Functional Genomics

Nov 29, 2017

Sarah Gagliano
Sarah.Gagliano@umich.edu

Biostat666: Statistical methods in human genetics

Goals for this lecture

- Provide an overview of functional genomics
- Understand its importance/uses in the context of GWAS

Outline

- Functional Genomics 101
- Using functional genomics to prioritize risk variants

Outline

- Functional Genomics 101
- Using functional genomics to prioritize risk variants

DNA → mRNA → amino acids chain (protein)

Transcription Translation

phenotype

The "Central Dogma" of Biology

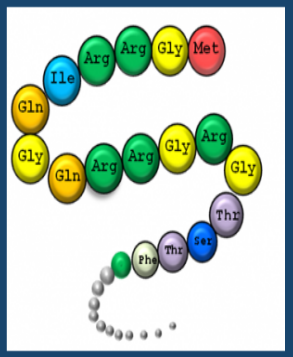
Coding variation

Genetic variation → amino acid → protein

phenotype

Coding variation

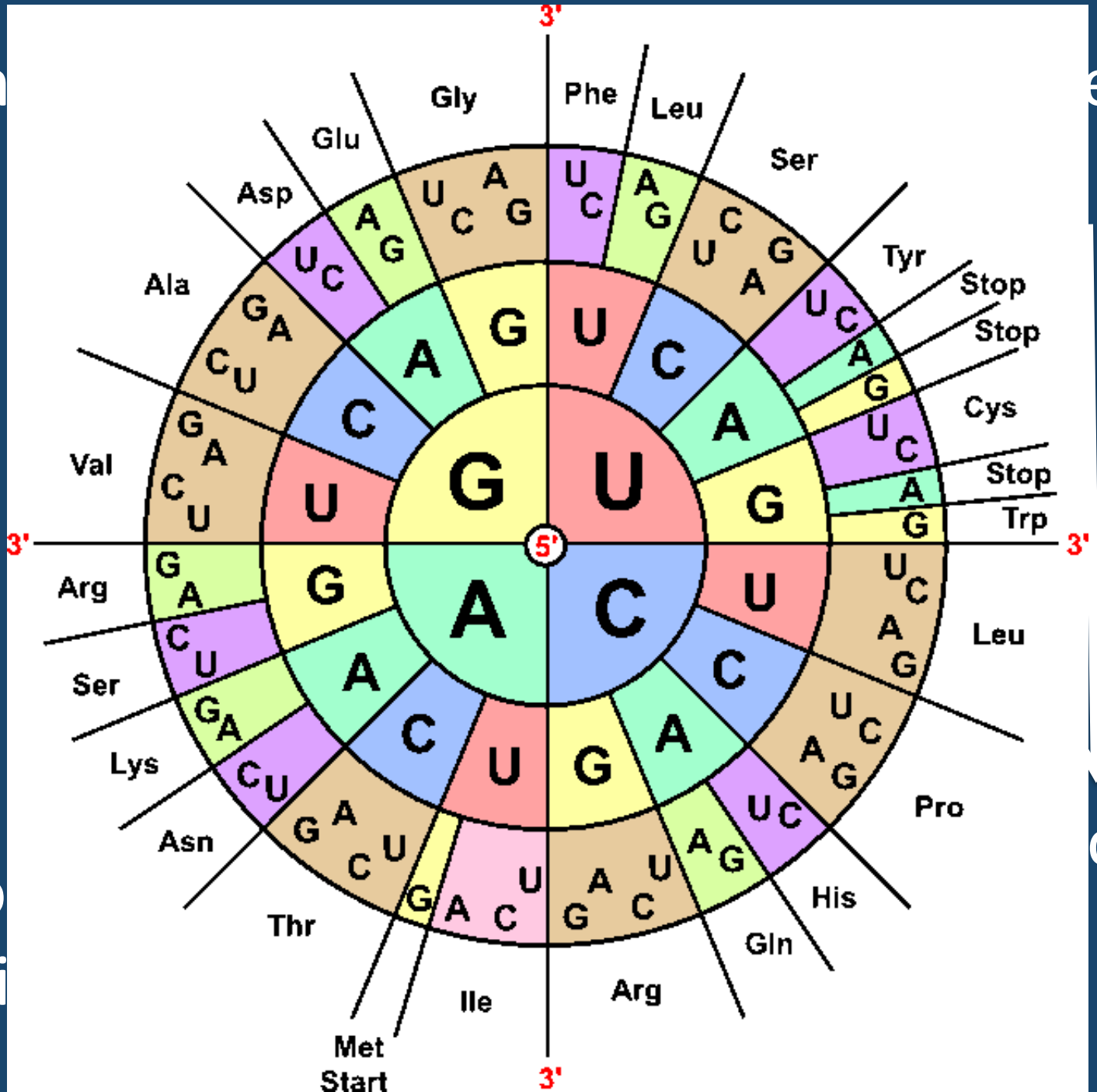
Genetic variation → amino acid → protein



phenotype

Coding variation

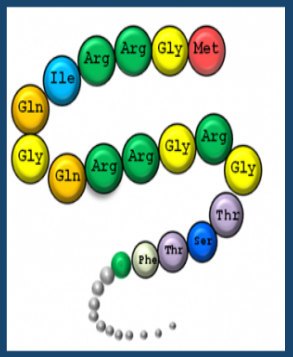
Gen ein



Co otype

vari

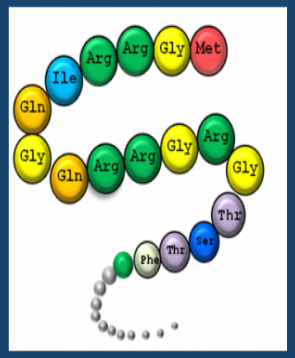
Genetic variation → amino acid → protein



phenotype

Coding variation

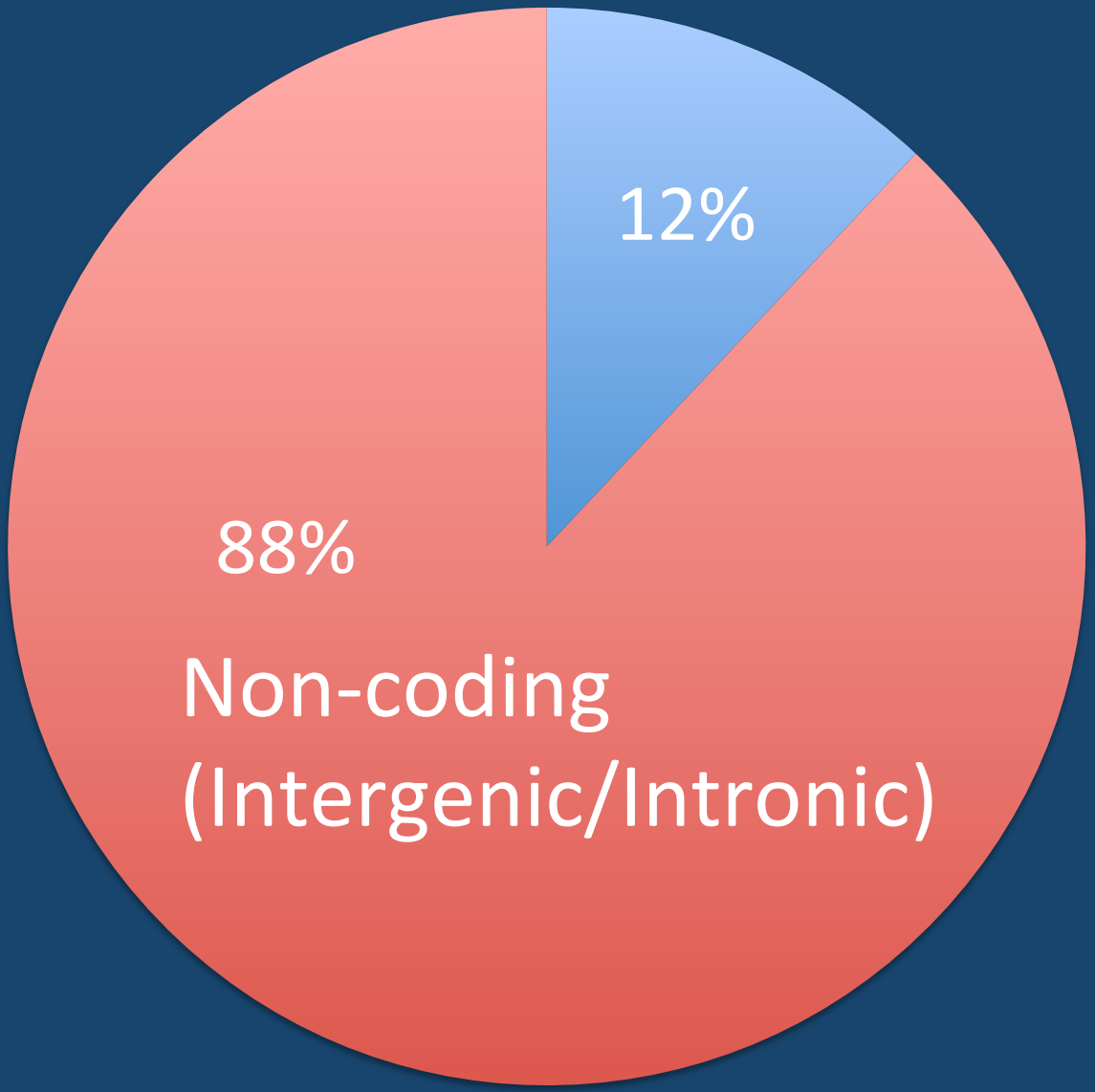
Genetic variation → amino acid → protein



phenotype

Coding variation

Most GWAS hits are non-coding



Hindorff et al. 2009

Scientists Find 'Junk DNA' Useful After All

Come on! Help me find it, I think this one controls hair loss!

ENCODE Project creates a map of functional regions

Non-coding regions are not “junk”!

