Expression QTLs (II)

BBMS 3009: Genome Science (First Semester, 2021)

Dr. Yuanhua Huang
School of Biomedical Sciences &
Department of Statistics and Actuarial Science

香 港 大 學
THE UNIVERSITY OF HONG KONG

Today’s learning objectives

- Review the methods for predicting eQTLs and allelic specific expression
- Understand the principles in machine learning to prioritise variants
- Discuss the use of gene regulatory network
- Understand the principles in systems genomics

Recall

Alternative model (Likelihood L_1) $y = \beta_0 + x_1\beta_1 + x_2\beta_2$

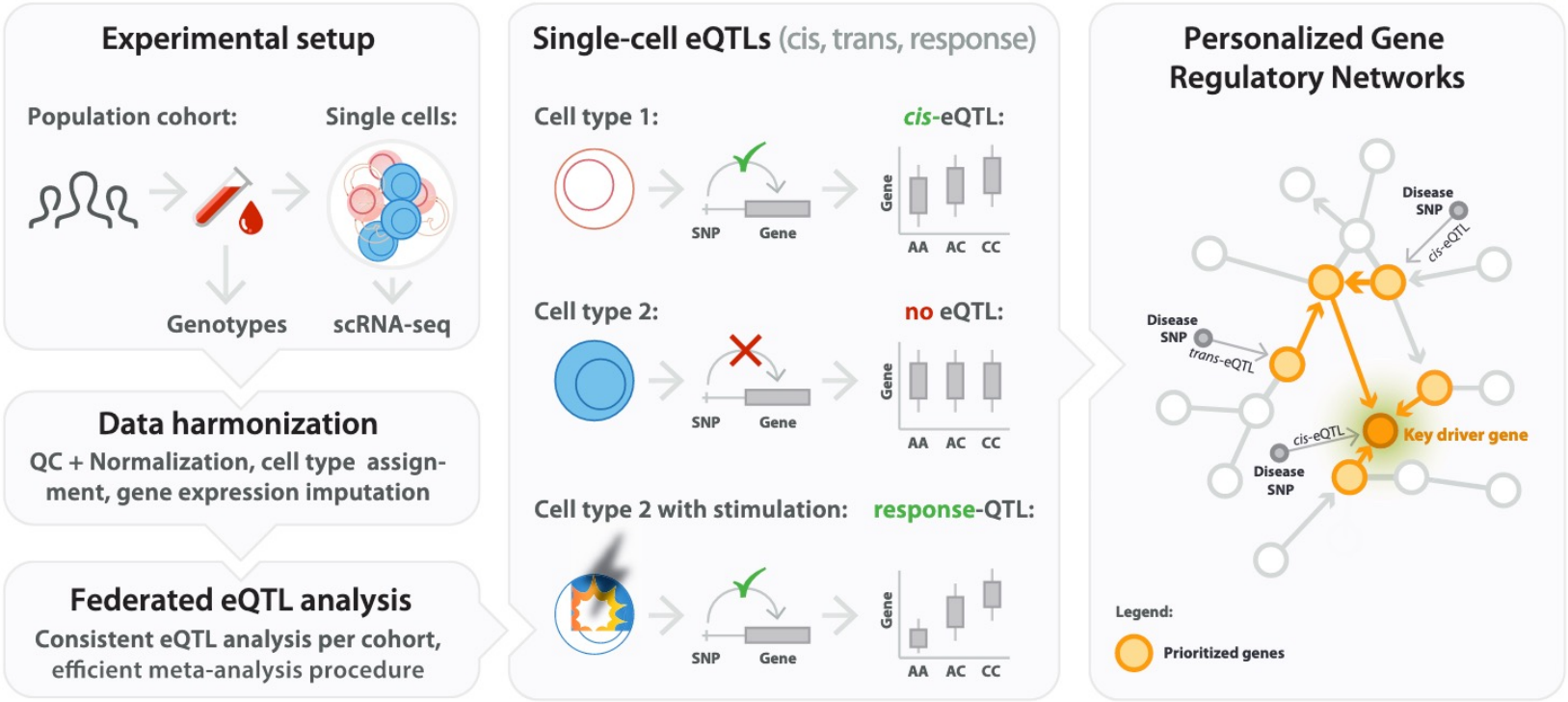
Null model (Likelihood L_0) $y = \beta_0 + x_1\beta_1$

