Post GWAS: integration of OMICS data via QTL studies

Amna Khamis

amna.khamis@cnrs.fr

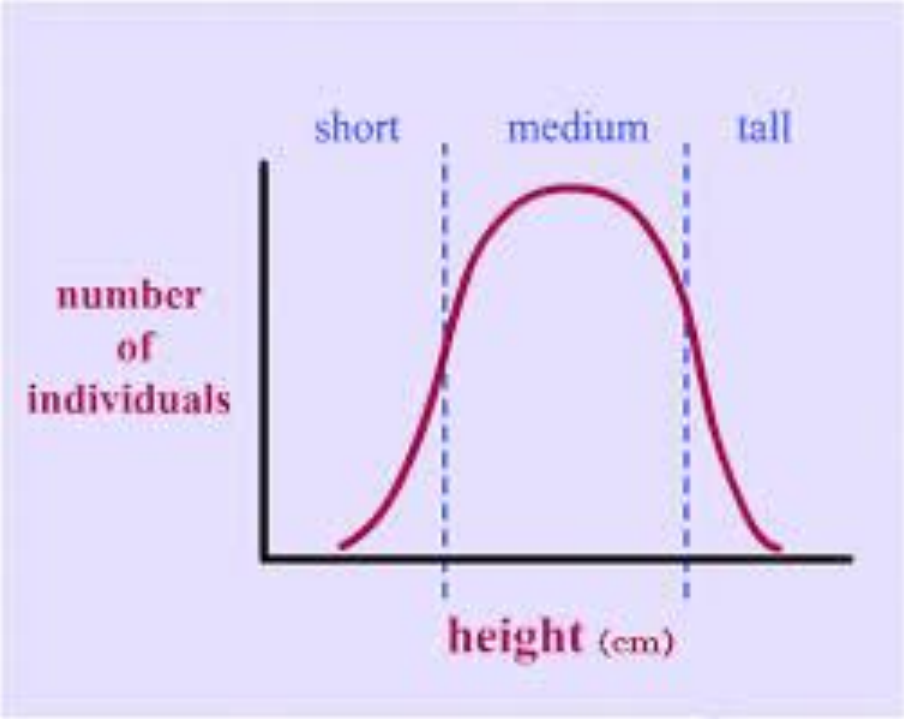
Today’s lecture

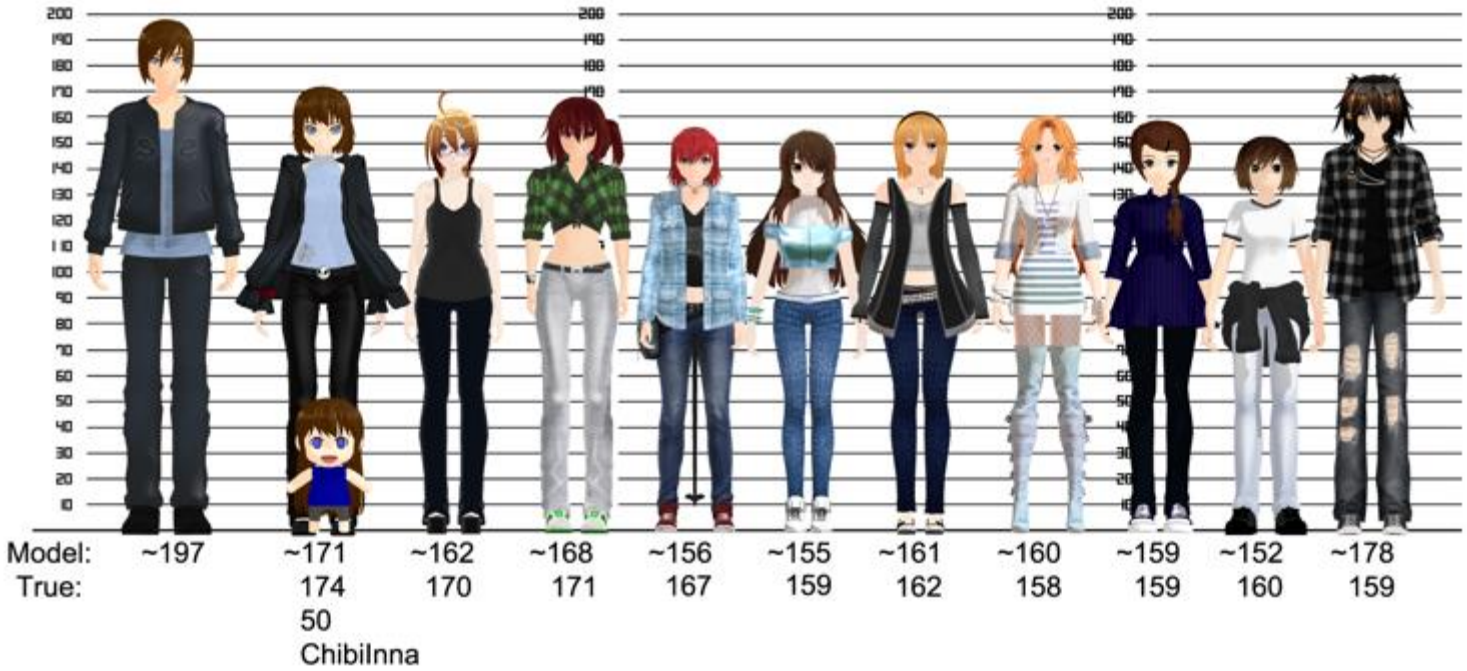
- Expression quantitative trait loci
- Combining –omics with genotyping
  - Epigenetics
  - Types of RNA + genomics
  - Protein quantitative trait loci
  - Metabolite quantitative trait loci