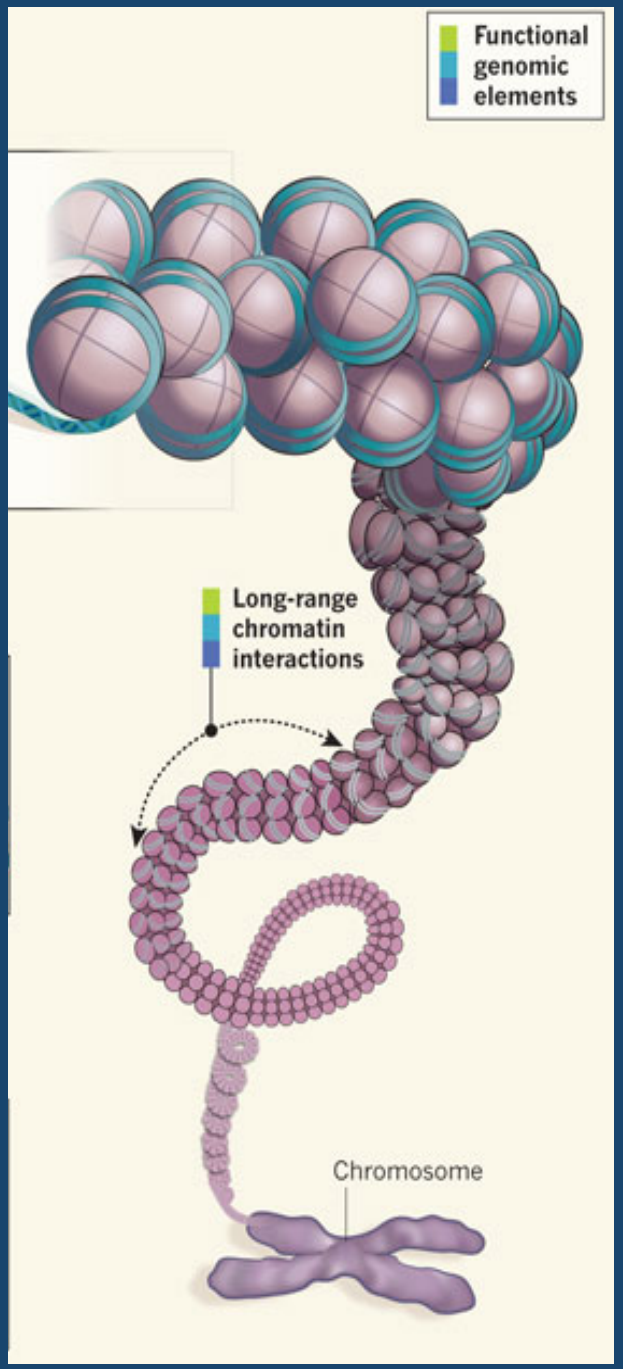


Figure from Ecker et al. (2012) *Nature*, 489: 52-55

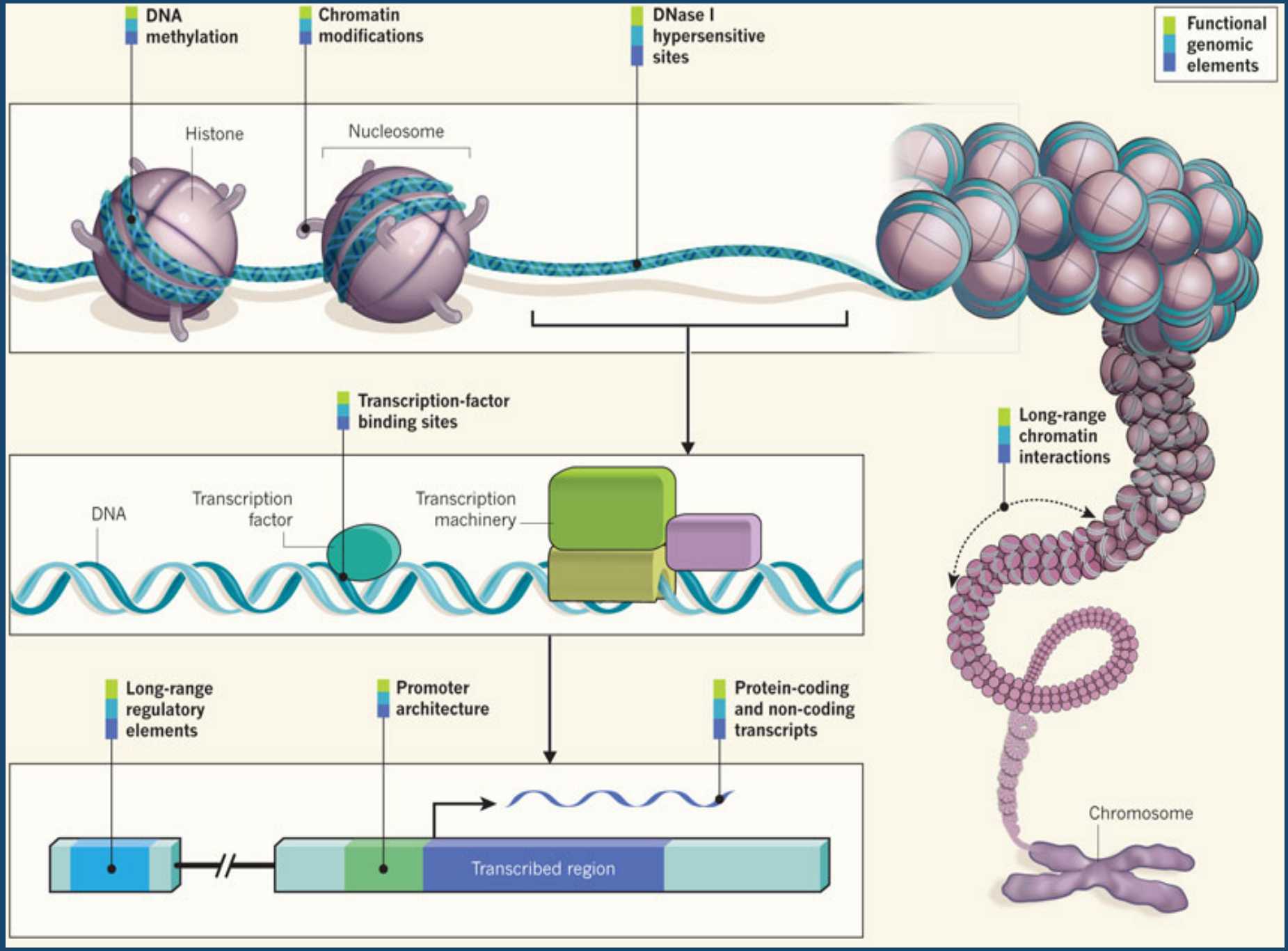


Figure from Ecker et al. (2012) *Nature*, 489: 52-55

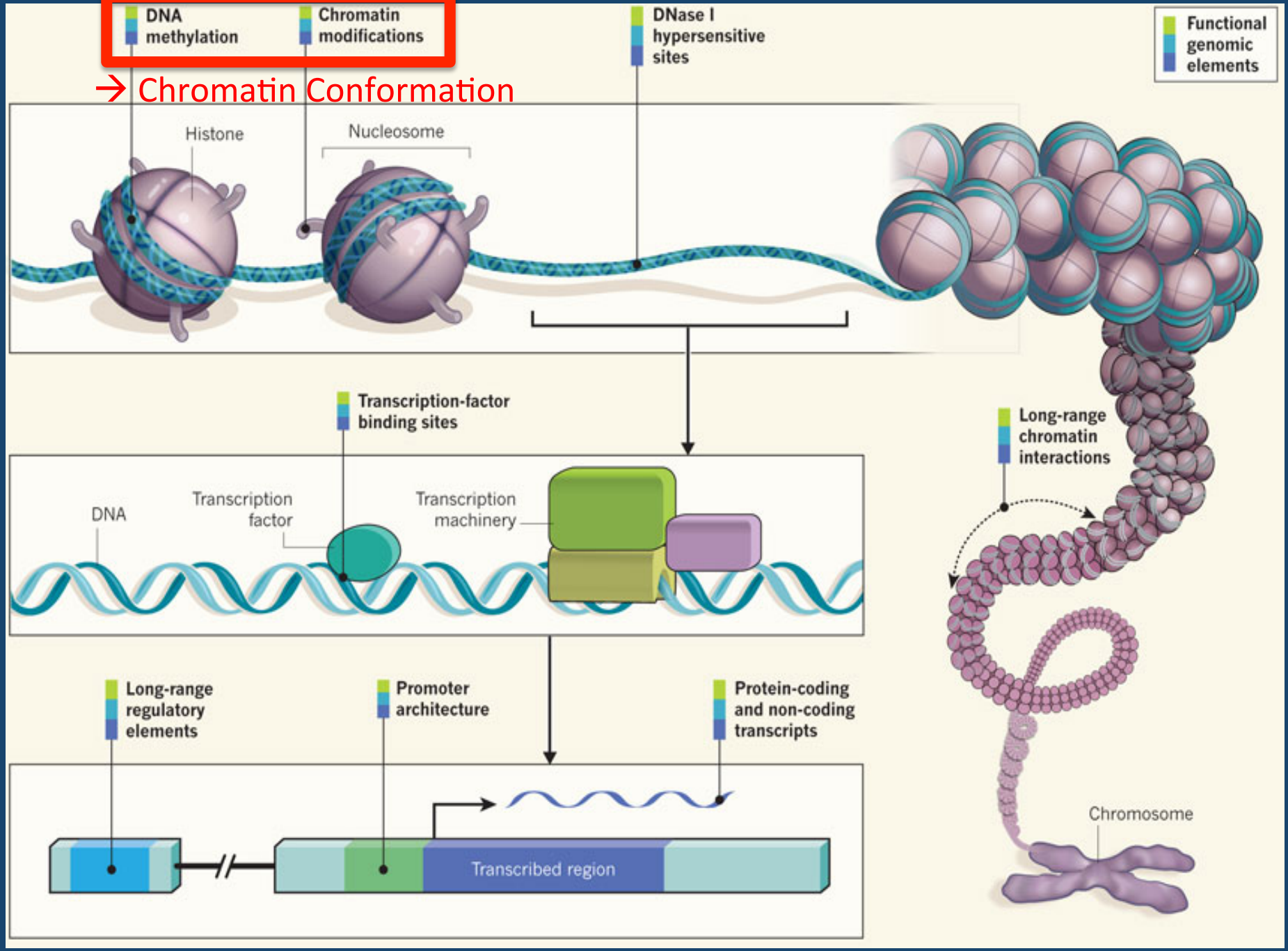
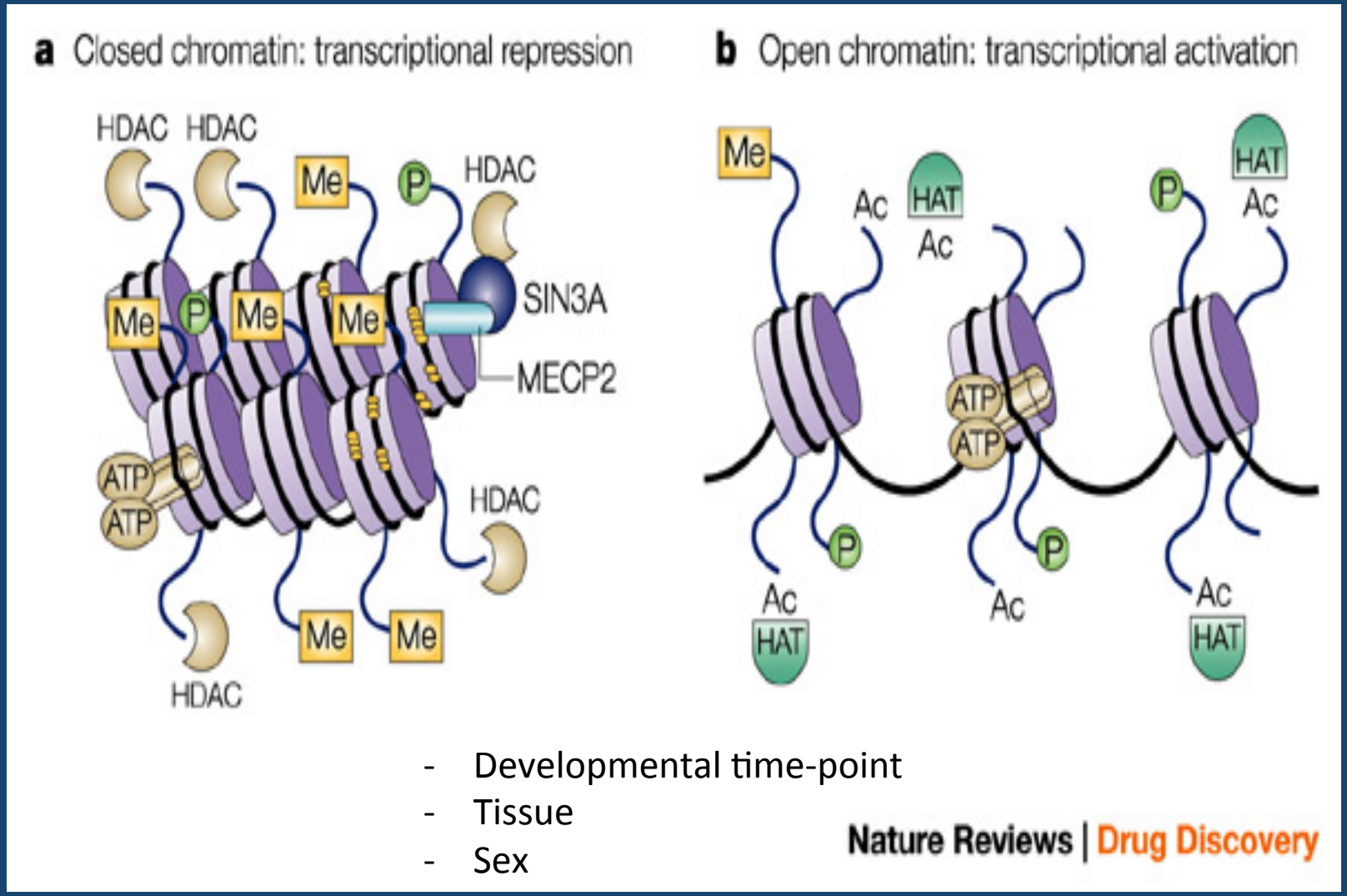


Figure from Ecker et al. (2012) *Nature*, 489: 52-55

Marks associated with closed vs. open chromatin



- Developmental time-point
- Tissue
- Sex

Histone Modifications

Histone marks define functional regions

Mark of an active enhancer

UCSC Genome Browser

Enhancer-Promoter Looping

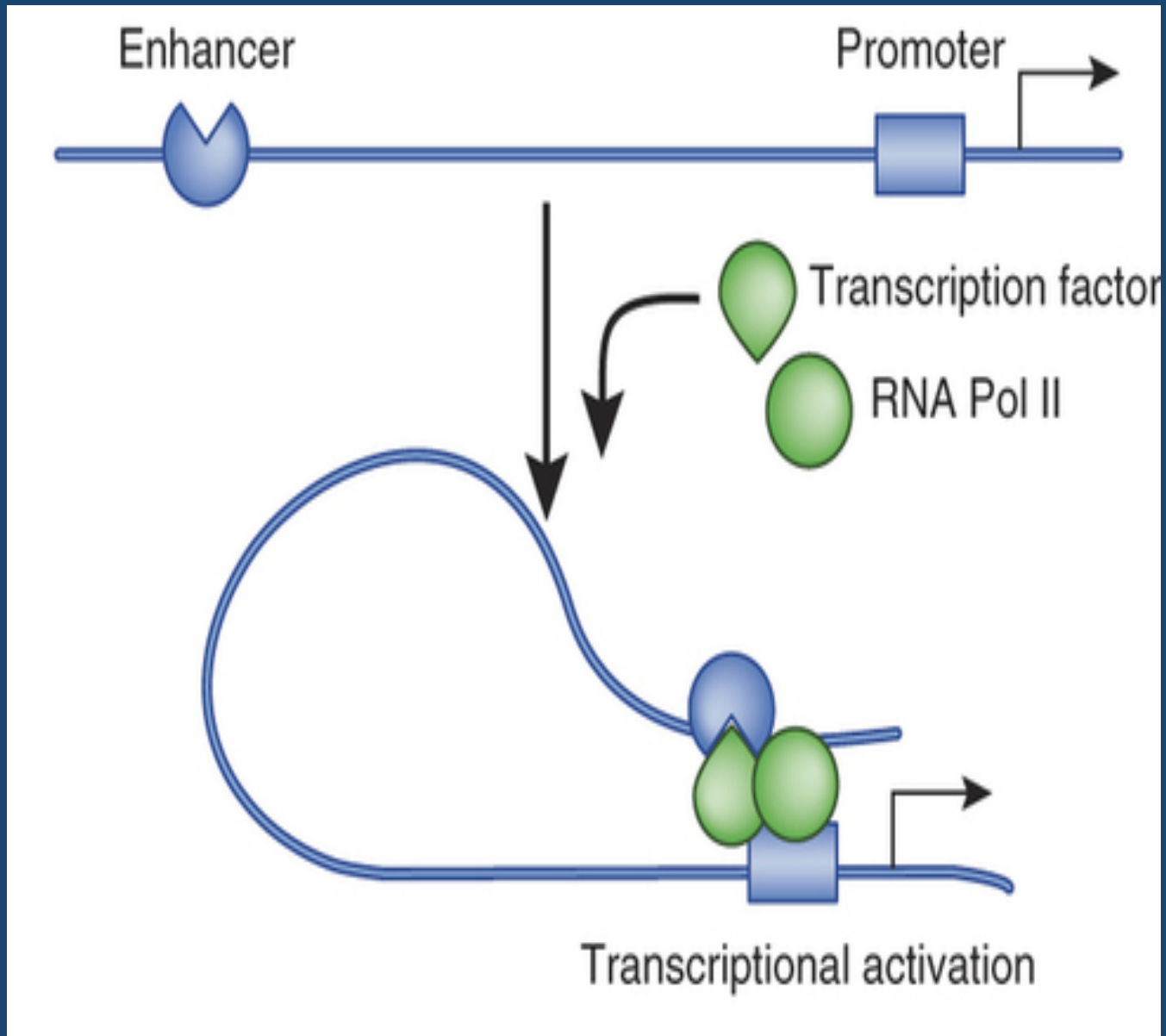
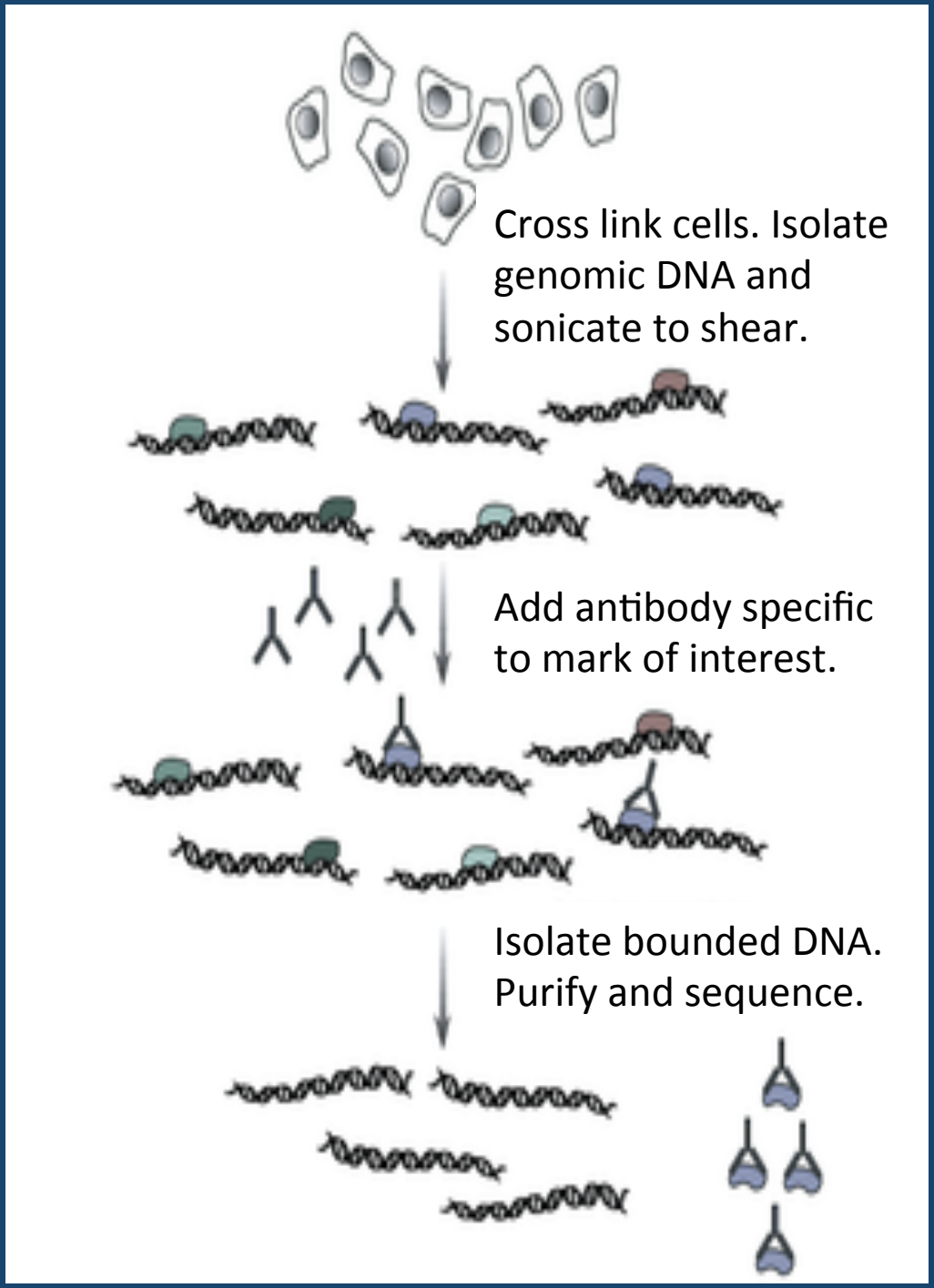