- Differential gene expression
  - x_2 is condition, e.g., treated or untreated
  - Usually small sample size; around 3 replicates
  - Generalized linear model – negative binomial for raw counts
- Expression QTLs (eQTL)
  - x_2 is genotype value 0, 1, 2, from cis- or trans- SNPs
  - Usually, hundreds of samples
  - log(RPKM + small_value) or log(TPM + small_value); Gaussian-like

mean	Size factor	variable
x_0	x_1	x_2
1	0.6	0
1	1.3	0
1	0.9	1
1	1.1	1

- **Hypothesis testing:**
  - Likelihood ratio test: likelihood ratio between two models
  - Wald test: mean and variance of β_2 in alternative model

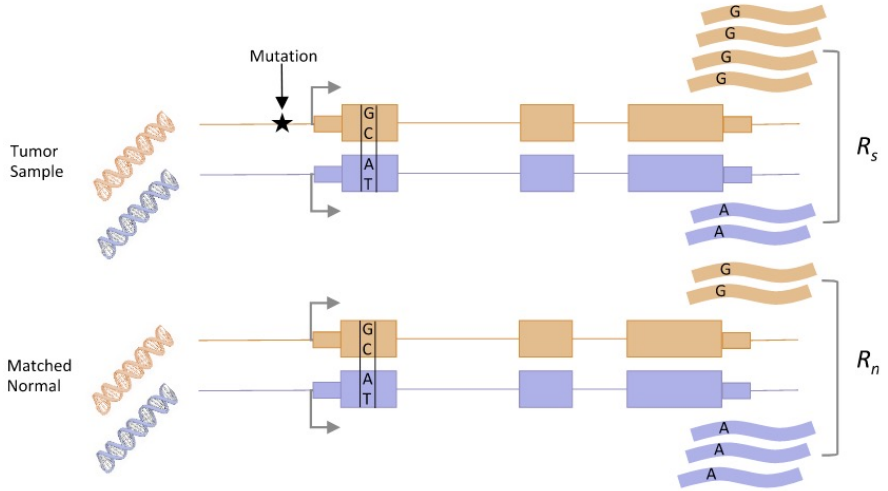
Single-cell eQTLs



van der Wijst, et al. elife, 2020

Allelic specific expression (ASE)

- Imbalance of the expression of two alleles
- One example reason: mutation in the promoter in one chromosome → affect factor binding → increase or reduce the expression → ASE
- Avoid mapping bias: use the right reference, possibly by masking the SNPs