Quantitative trait loci

- A locus is a region within the genome
- A phenotypic trait that can be measured quantitatively
- Attributed to the additive effect of many genes + environment

Quantitative traits – example Height

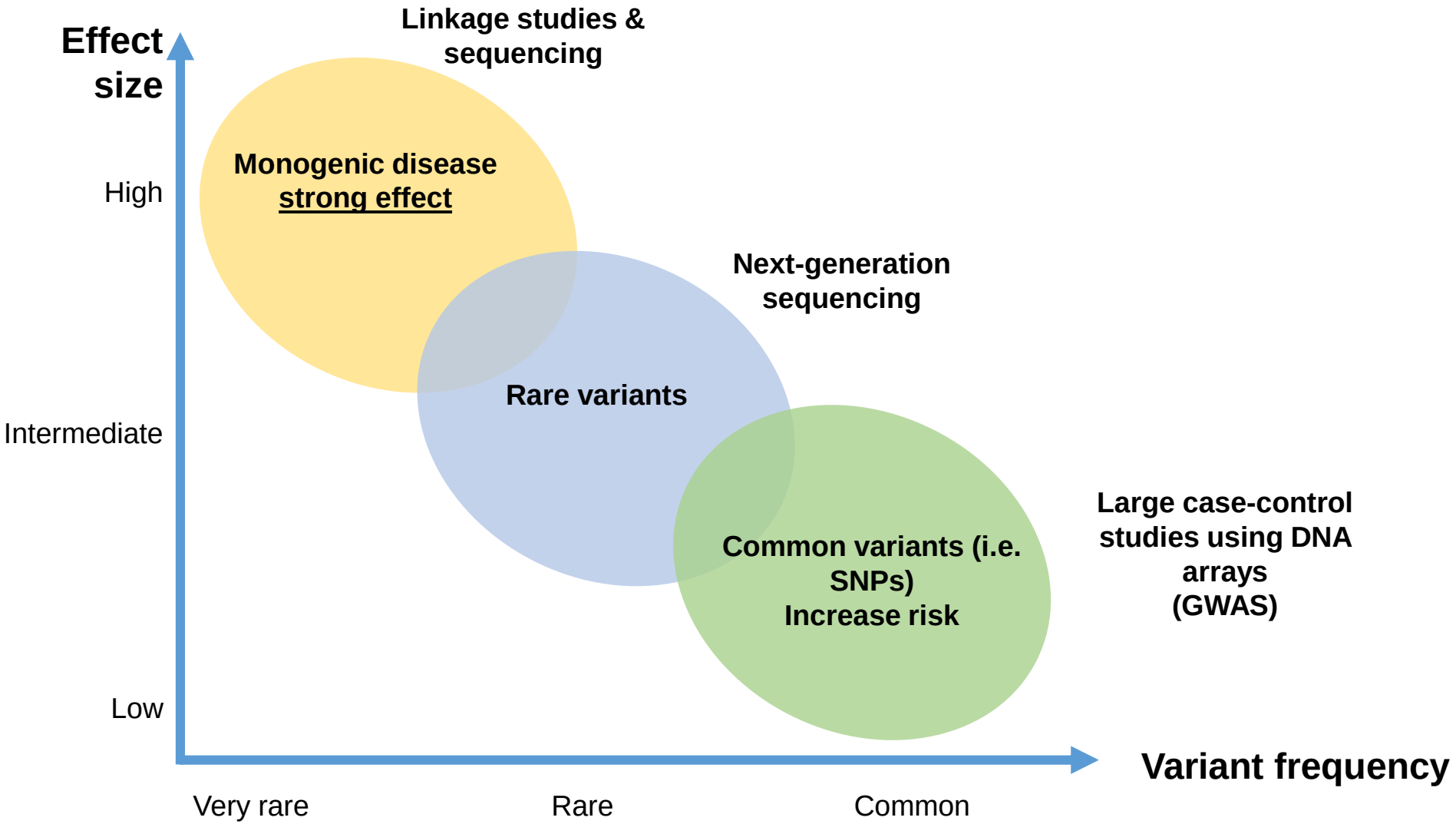




Mutation vs SNPs

Mutation	**Single Nucleotide Polymorphism**
Change of one nucleotide or more	A single nucleotide
Frequency is rare	Common in a population
A cause of **monogenic** disease	A cause of **polygenic** disease