Fig 2b from Cavalli & Misteli (2012) *Nature Structural & Molec Bio*

Detecting Histone Modifications

1. Wet lab
2. Impute

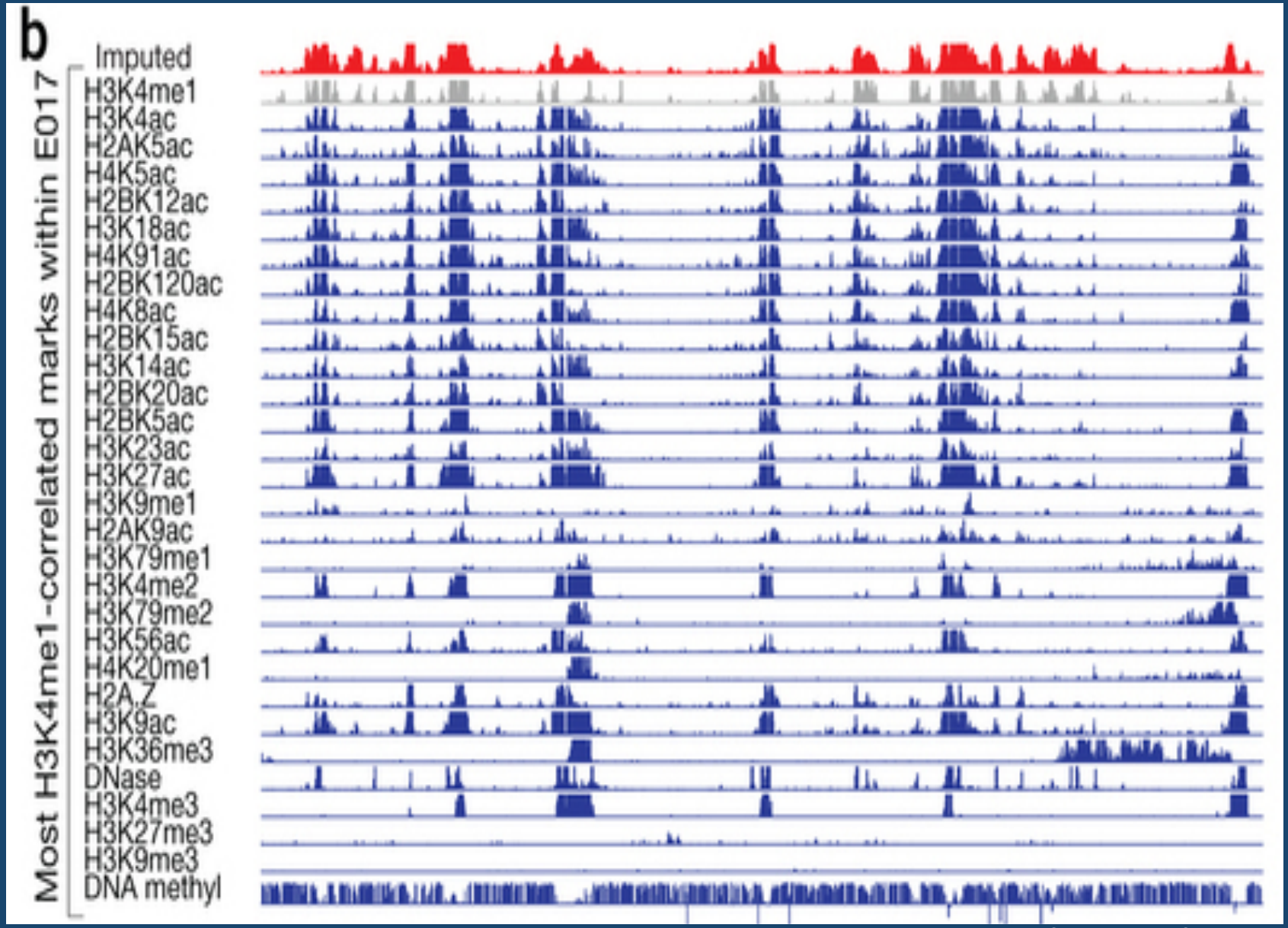
ChiP-seq to map out histone marks



Newer: Chromatin conformation capture techniques (e.g. Hi-C)

http://mmg-233-2013-genetics-genomics.wikia.com/wiki/Chromatin_Immunoprecipitation_(ChIP)

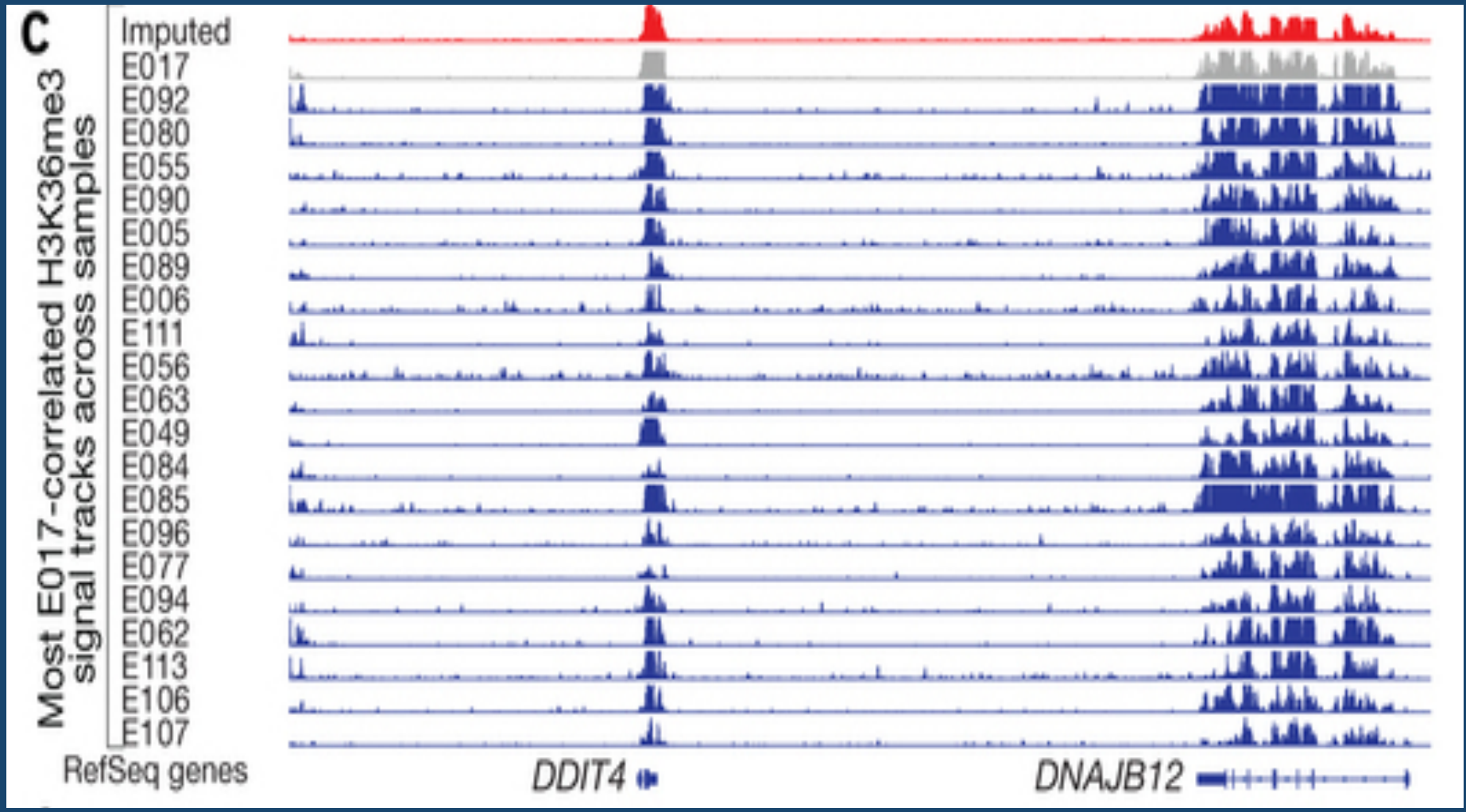
Correlations across marks per sample



Ernst & Kellis (2015) Nat Biotech

E017= fetal lung fibroblast

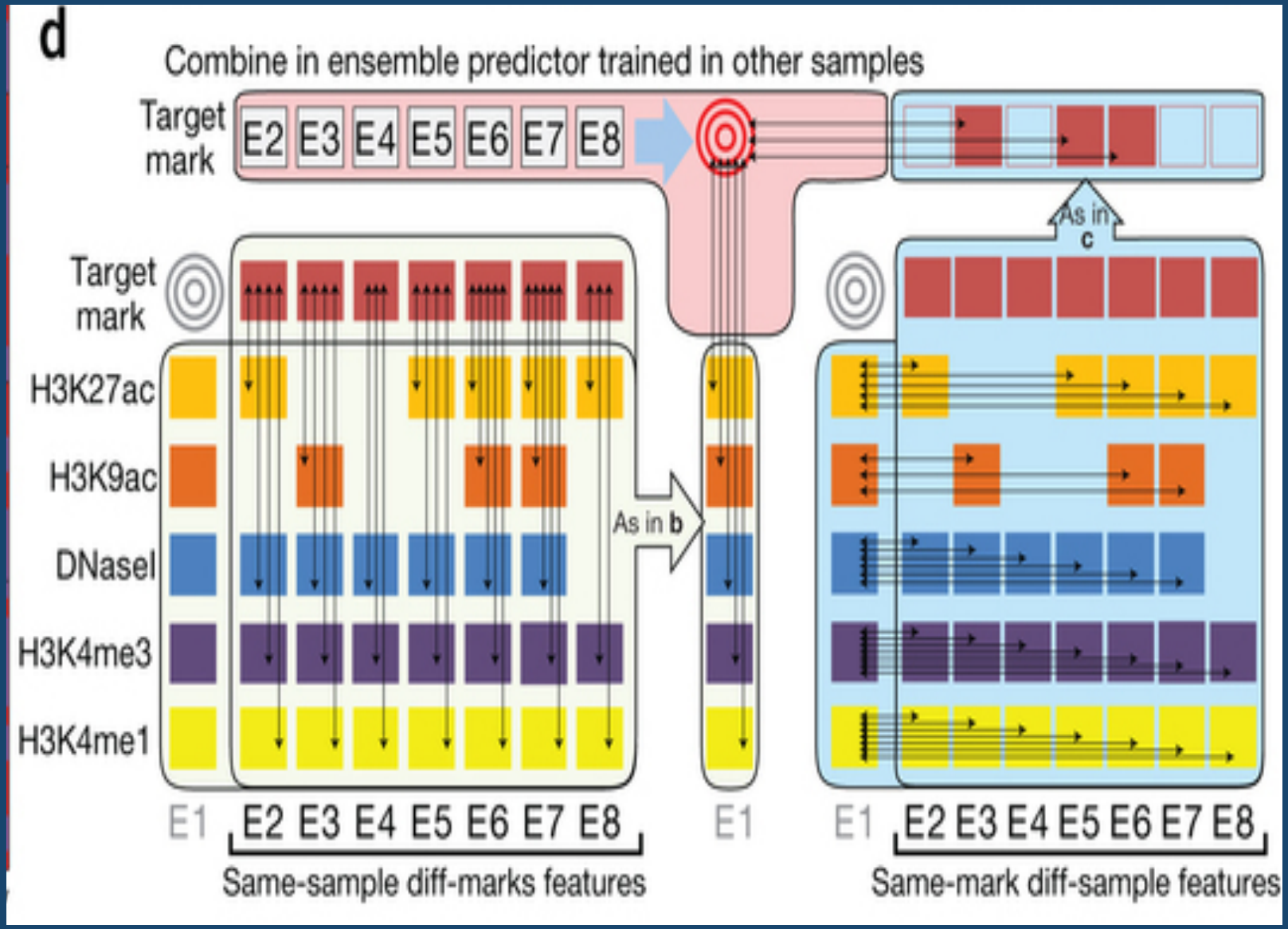
Correlations across samples per mark



Ernst & Kellis (2015) Nat Biotech

E017= fetal lung fibroblast

Use correlations btwn marks & btwn samples



Ernst & Kellis (2015) Nat Biotech

Ensemble of regression trees

Access to Histone Modifications

Roadmap Epigenomics Project (lots of tissues)

ENCODE (lots of cell lines)

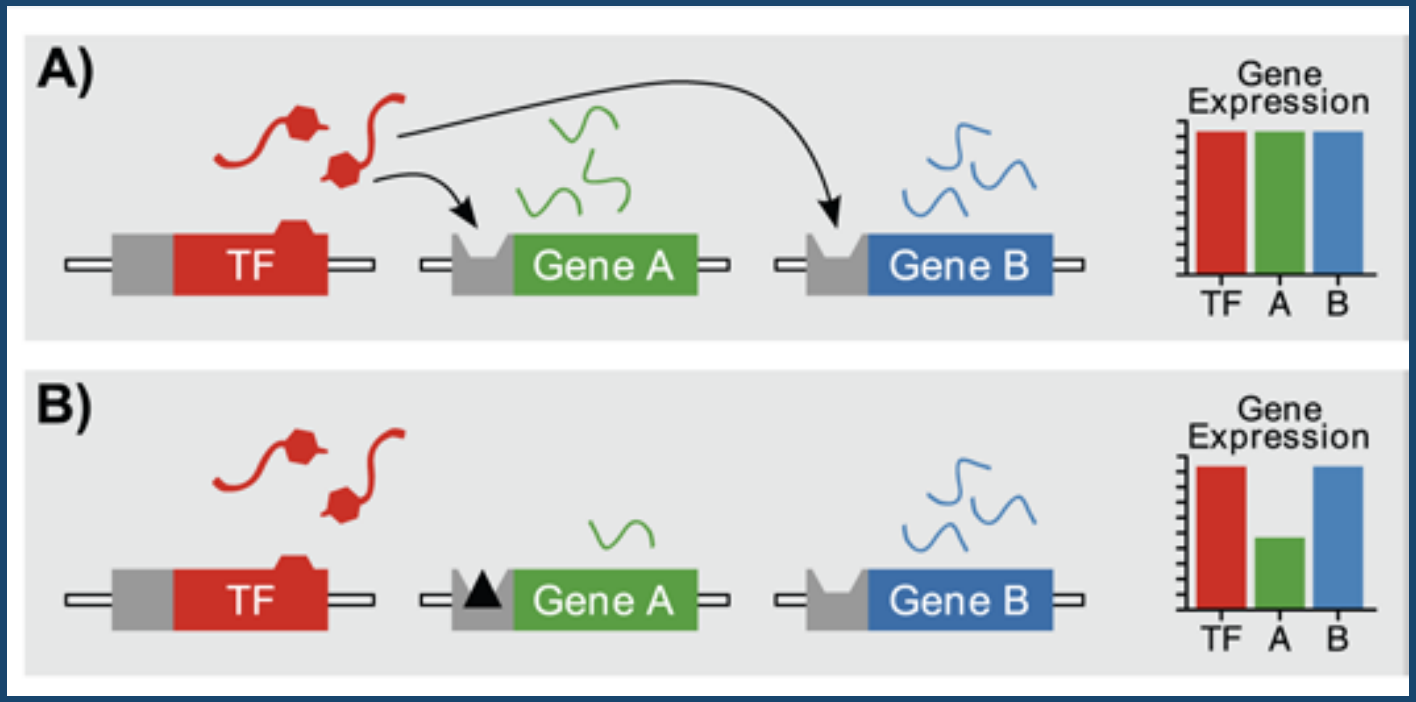
PsychENCODE (brain from control and/or psychiatric samples)