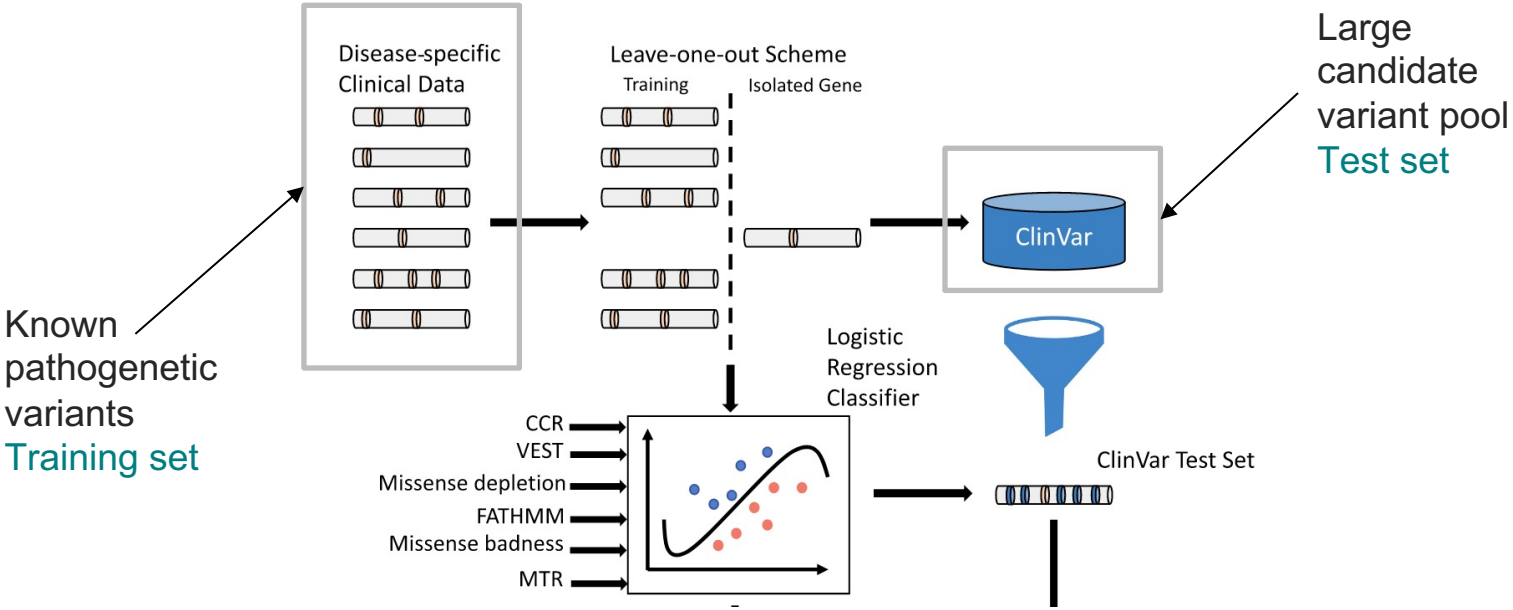


Castel et al., Genome Biology, 2015
Przytycki & Singh, 2020, Cell Systems10, 193–203

Think: relation between ASE and eQTLs

Predict pathogenic variants

- How can we predict pathogenic variants by the epigenomic markers and / or other genomic features?

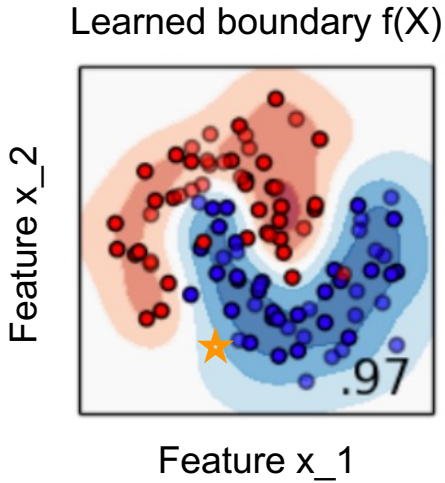


Evans et al, Genome Research, 2019

Machine learning methods

- Classification problem

- Training data (variants: positive & negative)
  - We have the pathogenic label: y
  - We have the feature vector: X



- Fitting machine learning model: $f(X)$
  - For example, by minimizing errors in training set $\sum_{i=1}^{N}(f(X_i) - y_i)^2$

