missMethyl.html>

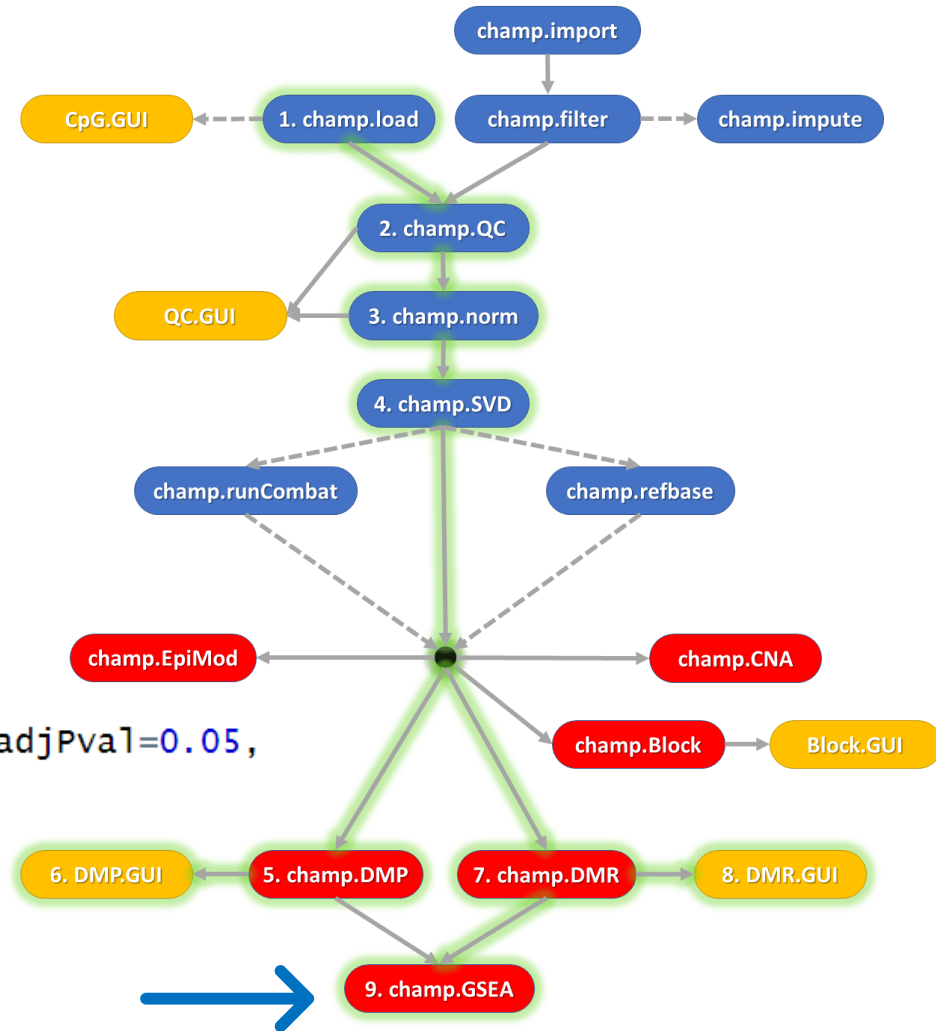
<http://oshlacklab.com/methyl-geneset-testing/index.html>

Maksimovic et al. 2021 Genome Biology (PMID: 34103055)

ebGSEA

- Adapts a global test to directly rank genes by overall differential methylation level (using all CpGs for that gene).
- **champ.GSEA**

```
> champ.GSEA(beta=myNorm, arraytype="450K", adjPval=0.05,  
method="ebayes", cores=5)
```

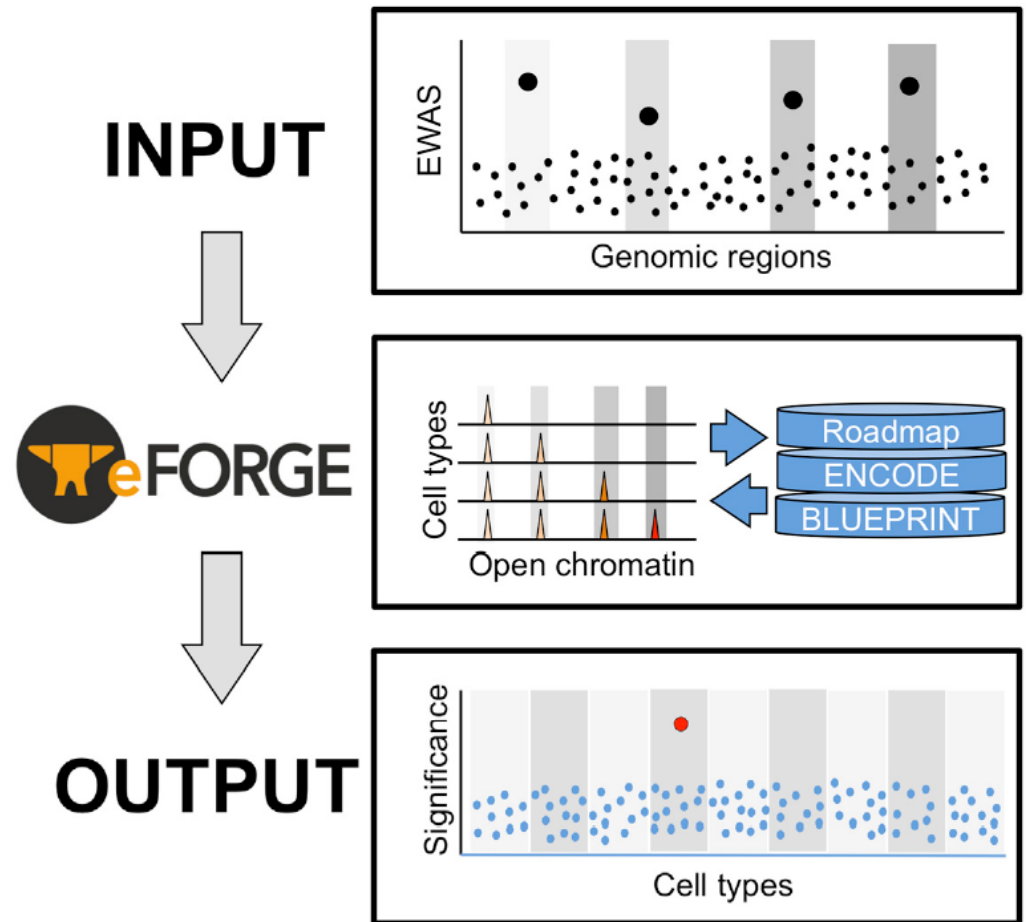


<https://github.com/aet21/ebGSEA>

<https://www.bioconductor.org/packages/release/bioc/html/ChAMP.html>

eFORGE

- Experimentally derived Functional element Overlap analysis of ReGions from EWAS.
- Standalone (Perl /python) & web-based tool.
- Detects enrichment of DNase I hypersensitive sites & chromatin states reference samples (tissues, primary cell types, & cell lines)
- Input CpG probes or bed file.