UCSC Genome Browser to visualize

eQTLs

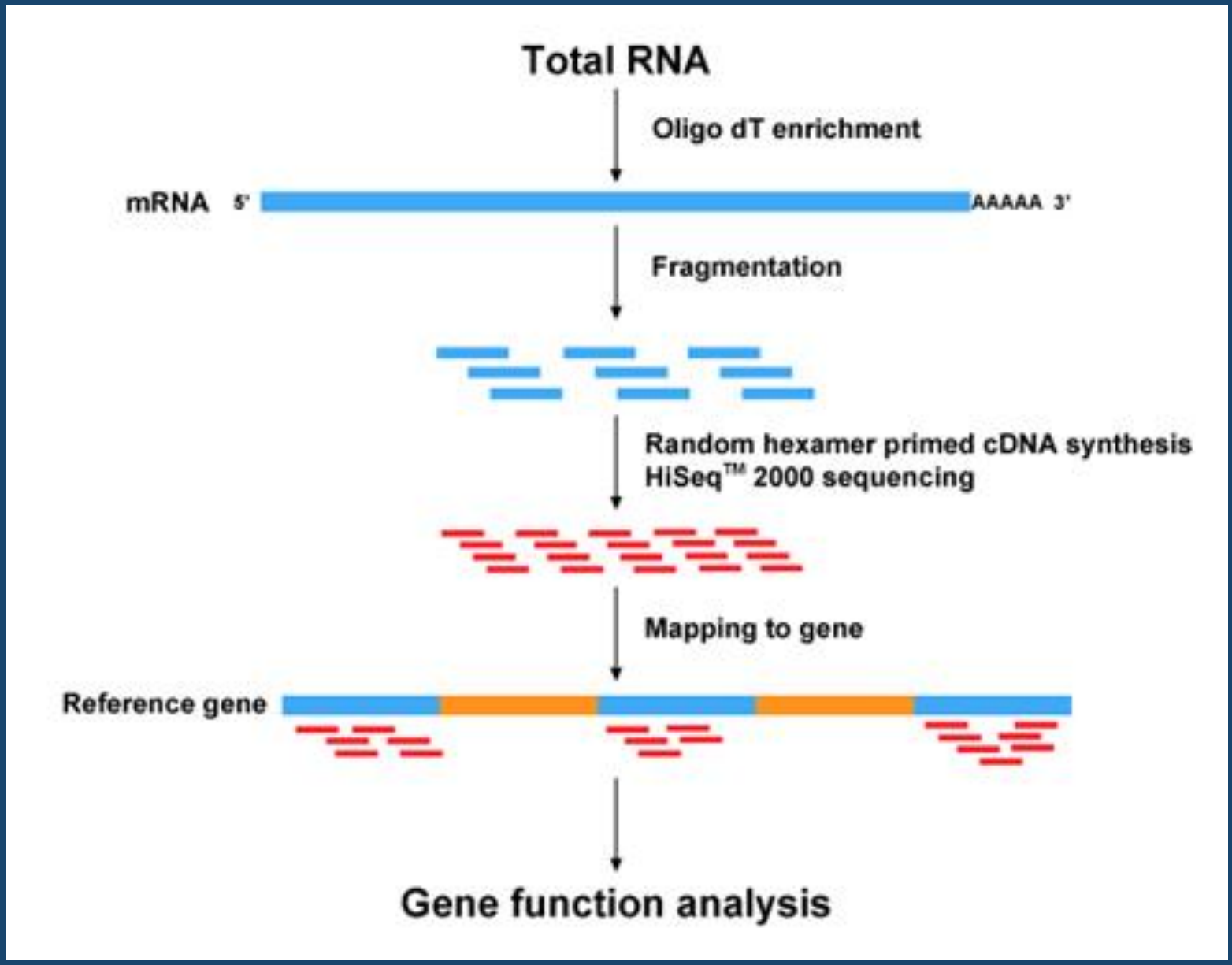
What is an expression quantitative trait locus?

Expression quantitative trait loci (eQTLs) = DNA loci that regulate expression levels of RNAs



Adapted from: Wolen AR, Miles MF. Identifying gene networks underlying the neurobiology of ethanol and alcoholism. Alcohol Res. 2012;34(3):306-17.

RNA-seq to identify eQTLs



Lundberg

Access to eQTLs

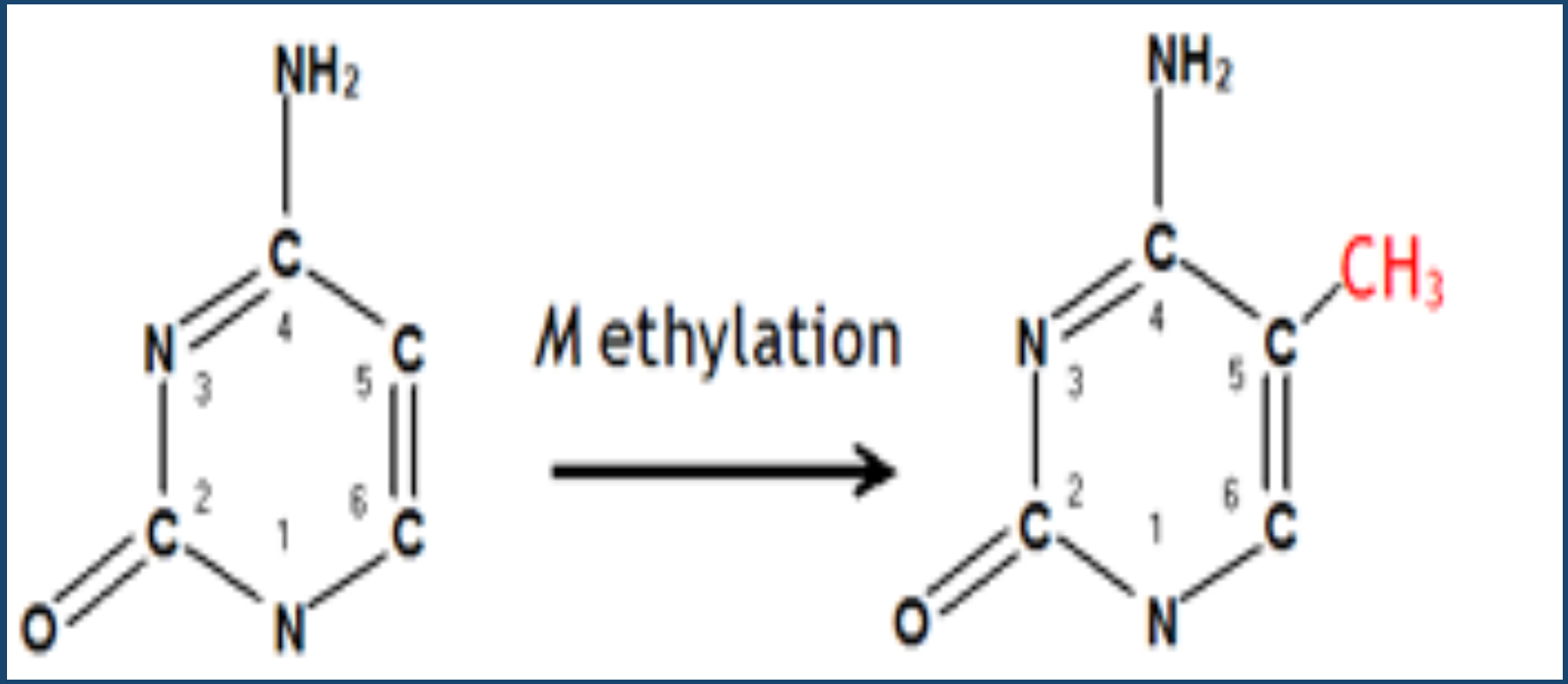
GTEx (44 tissues with eQTLs)
http://www.gtexportal.org/

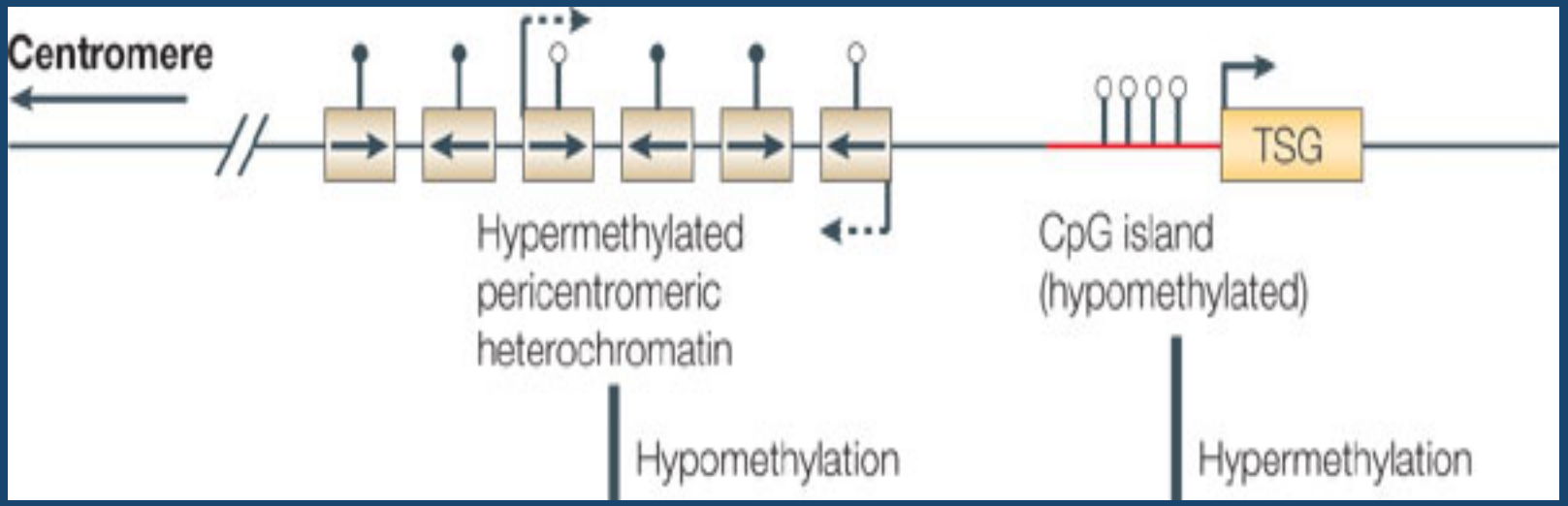
UKBEC (10 brain tissues with eQTLs)
http://braineac.org/

Individual studies

DNA methylation

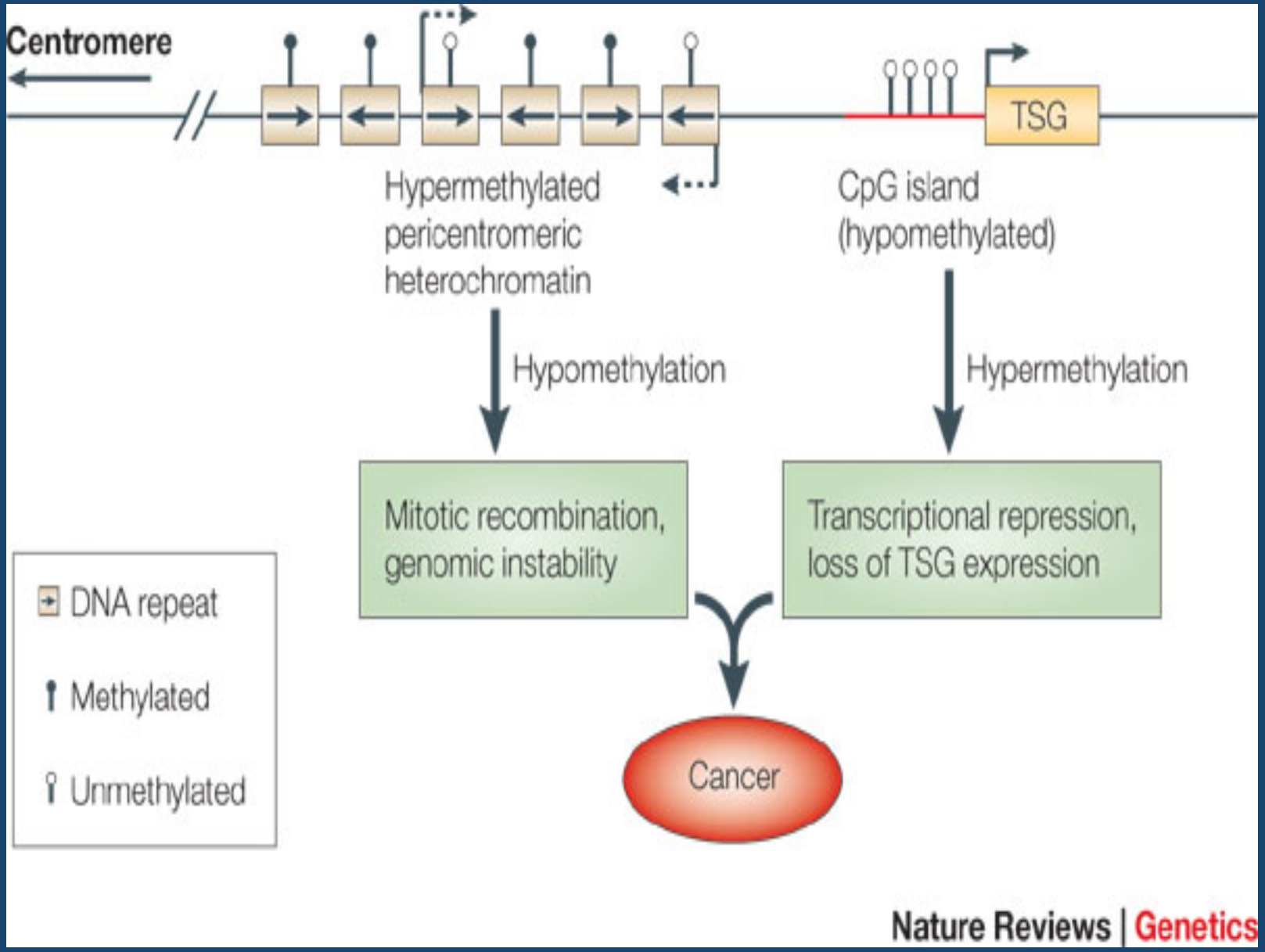
5-methyl-Cytosine



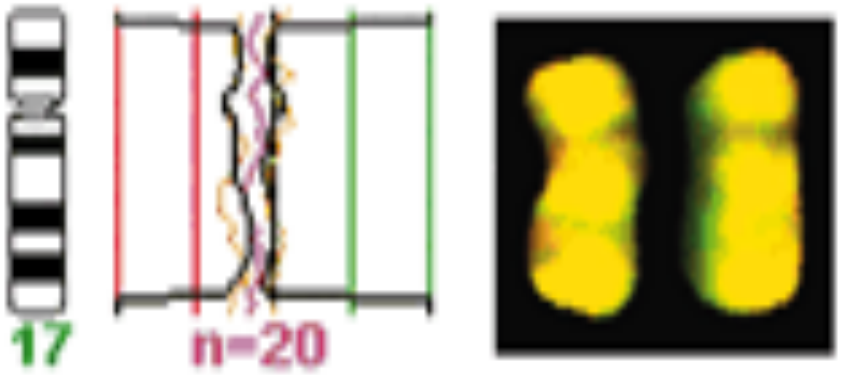


TSG= tumor suppressor gene

Robertson (2005) *Nature Review Genetics*


Centromere
TSG
Hypermethylated
pericentromeric
heterochromatin
CpG island
(hypomethylated)
Hypomethylation
Hypermethylation
Mitotic recombination,
genomic instability
Transcriptional repression,
loss of TSG expression
Cancer
DNA repeat
Methylated
Unmethylated
Nature Reviews | Genetics