- Predicting the pathogenic state of new variant $f(X_{new})$

Machine learning methods

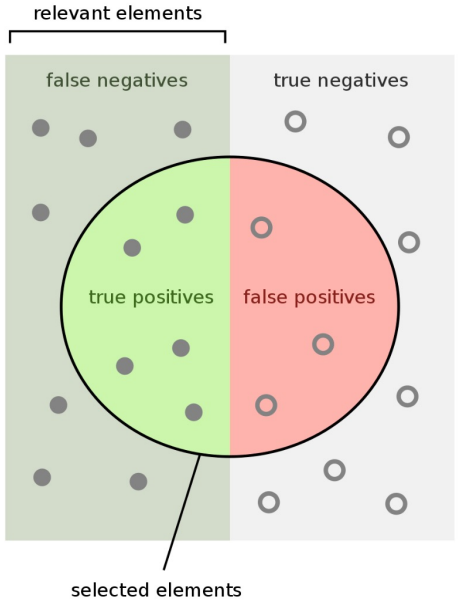
- Predictive features (need exploration and prior knowledge)
  - Epigenetic marks
  - Functional annotations
  - Sequence motifs (or factor binding peaks)
  - Conservation across species
- Models
  - Logistic regression
  - Artificial neural network (deep learning)
  - Support vector machines
  - Random forest (decision tree based)
  - Many more, incl. ensemble or stack models

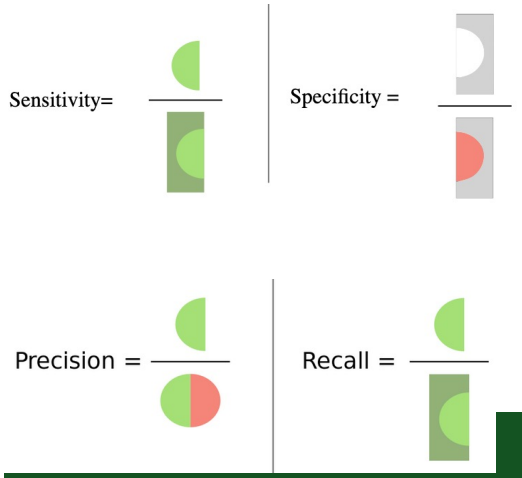
Different models may have advantages to a certain types of data distributions

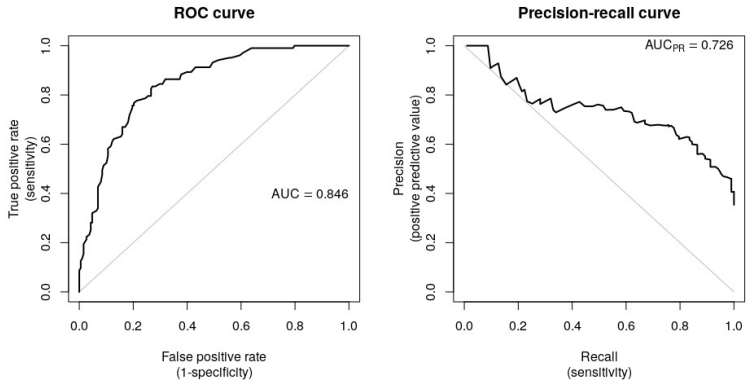
https://en.wikipedia.org/wiki/Precision_and_recall
https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Performance evaluation

- Cross-validation: 10-fold (assessing generalization)
  - Training set, (validation set), test set
- Sensitivity vs Specificity; Precision vs recall
  - Receiver Operating Characteristic (ROC) curve; Precision recall curve (PRC)
- Avoid systematic bias: check balance of positive vs negative samples