<https://eforge.altiusinstitute.org/>
<https://eforge-tf.altiusinstitute.org/>

Breeze et al 2016 Cell Rep. PMID: 27851974;
Breeze et al 2019 Bioinformatics PMID: 31161210.



Data

Paste data:

```
cg12091331
cg12962778
cg16303562
cg16501235
cg18589858
cg18712919
cg18854666
cg21792432
cg22081096
cg25059899
cg26989103
cg27443224
```

Clear box

Upload file: No file selected.

Provide file URL:

Input Options

Input file format: ▾

Name (optional):

Platform: ▾

Analysis Options

Analyse data from: ▾

Proximity: ▾

Depletion:

Background repetitions (100-1000):

Significance threshold:

Strict:

Marginal:

eFORGE

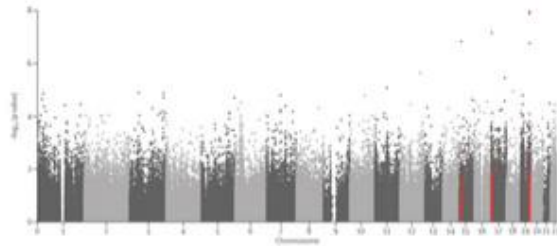
A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data

Web version max. 1K probes

<https://eforge.altiusinstitute.org/>

Breeze et al 2016 Cell Rep. PMID: 27851974;
Breeze et al 2019 Bioinformatics PMID: 31161210.

eFORGE



Published EWAS

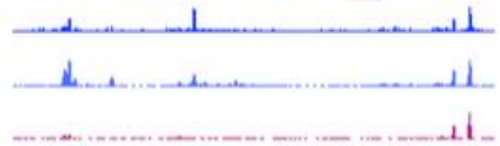
Match 1000 times



485,512 probes

Gene relationship
CpG island relationship

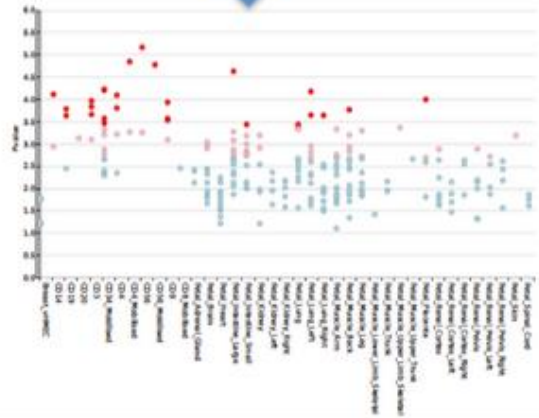
Overlap Overlap



DNase hotspots from ENCODE or Roadmap



Express enrichment
as $-\log_{10}(p \text{ value})$
of the EWAS MVPs
compared to the
450k array probes



*Manhattan plot from
BMI Dick et al. EWAS,
Lancet, 2014

GREAT: Genomic Regions Enrichment of Annotations Tool

GREAT: Genomic Regions Enrichment of Annotations Tool

GREAT predicts functions of *cis*-regulatory regions.

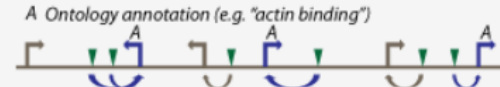
1. **Input:** A set of Genomic Regions (such as transcription factor binding events identified by ChIP-Seq).



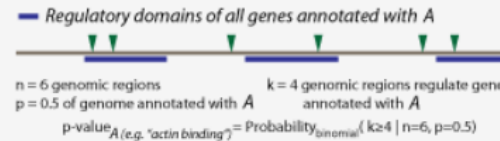
2. GREAT associates both proximal and distal input Genomic Regions with their putative target genes.



3. GREAT uses gene Annotations from numerous ontologies to associate genomic regions with annotations.



4. GREAT calculates statistical Enrichments for associations between Genomic Regions and Annotations.



5. **Output:** Annotation terms that are significantly associated with the set of input Genomic Regions.

	Ontology term	p-value
SRF peaks regulate genes involved in:	Actin cytoskeleton	10^{-9}
	FOS gene family	10^{-8}
	TRAIL signaling	10^{-7}

6. Users can create UCSC custom tracks from term-enriched subsets of Genomic Regions. Any track can be directly submitted to GREAT from the UCSC Table Browser.



Species Assembly

- Human: GRCh38 (UCSC hg38, Dec. 2013)
- Human: GRCh37 (UCSC hg19, Feb. 2009)
- Mouse: GRCm38 (UCSC mm10, Dec. 2011)
- Mouse: NCBI build 37 (UCSC mm9, Jul. 2007)

[Can I use a different species or assembly?](#)

Test regions

- BED file: No file selected.
- BED data:

[What should my test regions file contain?](#)

[How can I create a test set from a UCSC Genome Browser annotation track?](#)