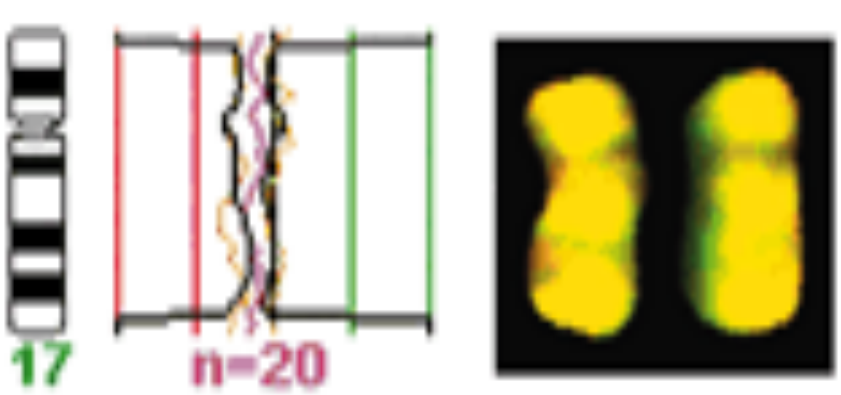
DNA Methylation changes over time



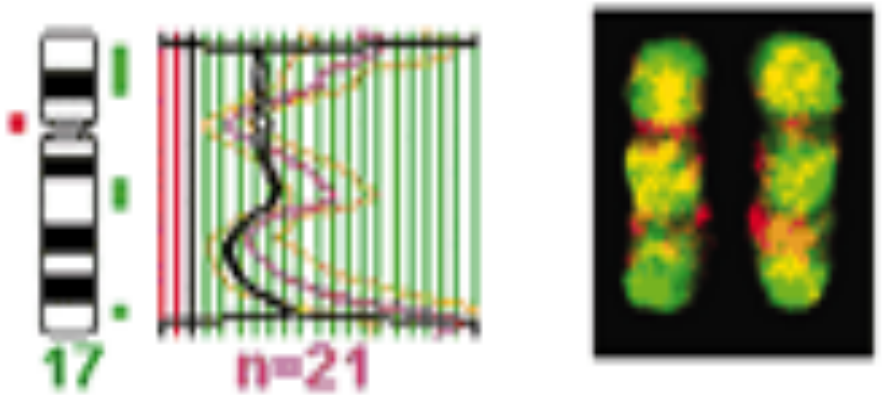
3-year-old twins

Fraga et al. (2005) PNAS

DNA Methylation changes over time



3-year-old twins



50-year-old twins

Fraga et al. (2005) PNAS


F·A·O·SCHWARZ

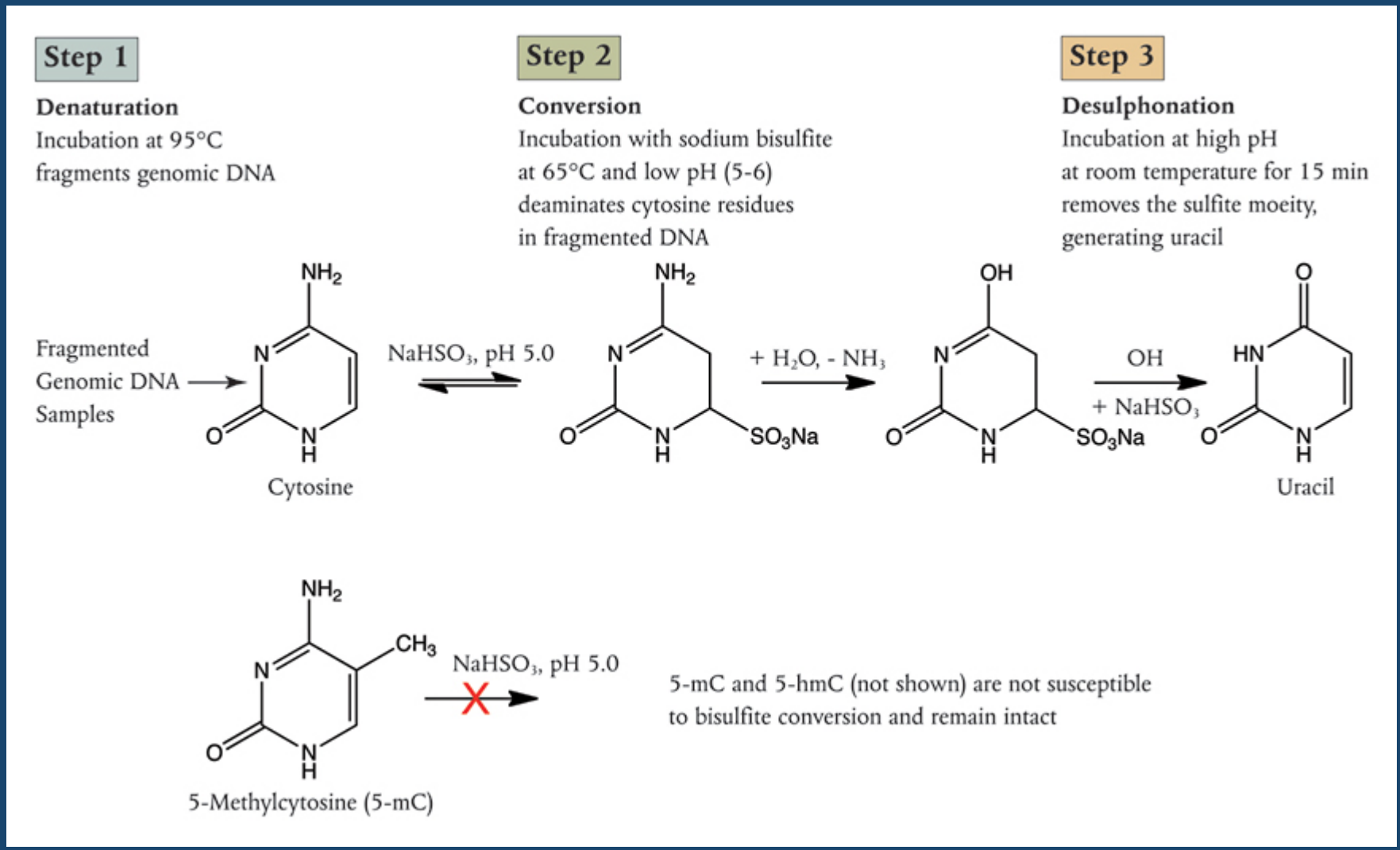
Arrays to assess DNA CH_3

Illumina HumanMethylation450K

Illumina EPIC (>850K CH_3 sites)

- >90% content on 450K
- New: CH_3 sites in ENCODE open chromatin & enhancers

(Gold Standard) Bisulfite Conversion of gDNA

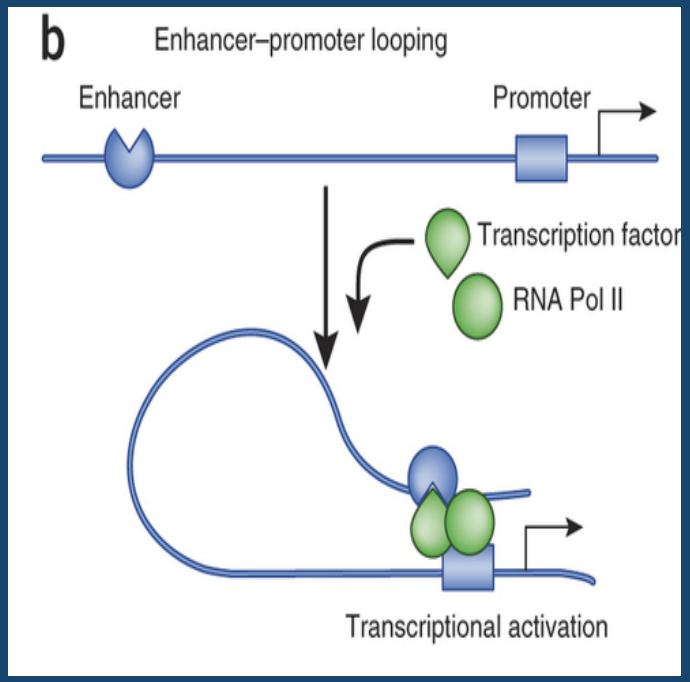


Genetic variation → gene regulation

phenotype

Non-coding variation

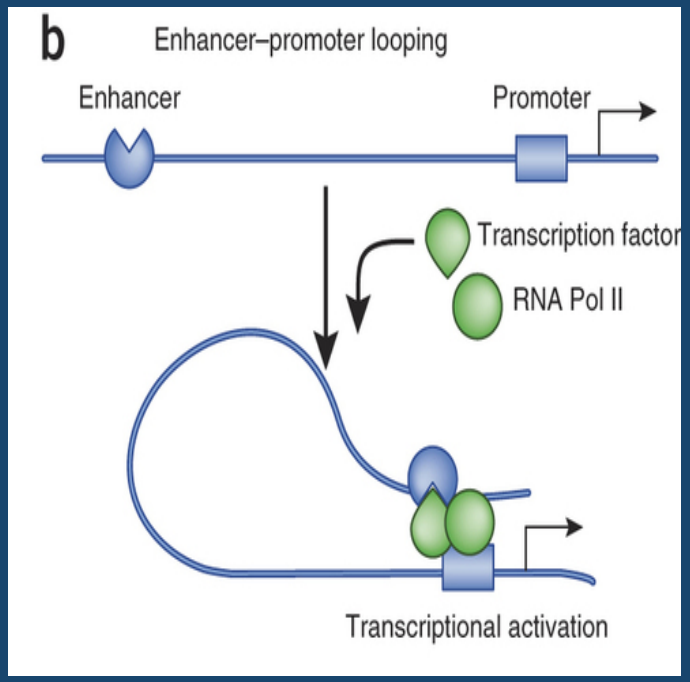
Genetic variation → gene regulation



phenotype

Non-coding variation

Genetic variation → gene regulation



phenotype

Non-coding variation

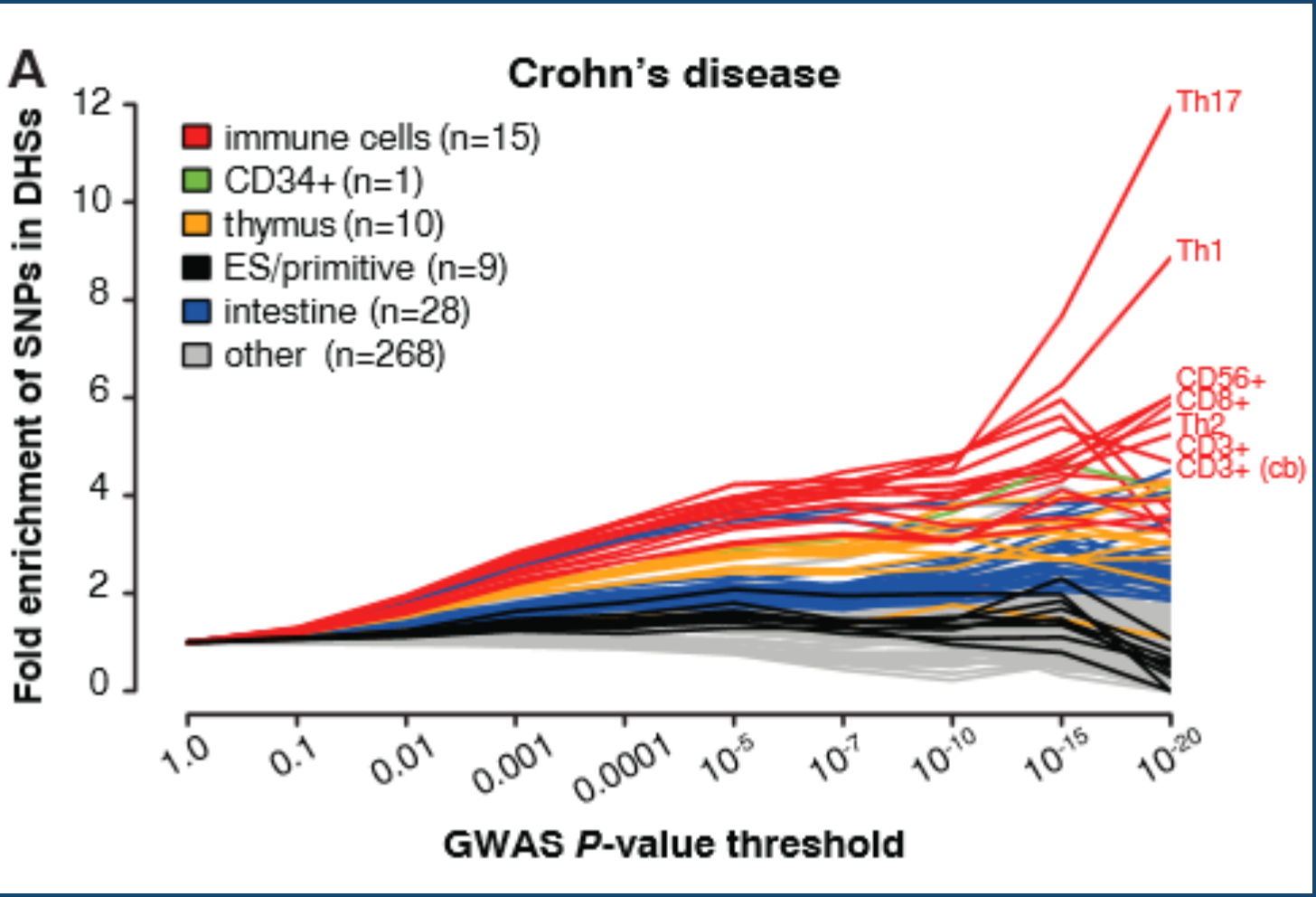
Key Points

- Most variants identified through GWAS are noncoding
- Non-coding DNA help regulate transcription (e.g. histone marks, eQTLs, DNA methylation)
- Functional annotations are dynamic

Outline

- Functional Genomics 101
- Using functional genomics to prioritize risk variants

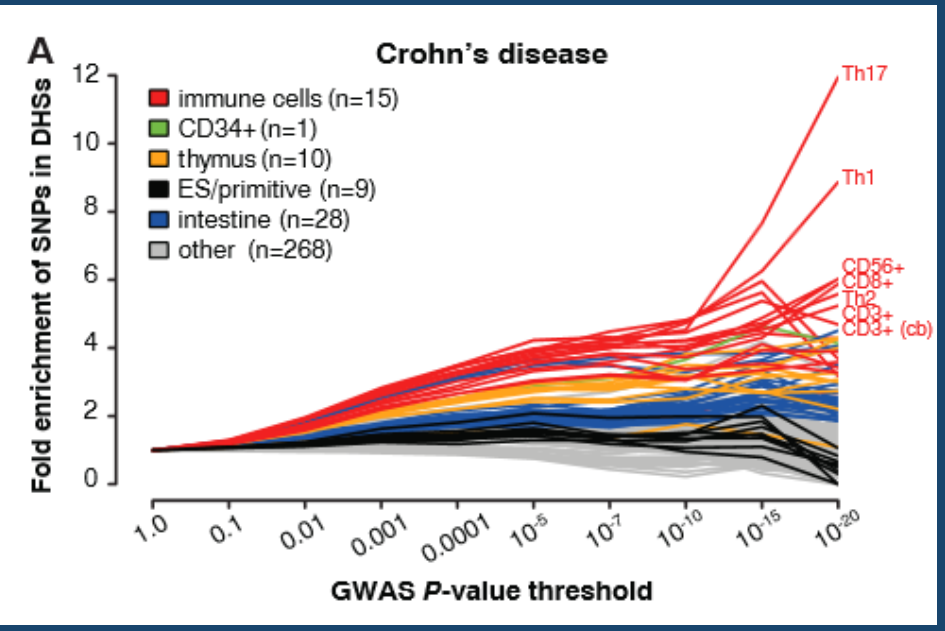
GWAS hits overlap with functional regions