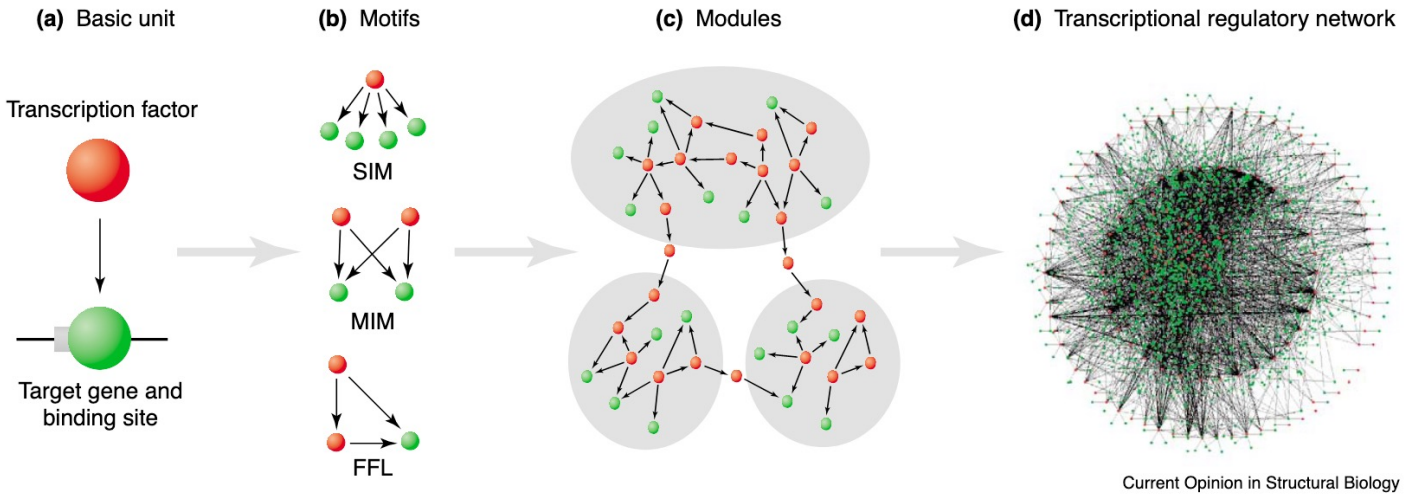






Revisit gene regulatory network (GRN)

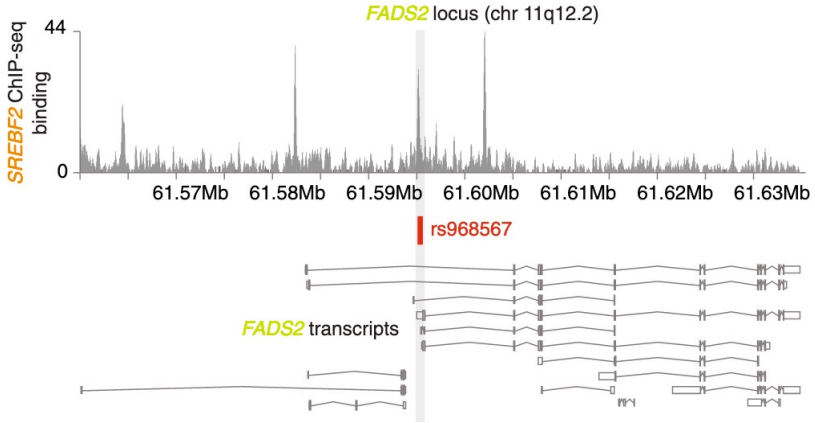
- Recall: discussion of GRN on the motivation of transcriptomics
  - Systematic way to view the transcriptome instead of single genes
- Transcription factor
  - Binding to promoter region of other genes
  - Turn on or off the target gene: the right cell, right time, right amount
  - Estimated up to 1600 TFs in the human genome

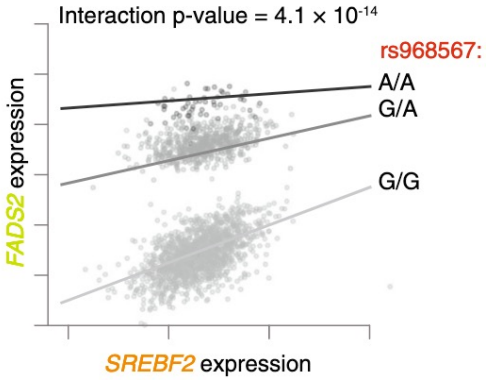


Babu et al. 2004
Vaquerizas et al, 2009

Gene regulatory network with genetics

- Genetic variants interacting with transcriptional regulation
  - Genetic variants may affect the TF bindings
  - Potential mechanism of eQTLs and Allele specific expression
- CRISPR perturbations can be used to test the regulatory roles, e.g., Perturb-seq, CROP-seq