Background regions

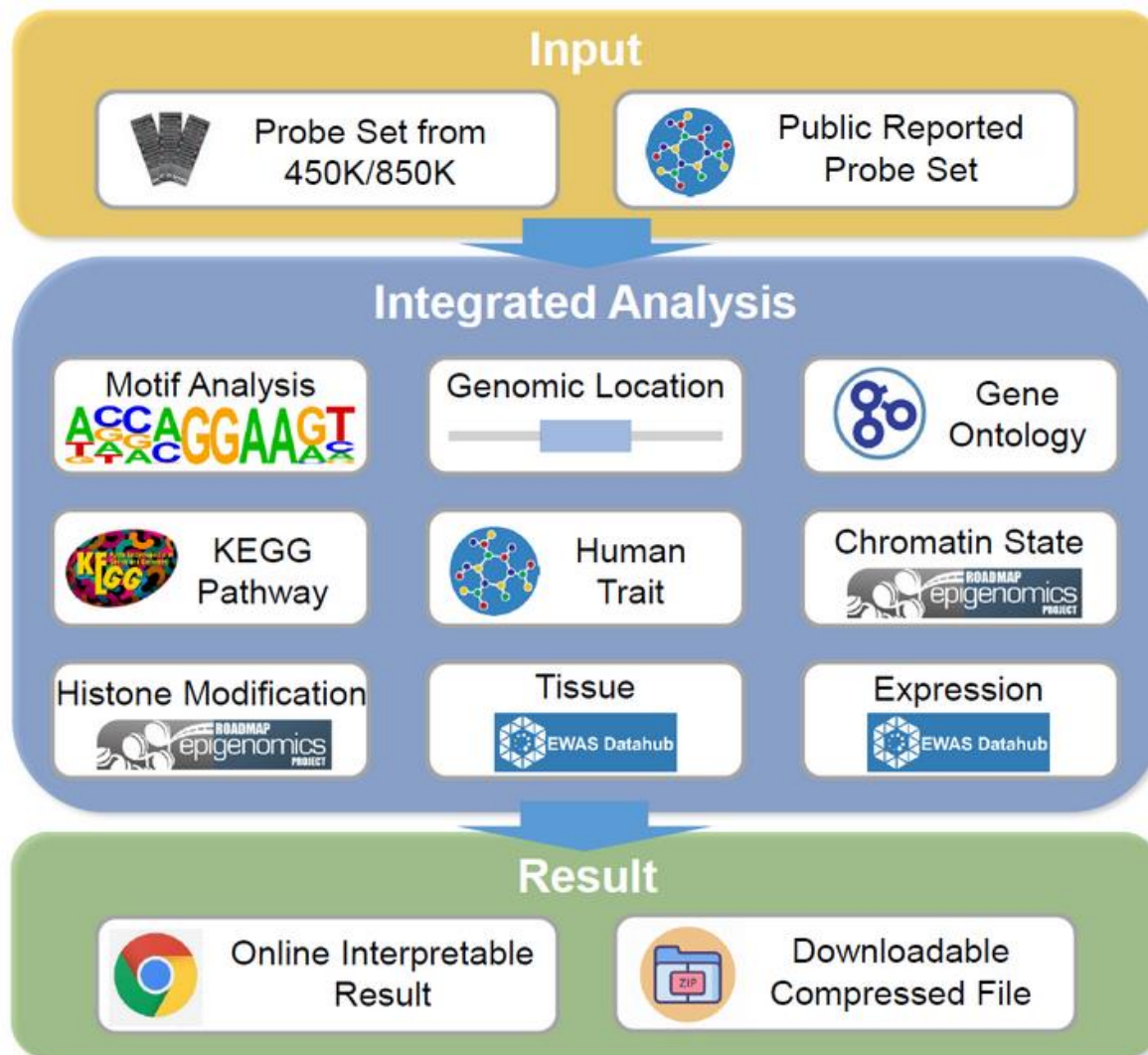
- Whole genome
- BED file: No file selected.
- BED data:

[When should I use a background set?](#)

[What should my background regions file contain?](#)

EWAS Toolkit

EWAS Open Platform



EWAS Toolkit

EWAS Open Platform

Web toolkit for epigenome-wide association studies



国家生物信息中心
China National Center for Bioinformatics

Data Resources

Computing Analysis

Data Network

EWAS Atlas
@EWAS Open Platform

Browse

EWAS Toolkit

Downloads

Statistics

API

Help

EWAS Data Hub



EWAS Toolkit @ EWAS Open Platform

a web toolkit for epigenome-wide association study

Enrichment & Annotation

Network Visualization

Correcting Batch Effects - GMQN

Input File

No file selected.

#Example File

Trait From EWAS Atlas:

Input Probe ID:

Clear Input

cg23201812
cg18014789
cg11532433
cg02091781
cg11252953
cg01323777
.....

Input Job ID:

#example: vitamin B12 supplementation related DMP (PMID:29135286)