Sickle cell anaemia, Klinefelter's disease etc.	Type 2 diabetes, height, weight

Genetic architecture of disease



Genome Wide Association Studies (GWAS) Overview

Population
Cases vs controls

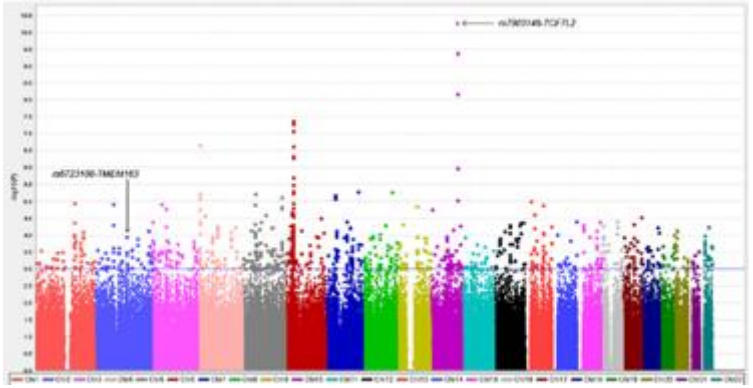
DNA Genotyping

Genes associated
with disease/trait

Drug/biomarker
discovery

Cases

SNP Arrays



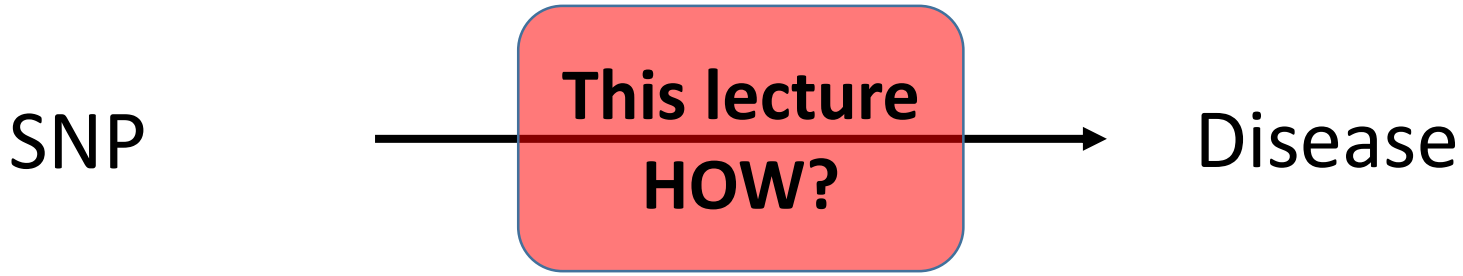
Cardiovascular disease,
type 2 diabetes, height,
weight, cancer…

Disease/biology understanding?