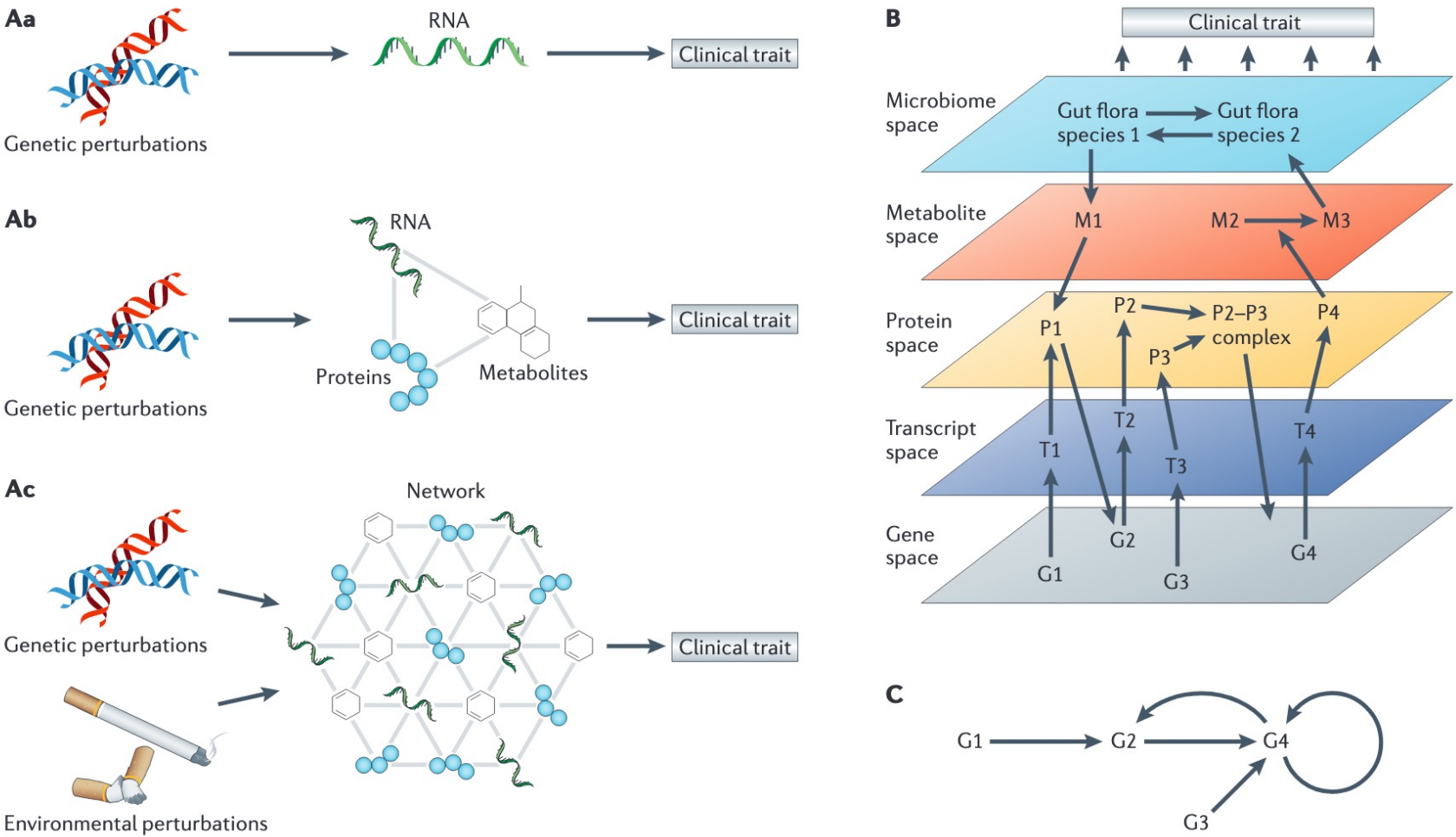


van der Wijst, et al. Genome Med, 2018

Systems genomics

- Why do we need Systems genomics?
- Conventional genetic analysis approaches failed to
  - Biologically interpret many statistical significances
  - Find out all disease susceptibility DNA sequence variations
  - Obtain a full picture of the development of human diseases