Background: 450K EPIC/850K Others

EWAS Toolkit

EWAS Open Platform

EWAS Atlas
@EWAS Open Platform

Browse

EWAS Toolkit

Downloads

Statistics

API

Help

EWAS Data Hub

Trait ✓

Genomic Location ✓

Gene Ontology ✓

KEGG Pathway ✓

Chromatin State ✓

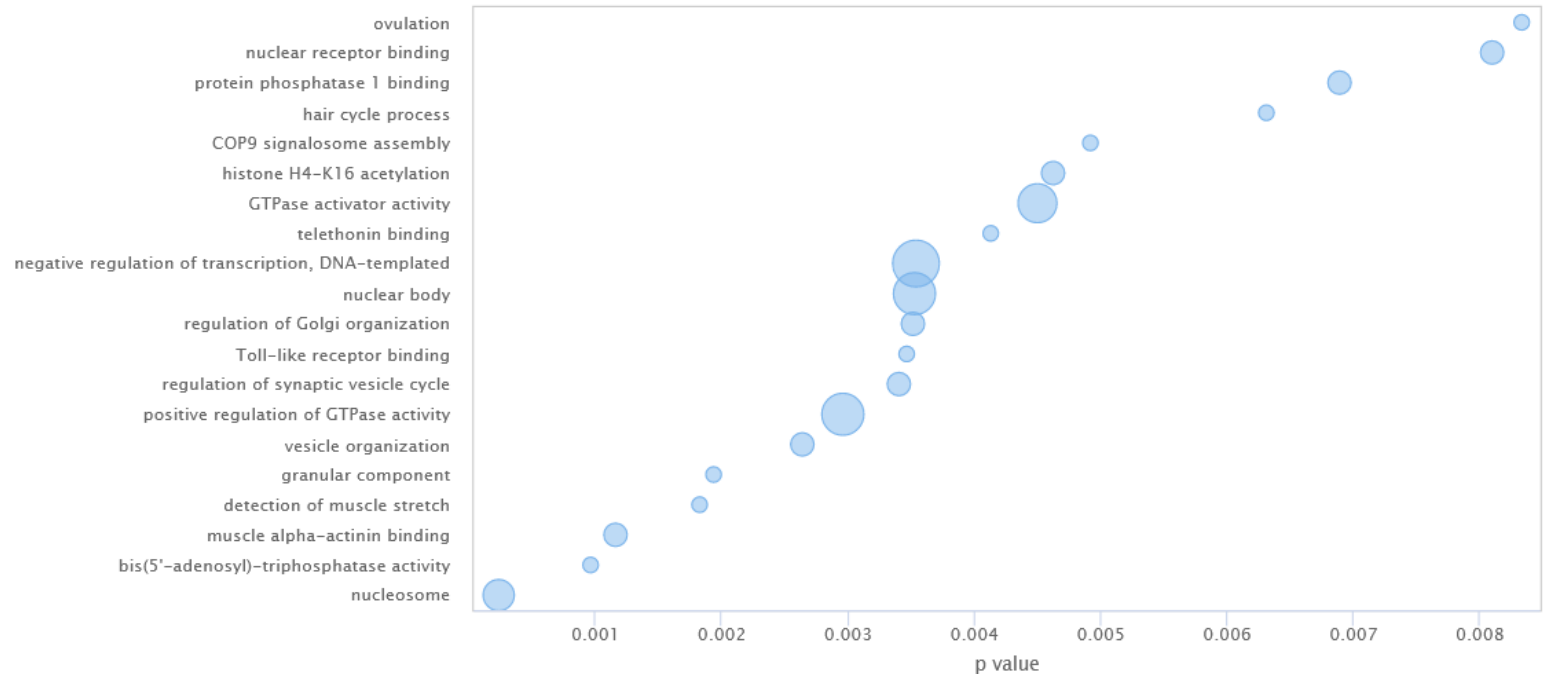
Histone Modification ✓

Tissue ✓

Expression Regulation ✓

Motif ✓

Gene Ontology Enrichment



Xiong et al 2022. Nucleic Acids Res. PMID: 34718752

<https://ngdc.cncb.ac.cn/ewas/documentation#/toolkit>

Outline

1. DNA methylation & EWAS

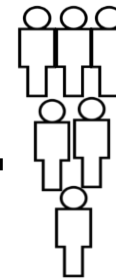
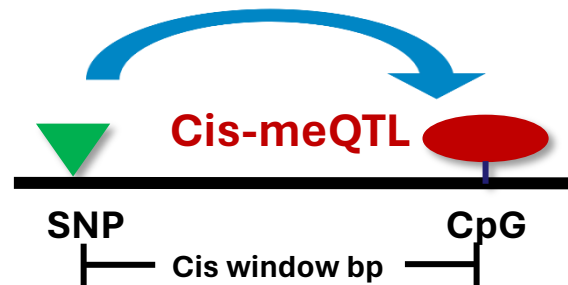
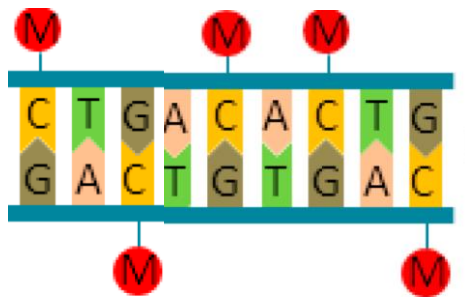
2. Post-EWAS analyses

- Overrepresentation & Gene Set Enrichment Analyses
- Other enrichment analyses
- Molecular quantitative trait loci
- Cell type proportions deconvolution & analyses
- Result annotation with other genomics databases



Molecular Quantitative Trait Loci (molQTL)

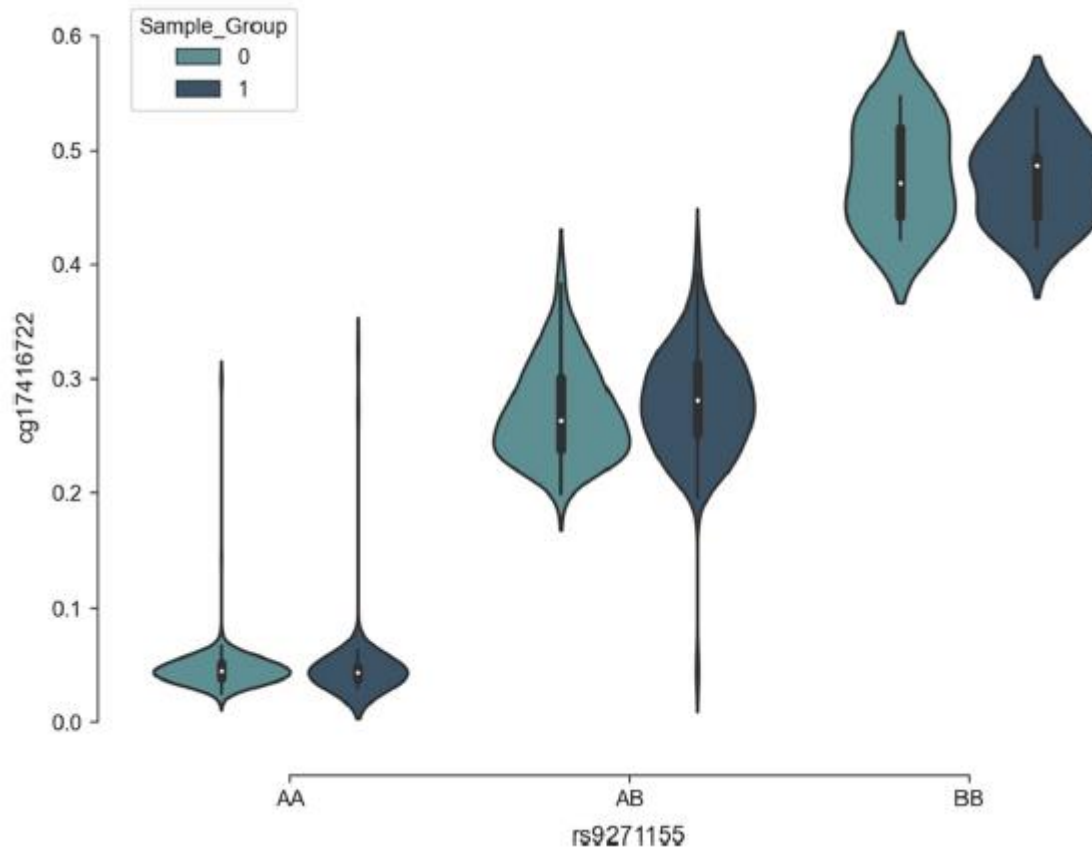
- Methylation quantitative trait loci (meQTL)
- $\approx 45\%$ of CpGs in 450K array



SNP	%Meth.
AA	30
AG	35
GG	45

Molecular Quantitative Traits	
SNP \rightarrow methylation	mQTL or meQTLs
SNP \rightarrow gene expression	eQTLs
5mC \rightarrow expression	eQTM

Methylation Quantitative Trait Loci



Genetics of DNA Methylation Consortium (GoDMC)

- A multi-cohort meta-analysis of **blood meQTLs**
- >32,851 participants
- Database <http://mqtlldb.godmc.org.uk/>
- API <http://api.godmc.org.uk/v0.1>

GoDMC

Home Search Cohorts Resources API About

Cis and *trans* meta-analysis results from genome-wide scans of 420,509 DNA methylation sites

Search the GoDMC database

Example searches:

- rs7105015
- snp:6:16000000-17000000
- cg24851651
- cpg:6:16000000-17000000
- cg19104072,cg16950941

<http://www.godmc.org.uk/>

Min et al 2021 Nat Genet. PMID: 34493871

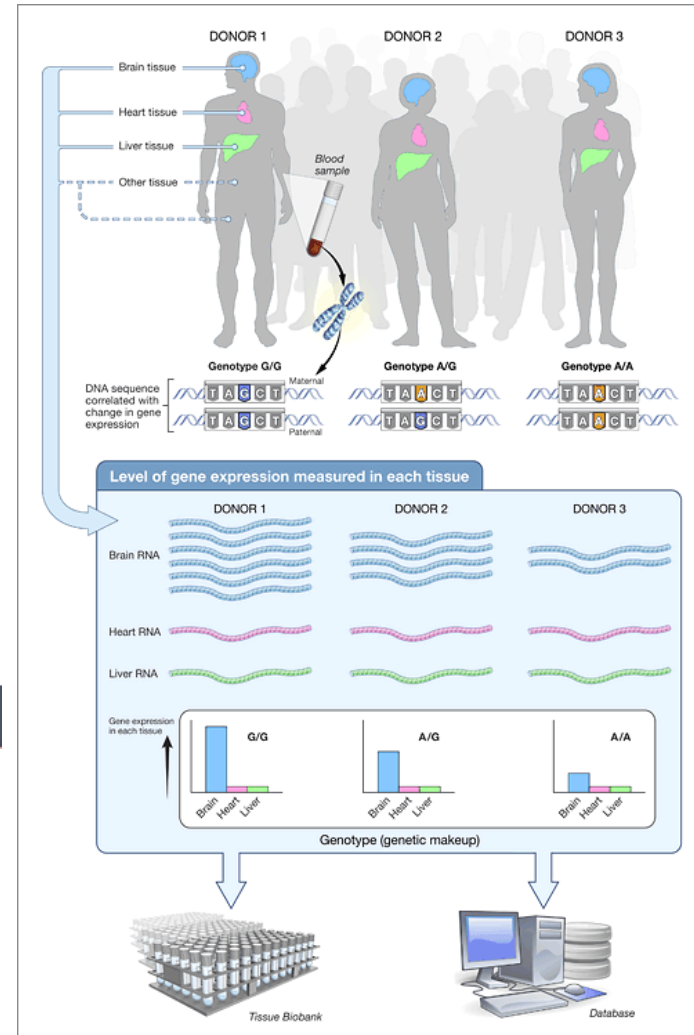
The Genotype-Tissue Expression (GTEx) Project

- Ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation.
- Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals.