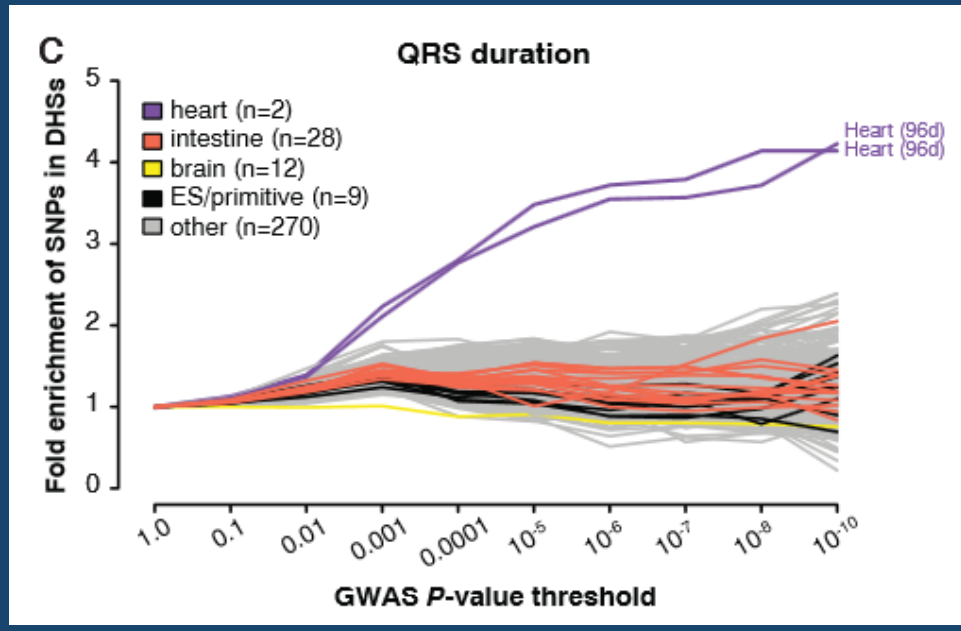


(Franke et al. 2010)
N SNPs= 938,703 N inds= 6,333 cases & 15,056 controls

Figure from Maurano et al. (2012) *Science*, 337: 1190-1195

GWAS hits overlap with functional regions



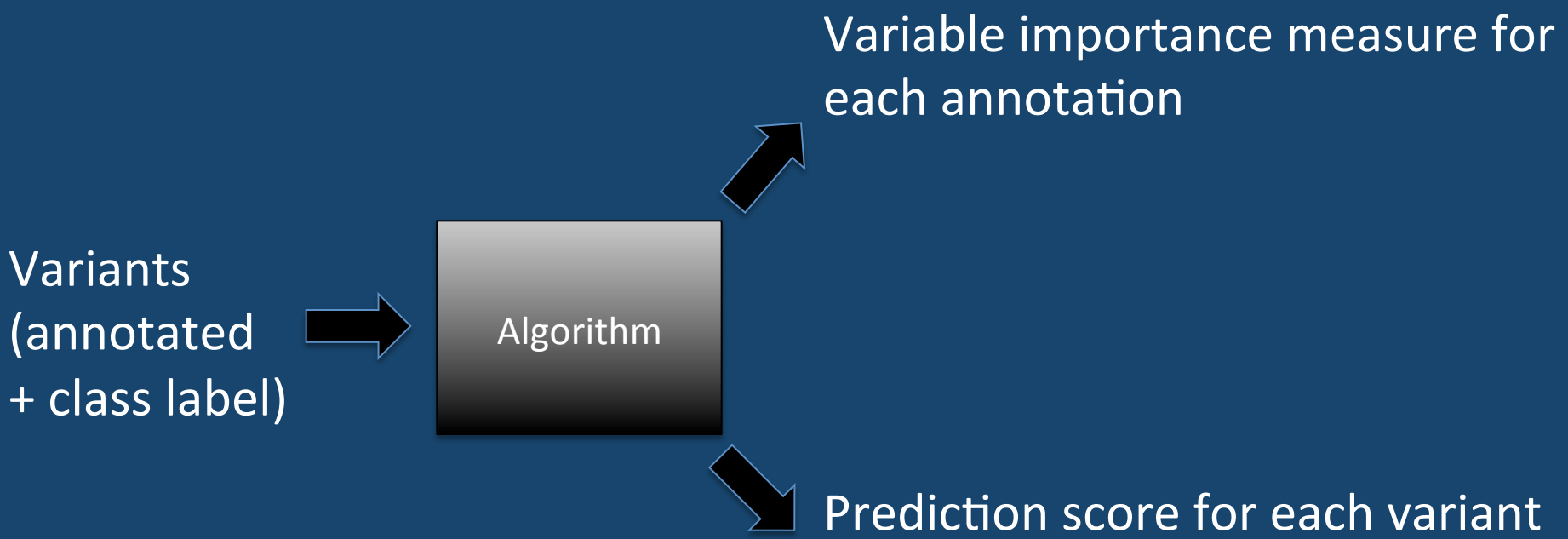


(Franke et al. 2010)
N SNPs= 938,703
N inds= 6,333 cases & 15,056 controls

(Sotoodehnia et al. 2010)
N SNPs ~2.5 M
N inds~ 40K

Figures from Maurano et al. (2012) *Science*, 337: 1190-1195

Use functional info to identify hits



Machine Learning Steps

Supervised vs. Unsupervised

Supervised vs. Unsupervised

Class assignment
(data are labelled)

Pattern discovery
(data aren't labelled)

A simple ML application

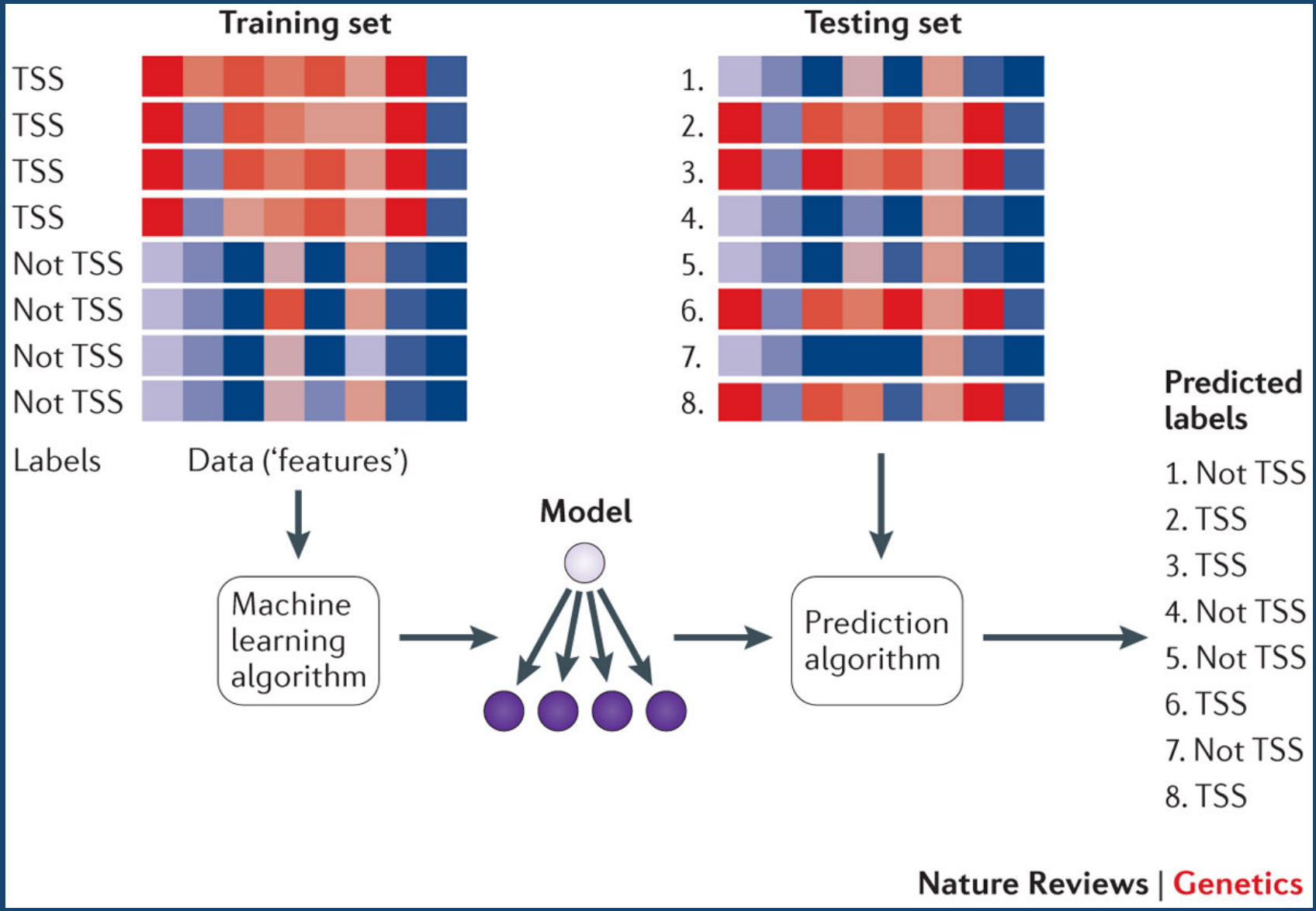


Figure 1 from Libbrecht and Noble 2015

Quiz! Supervised or unsupervised?

Quiz! Supervised or unsupervised?

Find novel TFBS based on annotated TFBS in the JASPAR database vs. non-TFBS sequences

Quiz! Supervised or unsupervised?

Find novel TFBS based on annotated TFBS in the JASPAR database vs. non-TFBS sequences

Identify clusters of similar tumors in cancer patients

Quiz! Supervised or unsupervised?

Find novel TFBS based on annotated TFBS in the JASPAR database vs. non-TFBS sequences

Identify clusters of similar tumors in cancer patients

Identify risk variants from other variants trained on genomic characteristics for *Human Gene Mutation Database* SNPs vs. 1KG (GWAVA, Ritchie et al. 2014)

Quiz! Supervised or unsupervised?

Find novel TFBS based on annotated TFBS in the JASPAR database vs. non-TFBS sequences

Identify clusters of similar tumors in cancer patients

Identify risk variants from other variants trained on genomic characteristics for *Human Gene Mutation Database* SNPs vs. 1KG (GWAVA, Ritchie et al. 2014)

Segment the genome into X chromatin states using chromatin mark patterns (Segway, Hoffman et al. 2012)

OPEN ACCESS Freely available online

PLOS | ONE

A Bayesian Method to Incorporate Hundreds of Functional Characteristics with Association Evidence to Improve Variant Prioritization

Sarah A. Gagliano[1,2], Michael R. Barnes[3], Michael E. Weale[4☯], Jo Knight[1,2,5*☯]

TECHNICAL REPORTS

nature genetics

A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher[1,5], Daniela M Witten[2,5], Preti Jain[3,4], Brian J O'Roak[1,4], Gregory M Cooper[3] & Jay Shendure[1]

BRIEF COMMUNICATIONS

Functional annotation of noncoding sequence variants

Graham R S Ritchie[1,2], Ian Dunham[1], Eleftheria Zeggini[2] & Paul Flicek[1,2]

TECHNICAL REPORTS

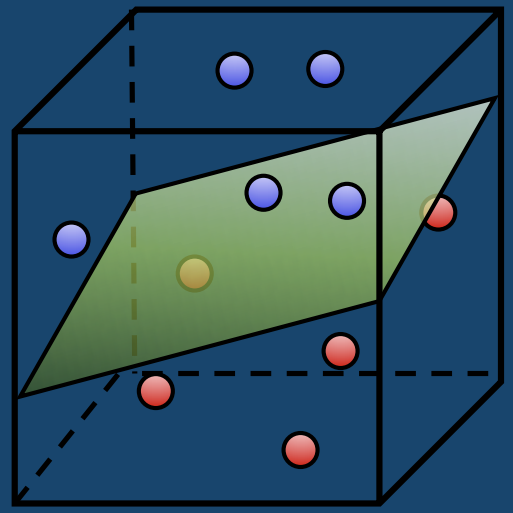
nature genetics

A method to predict the impact of regulatory variants from DNA sequence

Dongwon Lee[1,4], David U Gorkin[1,3,4], Maggie Baker[1], Benjamin J Strober[2], Alessandro L Asoni[2], Andrew S McCallion[1] & Michael A Beer[1,2]

Support Vector Machines

Maximum-margin separating hyperplane in multi-dimensional space



Slide from Reena Ravji

Which line provides the best classifier?

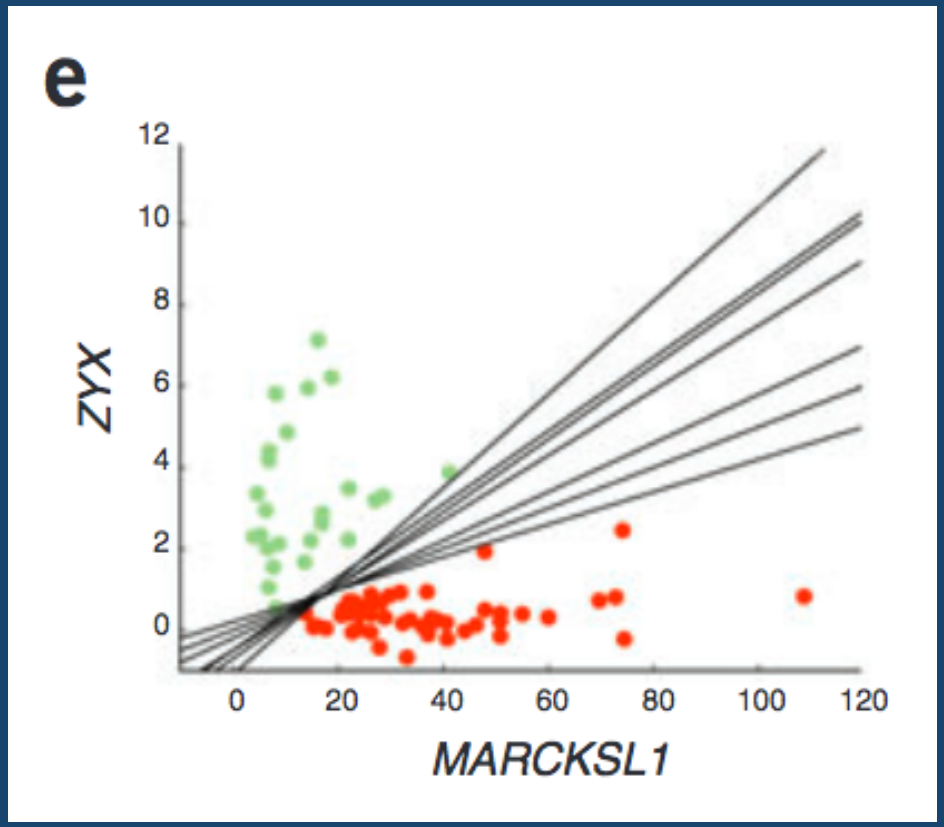


Figure 1 Noble 2016 *Nat.Biotech*

Which line provides the best classifier?