- Assays for multiple molecular layers
  - DNA sequence → Genomic data
  - Epigenetic markers → Epigenomic data
  - Gene expression → Transcriptomic data
  - Proteins → Proteomic data
  - Metabolites → Metabolomic data
  - Microbes → Microbiome data



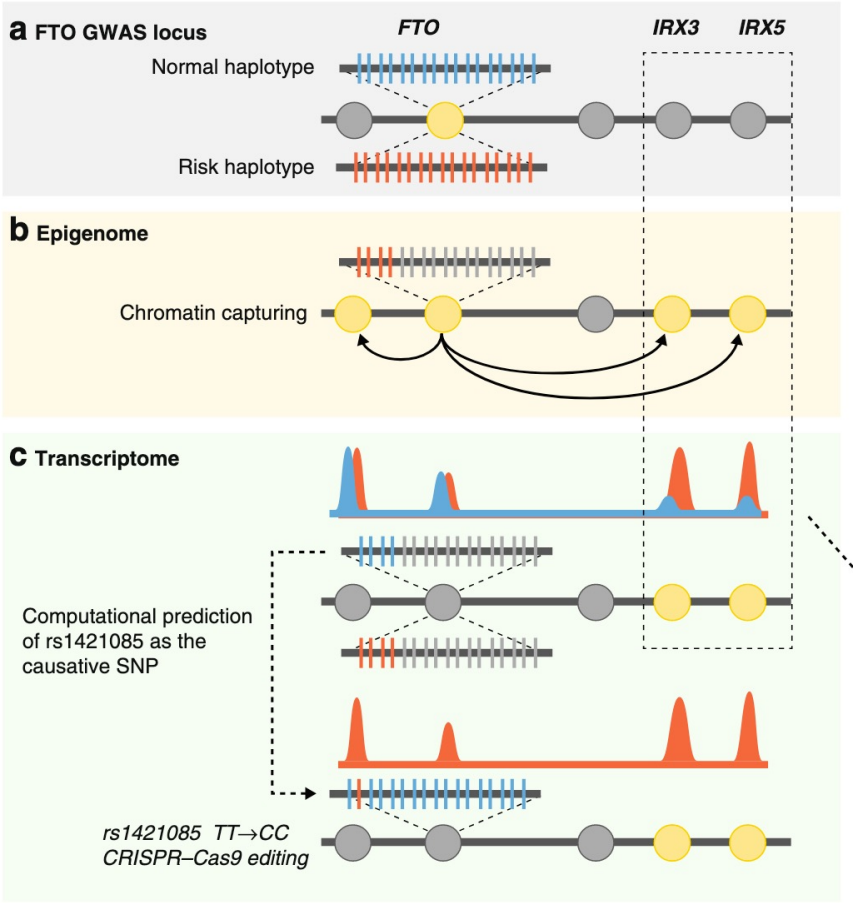
Civelek & Lusis (2014) Nat Rev Genet.

Systems genomics: example

- The *FTO* region harbours the strongest genetic association with obesity
- Unclear mechanistic basis

Approaches

- GWAS → associated SNPs
- Epigenetics & chromatin capturing (Hi-C) → genes interaction
- Transcriptome → eQTL; further prediction of causative SNP
- CRISPR/Cas9 validation



Claussnitzer, et al. N Engl. J Med. 2015
Hasin, et al. Genome Biology 2017

Questions?

- Review the methods for predicting eQTLs and allelic specific expression
- Understand the principles in machine learning to prioritise variants
- Discuss the use of gene regulatory network
- Understand the principles in systems genomics

Relevant reading

- Albert and Kruglyak. The role of regulatory variation in complex traits and disease, Nat Rev Gen, 2015