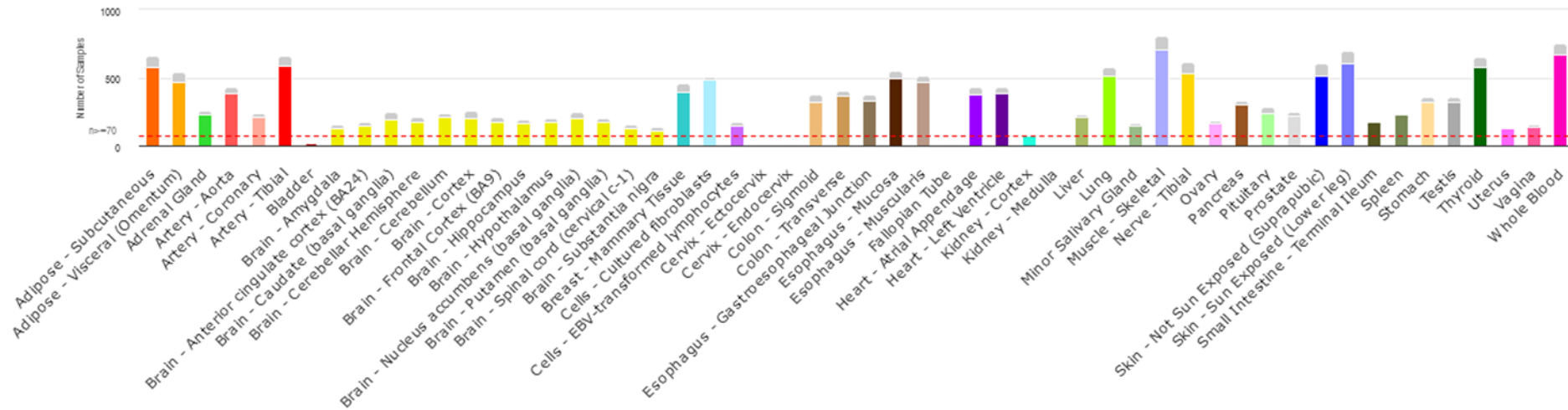


Home Downloads Expression Single Cell QTL IGV Browser Tissues & Histology

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



The Genotype-Tissue Expression (GTEx) Project

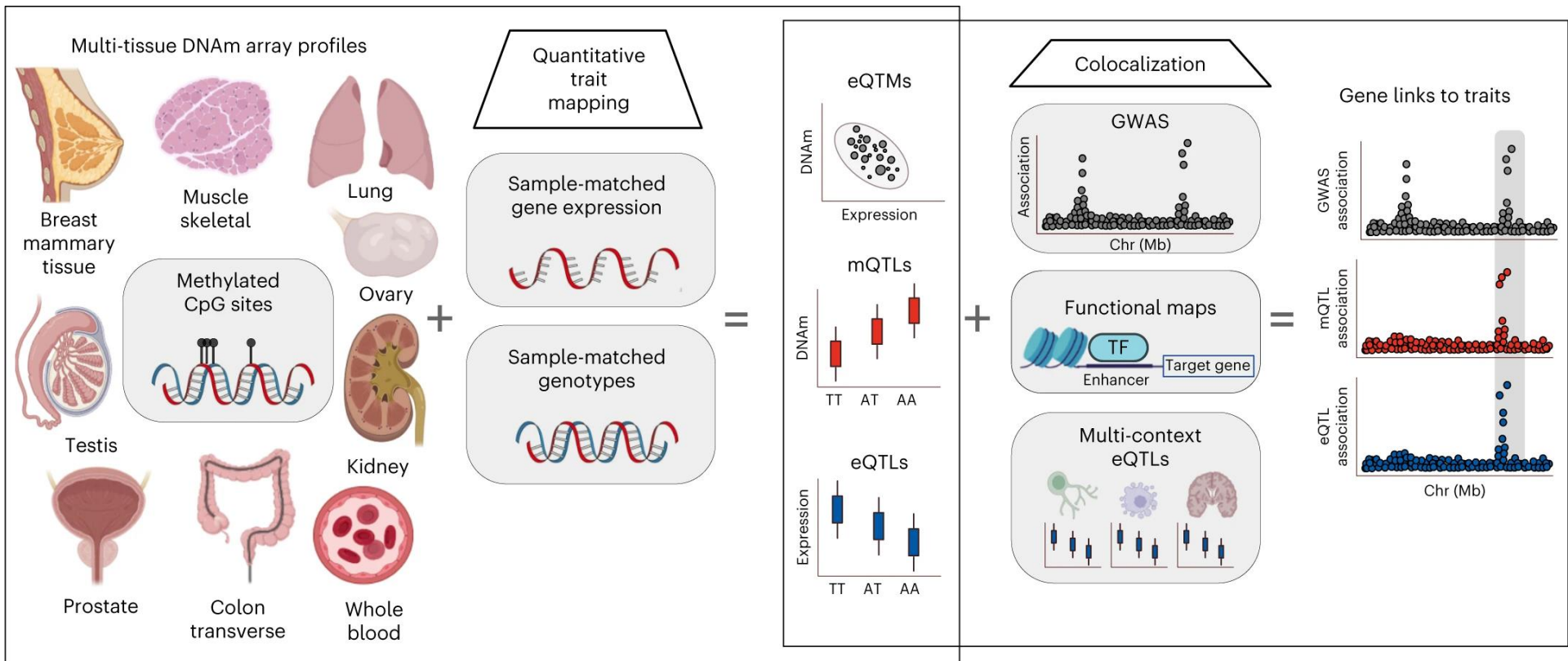


<https://gtexportal.org/home/>

<https://gtexportal.org/home/downloads/adult-gtex/ctl>



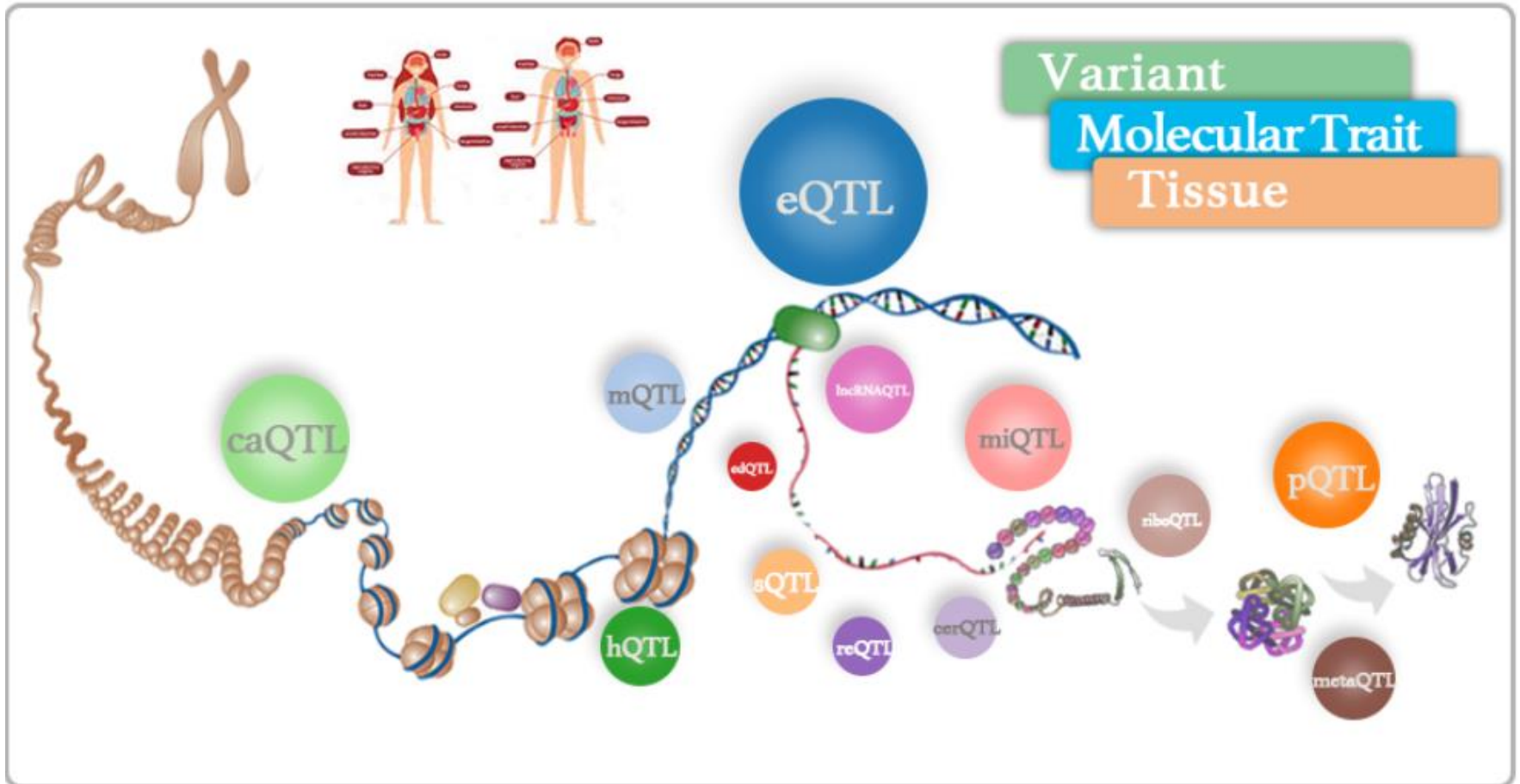
DNA methylation QTL mapping across diverse human tissues



<https://gtexportal.org/home/downloads/egtex/methylation>

Oliva et al. 2022 Nat Gen. PMID: 36510025

QTLbase



Outline

1. DNA methylation & EWAS

2. Post-EWAS analyses

- Overrepresentation & Gene Set Enrichment Analyses
- Other enrichment analyses
- Molecular quantitative trait loci
- Cell type proportions deconvolution & analyses
- Result annotation with other genomics databases