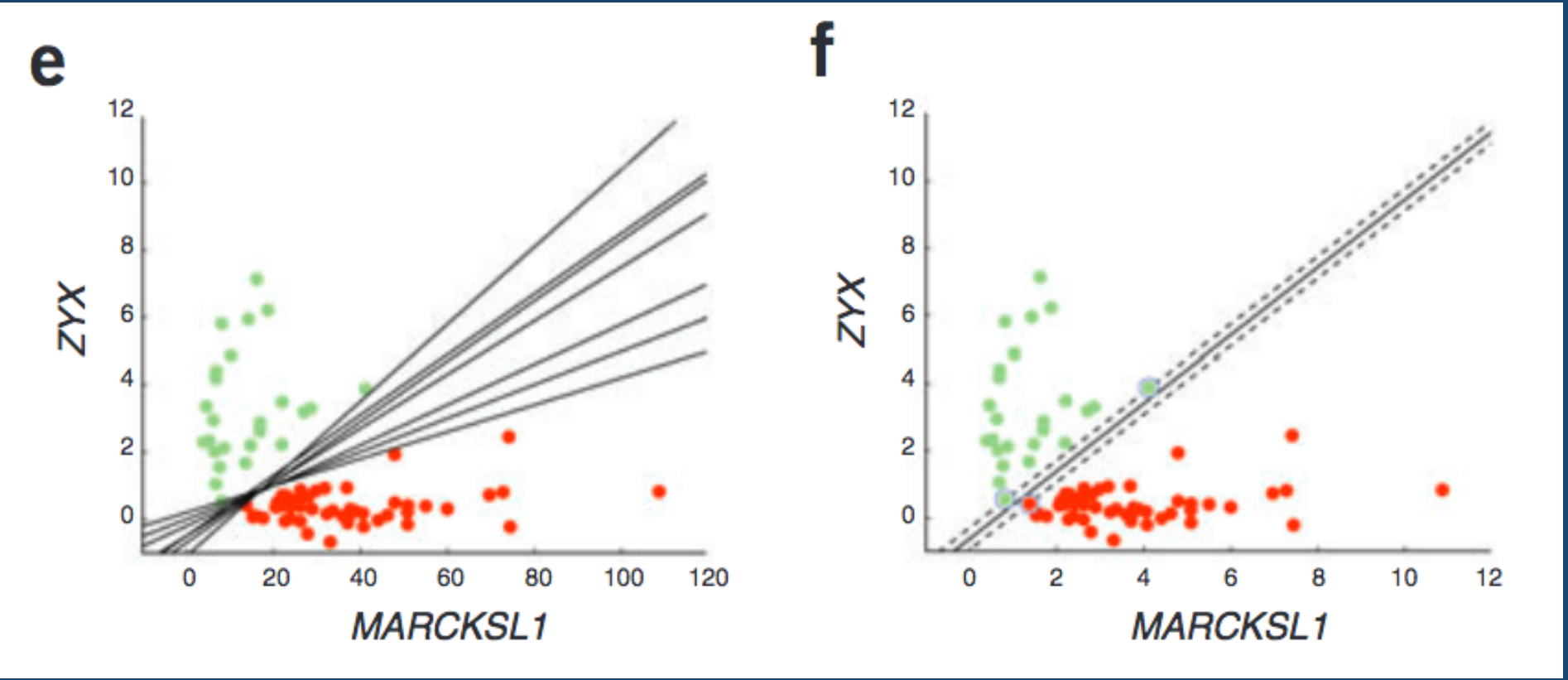


Figure 1 Noble 2016 *Nat.Biotech*

The “Soft Margin”

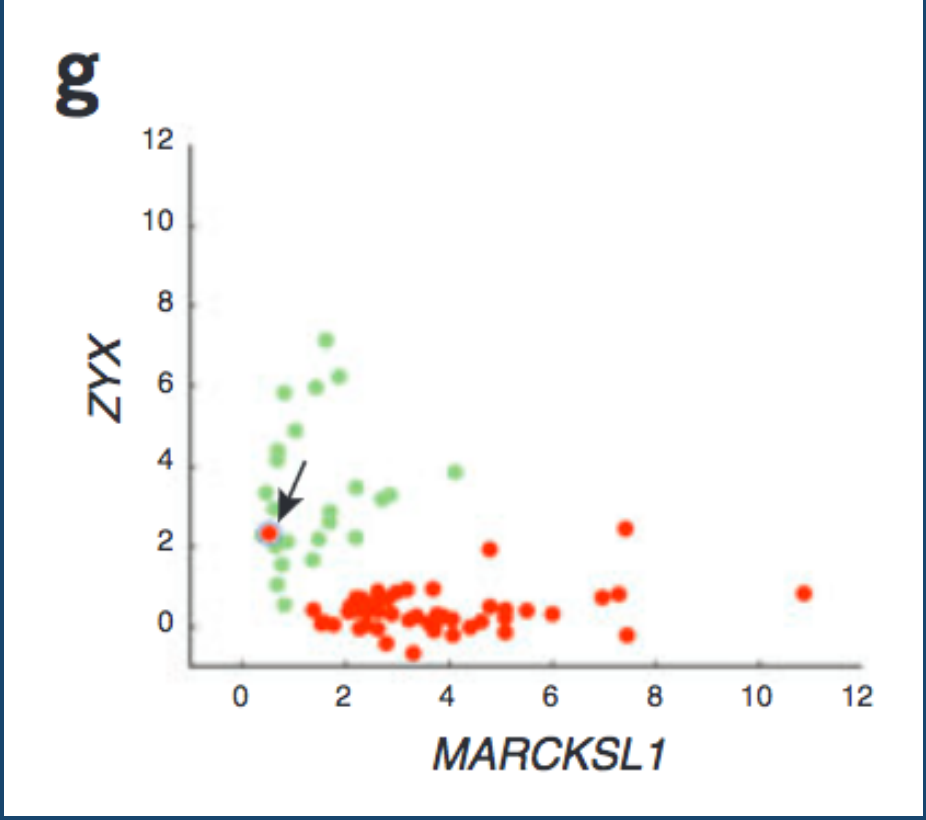


Figure 1 Noble 2016 *Nat.Biotech*

The “Soft Margin”

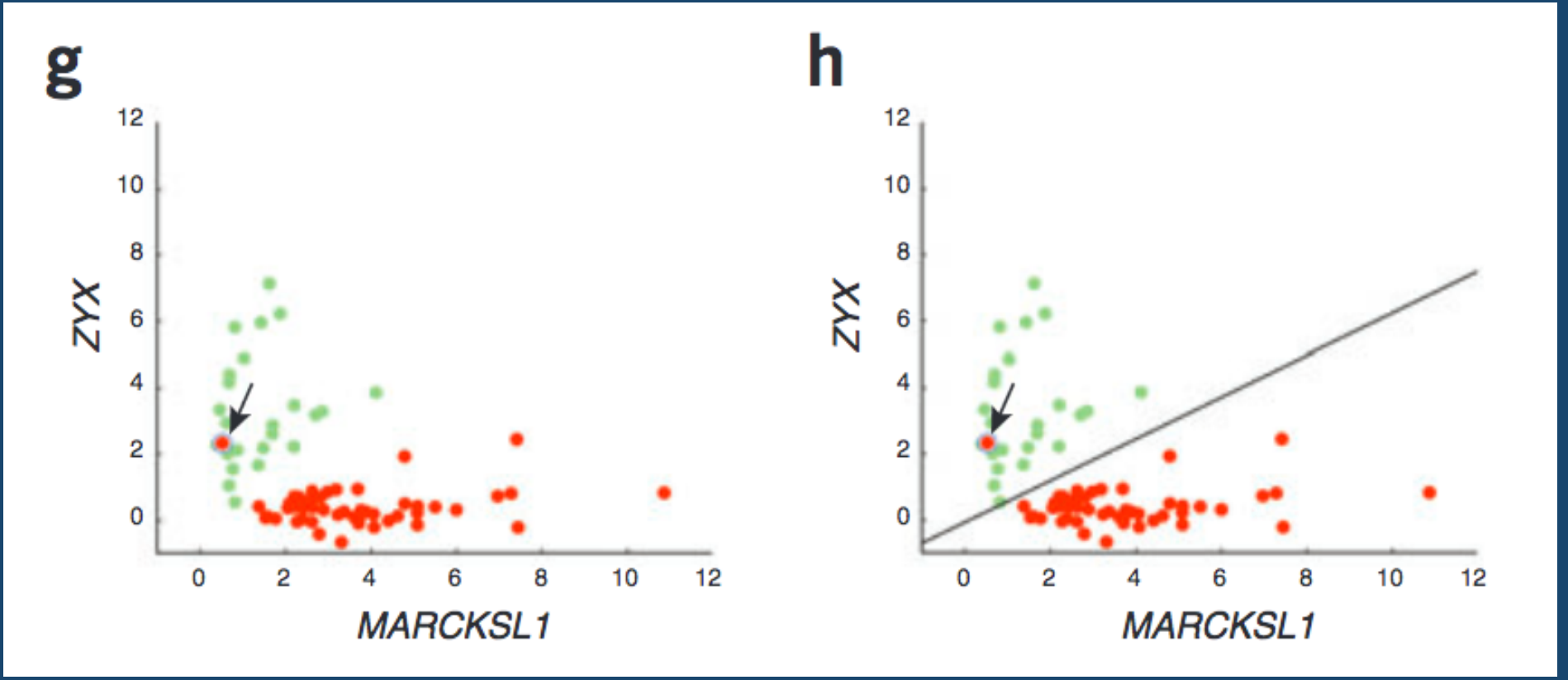
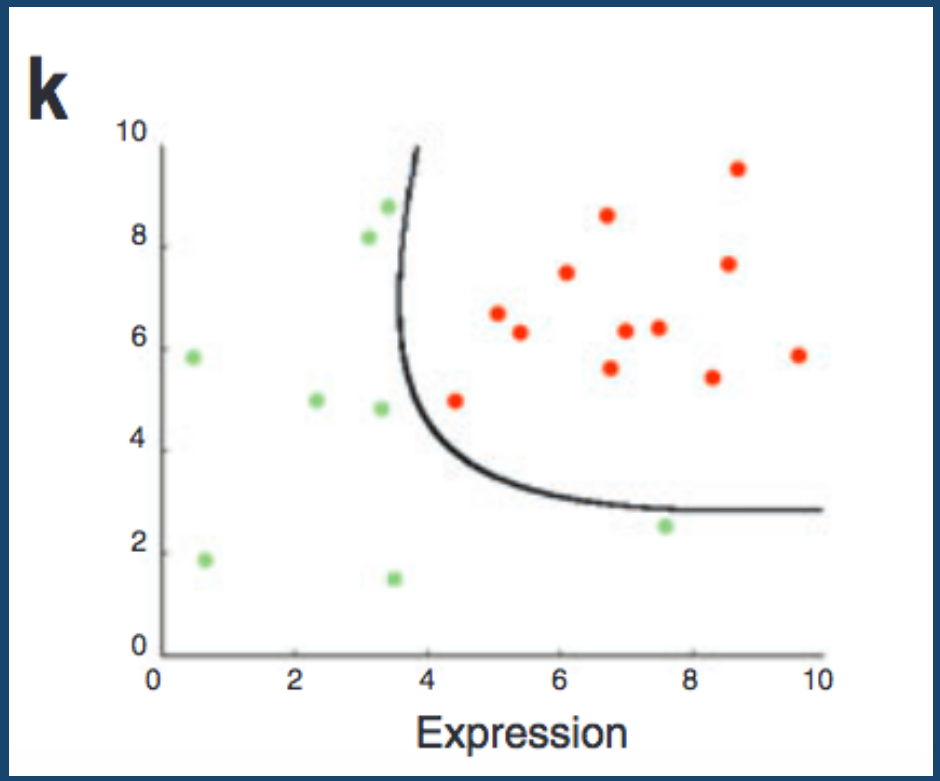


Figure 1 Noble 2016 *Nat.Biotech*

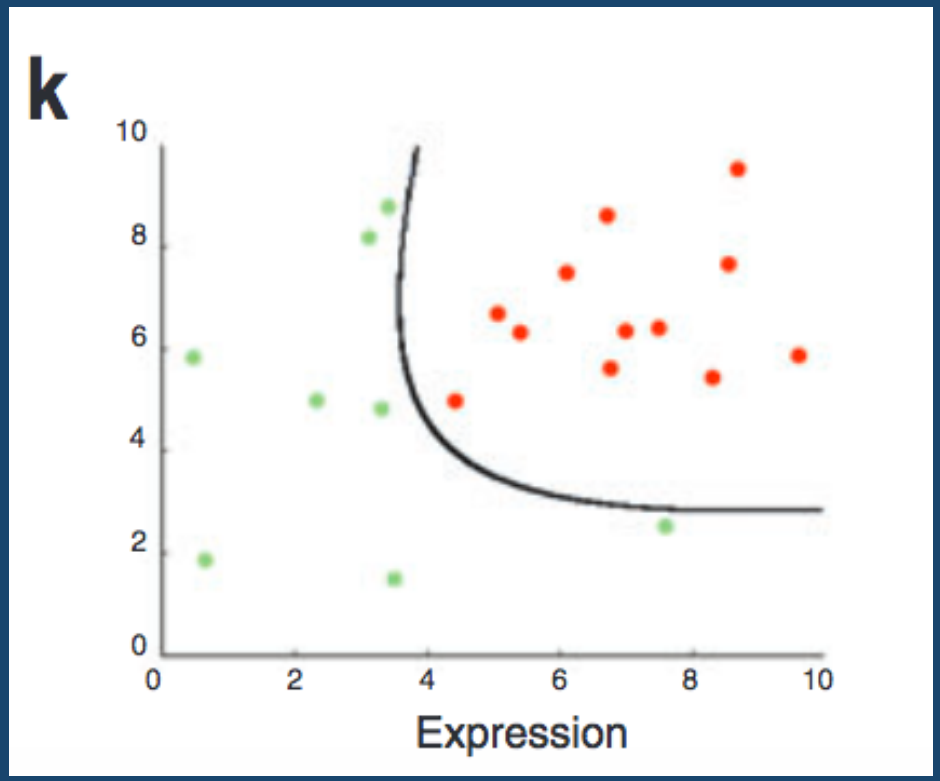
Linearly nonseparable? Kernel Trick



1. Kernel function to project the data in 2-D to 4-D space

Figure 1 Noble 2016 *Nat.Biotech*

Linearly nonseparable? Kernel Trick



1. Kernel function to project the data in 2-D to 4-D space
2. Project SVM hyperplane in 4-D to original 2-D space

Figure 1 Noble 2016 *Nat.Biotech*

1. Classifier
2. SVM protocol
3. Some results

TECHNICAL REPORTS

nature genetics

A general framework for estimating the relative pathogenicity of human genetic variants

Martin Kircher[1,5], Daniela M Witten[2,5], Preti Jain[3,4], Brian J O'Roak[1,4], Gregory M Cooper[3] & Jay Shendure[1]

1. Classifier

Simulated vs. Observed variants

Simulated (14.7 million variants):

- Empirical model of sequence evolution with local adjustment of mutation rates

Simulated vs. Observed variants

Simulated (14.7 million variants):

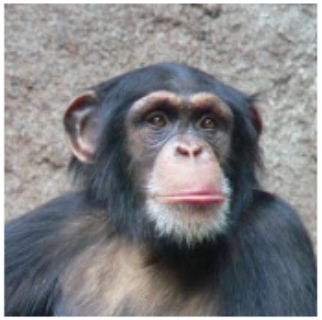
- Empirical model of sequence evolution with local adjustment of mutation rates

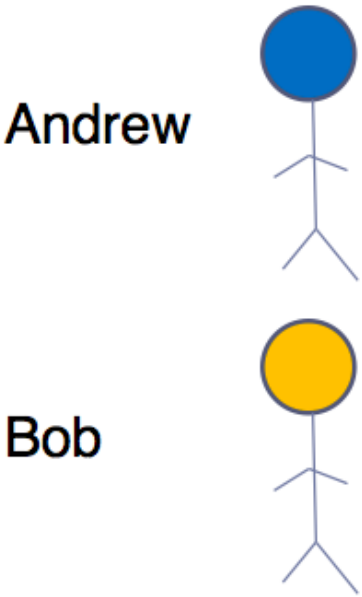
Observed (14.7 million variants):