Cellular Heterogeneity in EWAS

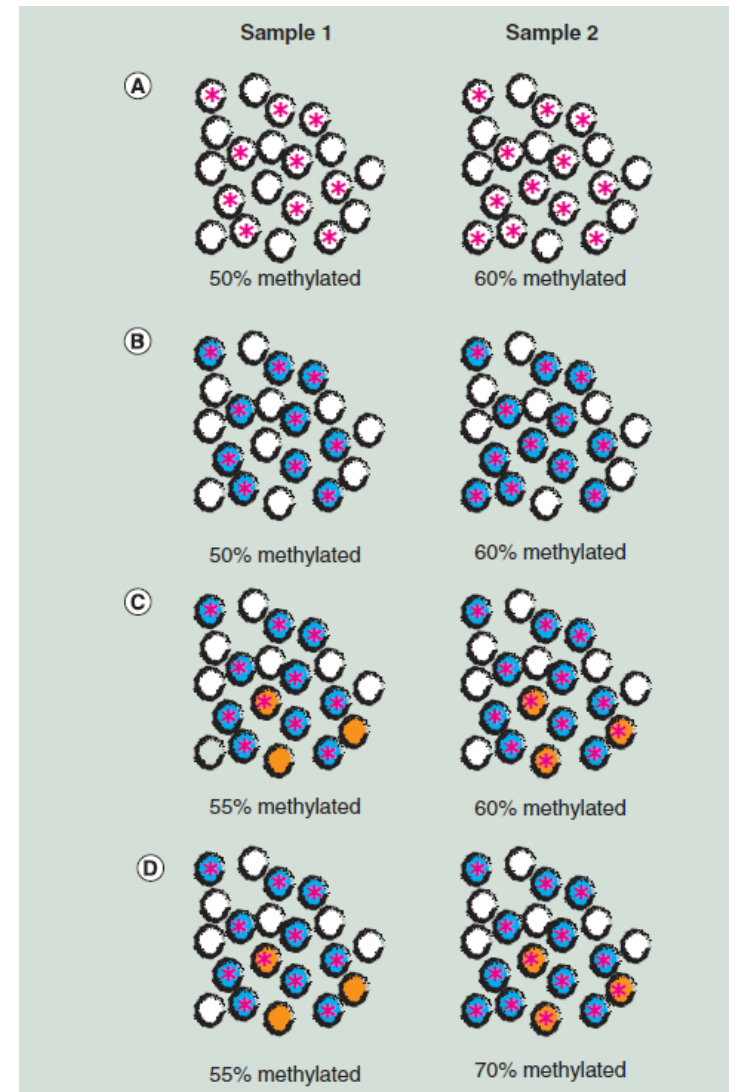
- Most biological samples are mixtures of cells with distinct methylation patterns.
- EWAS aim to identify DMPs, cell type proportions may also vary between cases and controls.
- Cell-type deconvolution methods:
 - Reference-free
 - Reference-based

R packages: EpiDISH, RefFreeEWAS

Houseman et al. 2012 Bioinformatics 22568884

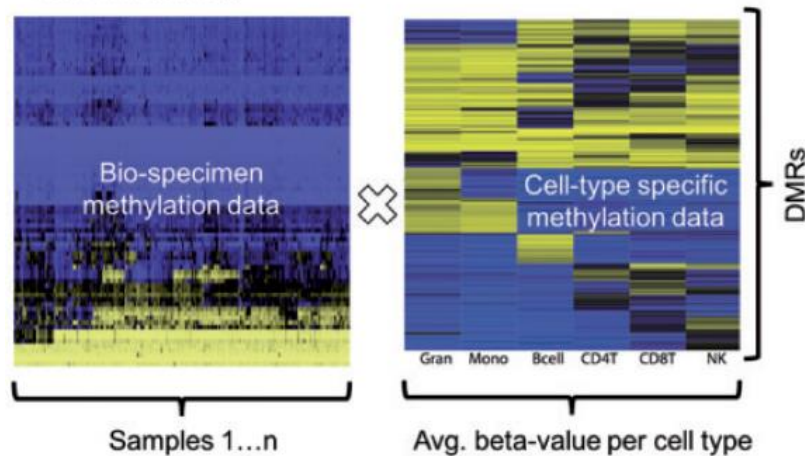
Houseman et al. 2019 Bioinformatics PMID: 27358049

Zheng et al 2018. Nat Methods. PMID: 30504870.



Cell proportion estimates in analyses

1. Reference-based cell type deconvolution using immune cell DMRs (e.g. Houseman method)



Result: a matrix of samples with estimated immune cell type proportions

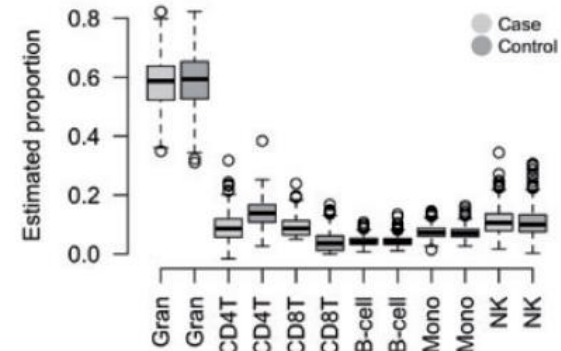
	Gran	Mono	B-cell	CD4T	CD8T	NK
Samples 1...n	Gran ₁	Mono ₁	B-cell ₁	CD4T ₁	CD8T ₁	NK ₁
	Gran ₂	Mono ₂	B-cell ₂	CD4T ₂	CD8T ₂	NK ₂

	Gran _n	Mono _n	B-cell _n	CD4T _n	CD8T _n	NK _n
	Immune proportion estimates for samples 1...n					

2. Use cell type proportion estimates in analyses of outcome or phenotype of interest.

Examples:

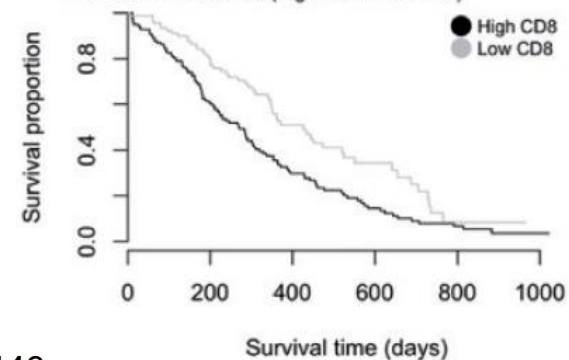
i. Compare cell type proportions in cases with controls



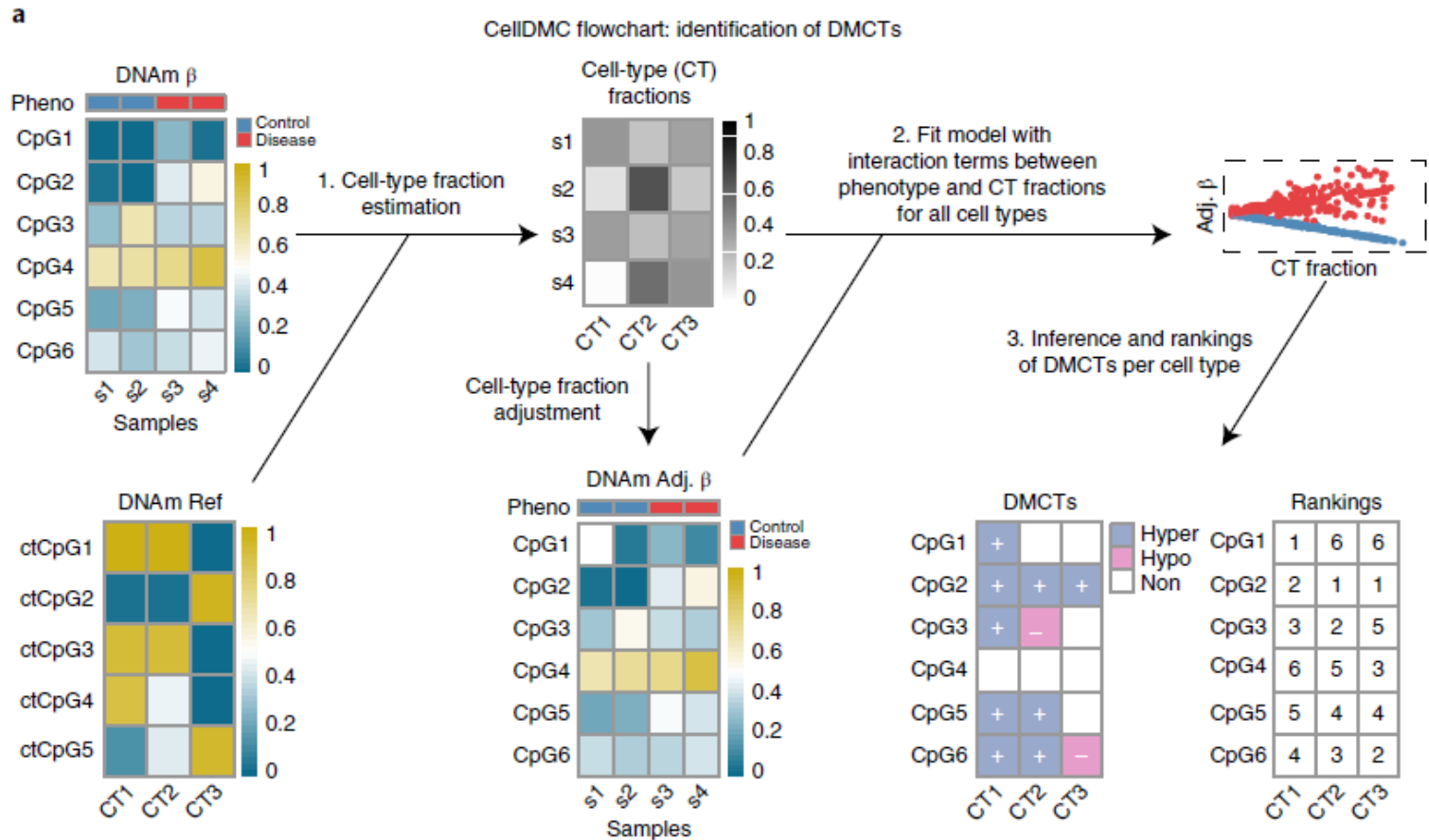
ii. Calculate cell type ratio (e.g. NLR), and test association of NLR with variables of interest (e.g. survival)

$$\text{NLR} = \frac{\text{Proportion of Granulocytes}}{\text{Proportion of Lymphocytes (B-cells + CD4T + CD8T + NK)}}$$

iii. Test relation of cell type proportion (e.g. CD8) or ratio with survival outcome (e.g. overall survival)



Identification of differentially methylated cell types in EWAS



EpiDISH::CellDMC

Outline

1. DNA methylation & EWAS

2. Post-EWAS analyses

- Overrepresentation & Gene Set Enrichment Analyses
- Other enrichment analyses
- Molecular quantitative trait loci
- Cell type proportions deconvolution & analyses
- **Result annotation with other genomics databases**



EWAS Catalog

EWAS Catalog

Search

About

Documentation

Download

Upload

University of
BRISTOL

MRC Integrative
Epidemiology
Unit



EWAS Catalog β

The MRC-IEU catalog of epigenome-wide association studies

Search



Examples: [cg00029284](#), [chr12:111731203](#), [FTO](#), [6:15000000-25000000](#), [body mass index](#), [27040690](#).

[+Advanced](#)

Data last updated: 2023-12-22

<https://www.ewascatalog.org/>



EWAS Atlas

EWAS Open Platform

国家生物信息中心
China National Center for Bioinformation

Data Resources Computing Analysis Data Network Standards

EWAS Atlas @EWAS Open Platform

Browse EWAS Toolkit Downloads Statistics API Help EWAS Data Hub

EWAS Atlas @EWAS Open Platform

A knowledgebase of epigenome-wide association studies

Q

Examples: smoking, AHRR, cg05575921

Associations 675,142	Traits 797	Cohorts 3,613
Tissues/Cells 218	Studies 1,705	Publications 1078

Last update: new EWAS on metabolic syndrome (MetS) has been added online on January 5, 2024
New Database: EWAS Data Hub (A data hub of DNA methylation array data and metadata)
New Toolkit: EWAS Toolkit (A web toolkit for epigenome-wide association study)

Number of Publications

Year	Count
2010	1
2011	3
2012	5
2013	15
2014	29
2015	74
2016	115
2017	184
2018	305
2019	491
2020	677
2021	881
2022	973
2023	1,043
2024	1,078

Follow us: @EWAS_Open_Platform
Cite: EWAS Open Platform: integrated data, knowledge and toolkit for epigenome-wide association study. *Nucleic Acids Res.* 2021 [PMID=34718752]
EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res.* 2019 [PMID=30364969]

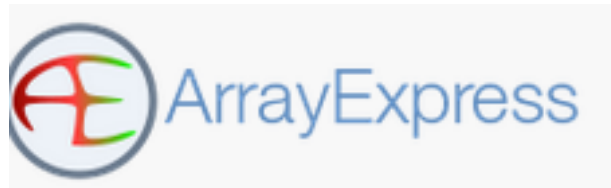
Public genomics data repositories



<https://www.ncbi.nlm.nih.gov/geo/>



<https://bioconductor.org/packages/release/bioc/html/GEOquery.html>



<https://www.ebi.ac.uk/biostudies/arrayexpress>



<https://ngdc.cncb.ac.cn/ewas/datahub/index>



Questions