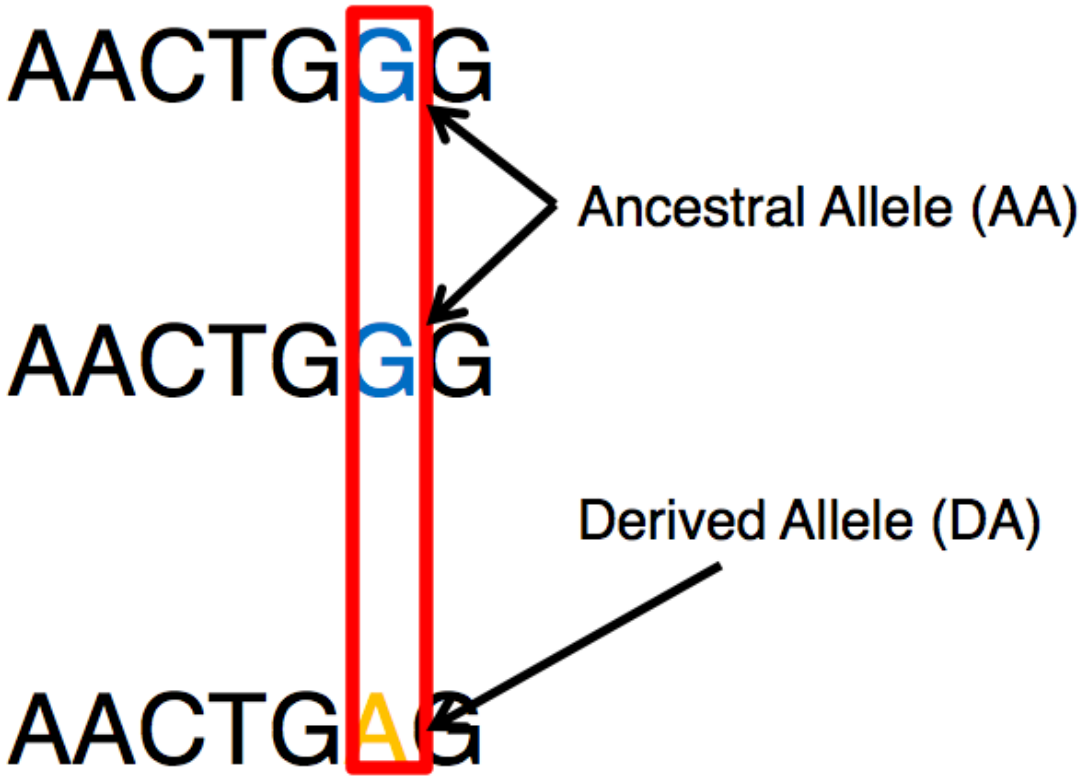
- Human derived allele >95% (exposed to many generations of natural selection)

1% reserved for testing

Ancestral Allele (AA) and Derived Allele (DA)







John Long (Kellis Lab) "Evidence of Purifying Selection in Humans"
https://math.mit.edu/research/highschool/primes/materials/2013/conf/9-2-Long.pdf

Peppered Moth Evolution

Ancestral

Derived

No pollution

Photo from: http://www.truthinscience.org.uk/tis2/index.php/evidence-for-evolution-mainmenu-65/127-the-peppered-moth.html

Peppered Moth Evolution

Ancestral

Derived

No pollution

Pollution

Photo from: http://www.truthinscience.org.uk/tis2/index.php/evidence-for-evolution-mainmenu-65/127-the-peppered-moth.html

2. SVM Protocol

Kircher's use of SVM

- Linear kernel
- Prior feature selection (univariate analysis)
- Imputed missing values
- Boolean variables for categorical variables
- Interaction terms (include the few that improve the two-feature linear regression models)

3. Some results

Performance in GWAS

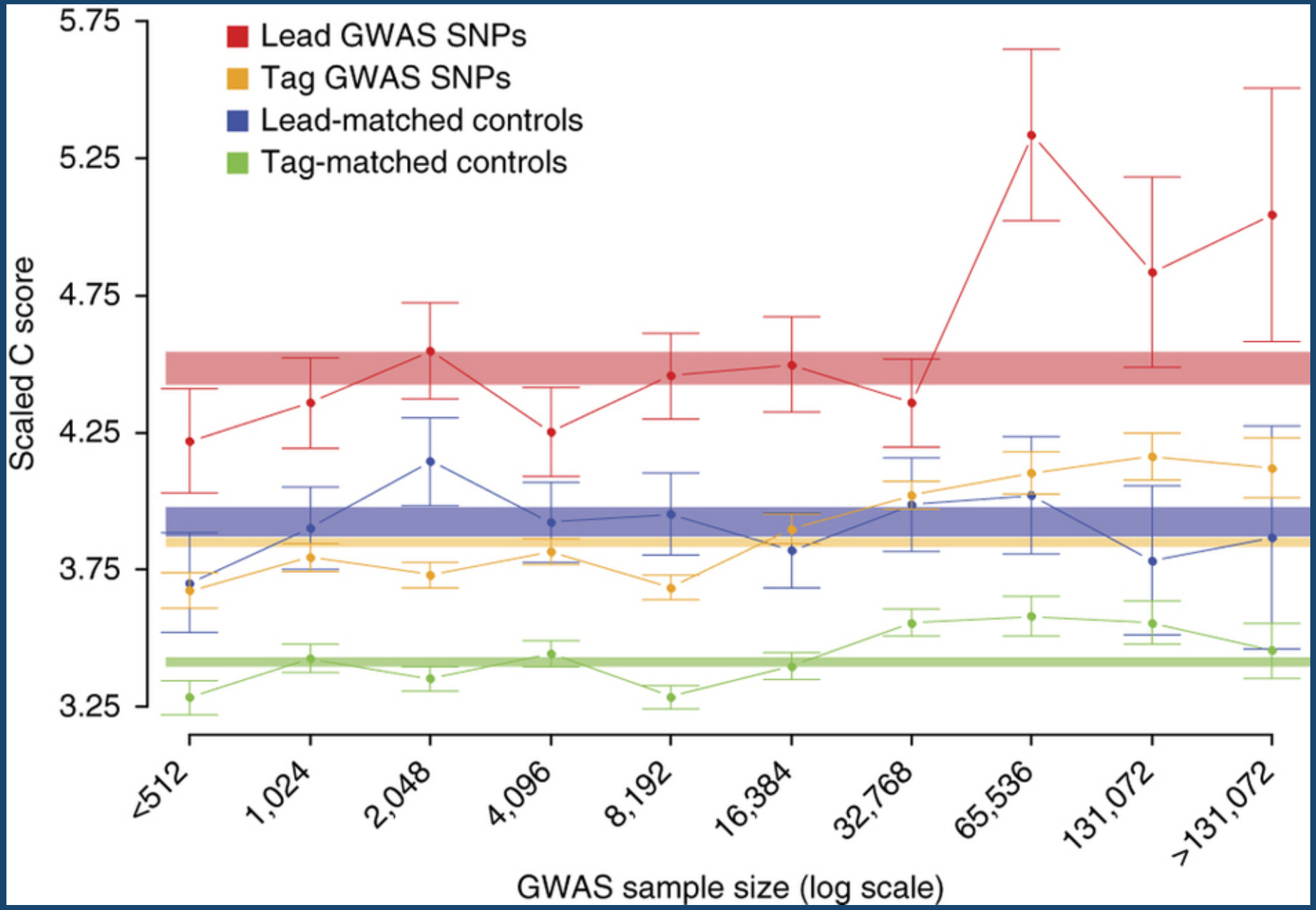
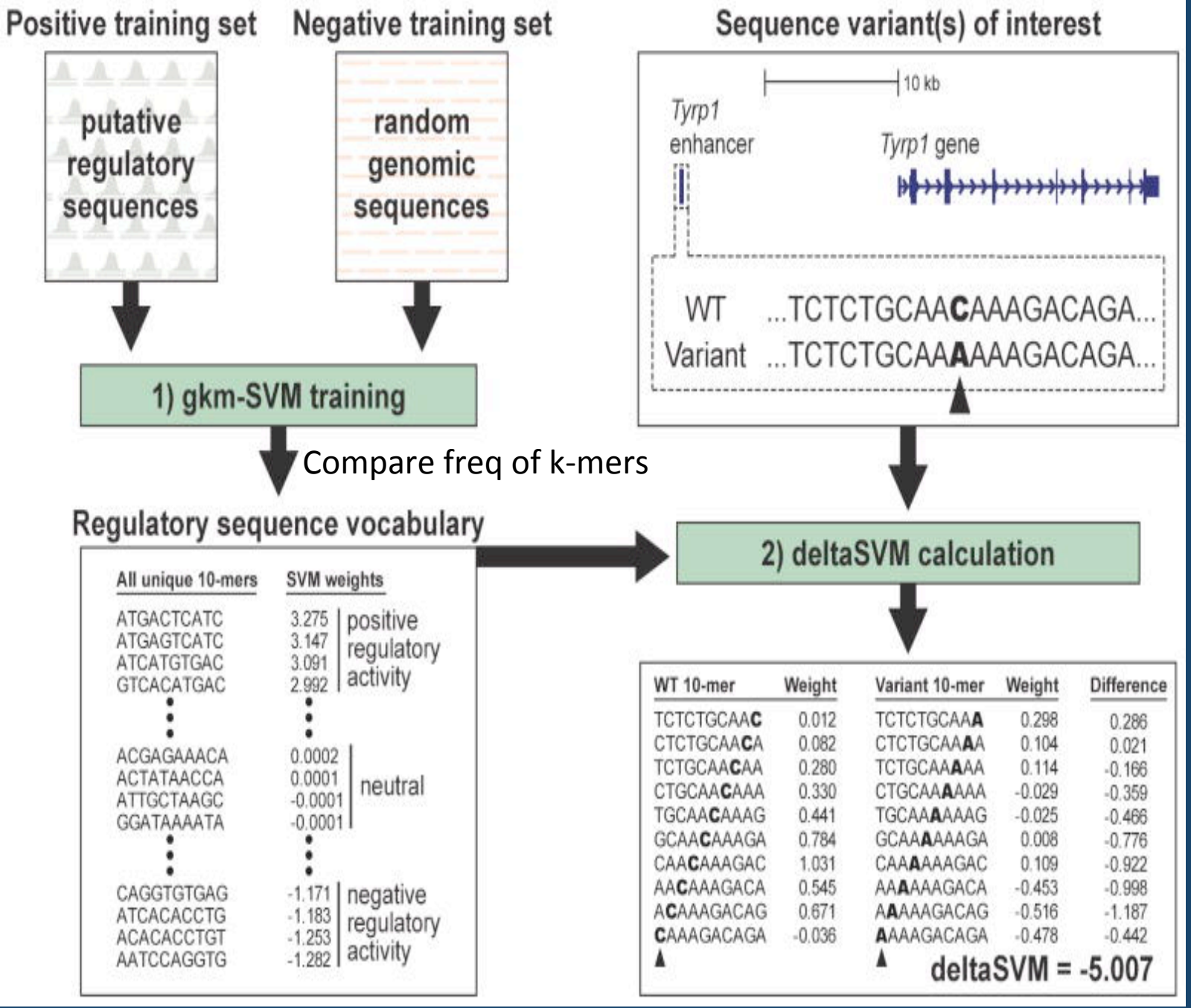


Figure 5 from Kircher *et al*. 2014 *Nat.Genetics*

deltaSVM

1. Classifier
2. SVM protocol
3. Some results

- Takes into account tissue-specificity



Cell-type-specific annots for training cell-type-specific models

Figure 1 from Lee *et al*. 2015 *Nat.Genetics*

deltaSVM

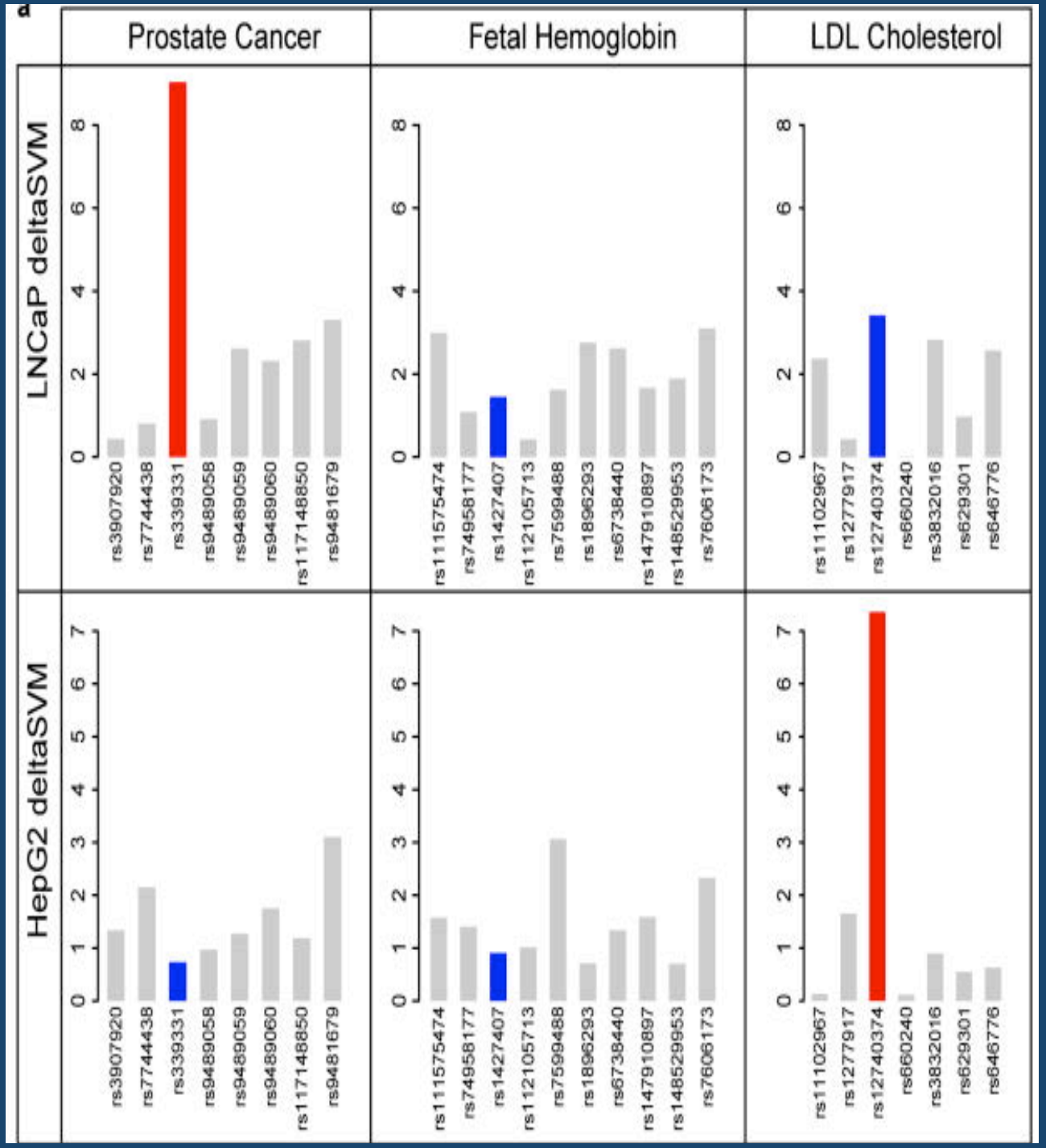
1. Classifier
2. SVM protocol
3. Some results

- Takes into account tissue-specificity

delta SVM values for 3 experimentally validated SNPs

human prostate adenocarcinoma

Hepatocytes (main cell in liver)



Key Points

- GWAS variants are enriched for functional information in a tissue-specific manner
- Machine learning can be used to find patterns in functional data to identify novel variants associated with disease