Controls

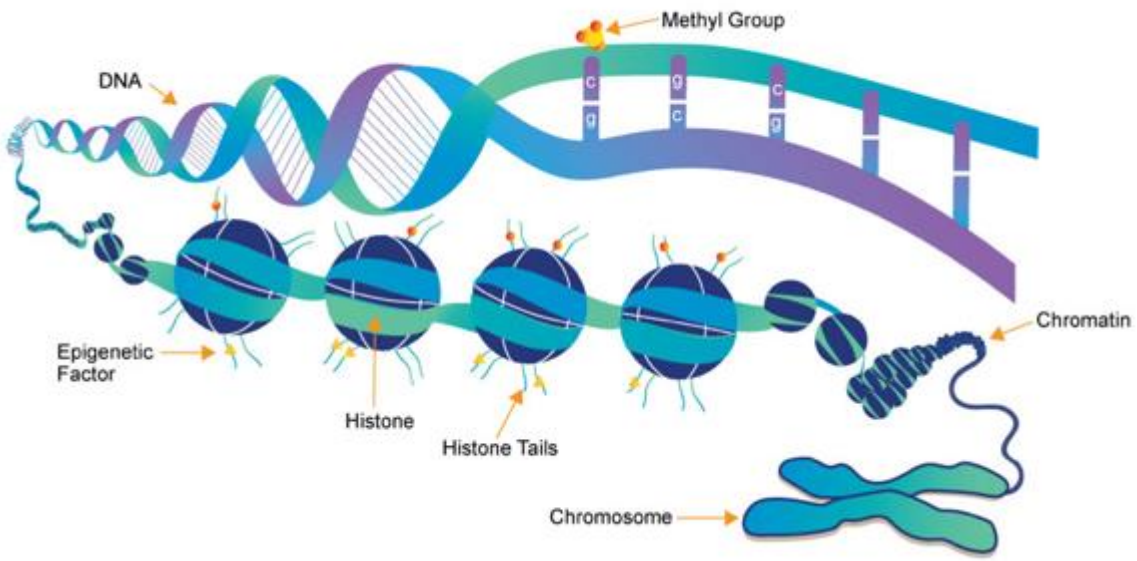
Genome wide association studies

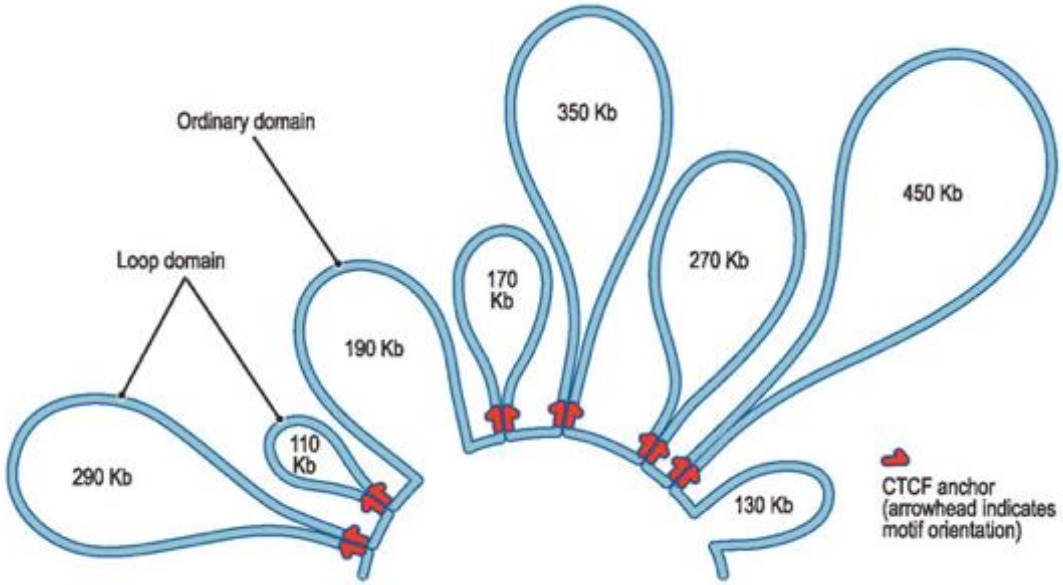
- > 10,000 SNPs associated with disease (e.g. cardiovascular disease, cancer, type 2 diabetes, BMI *etc.*)
- Imputation – increase power to detect these *loci*
- Most in non-coding regions
- Interpret effects of genetic variants in complex traits



Genome Wide Association Studies (GWAS) Overview

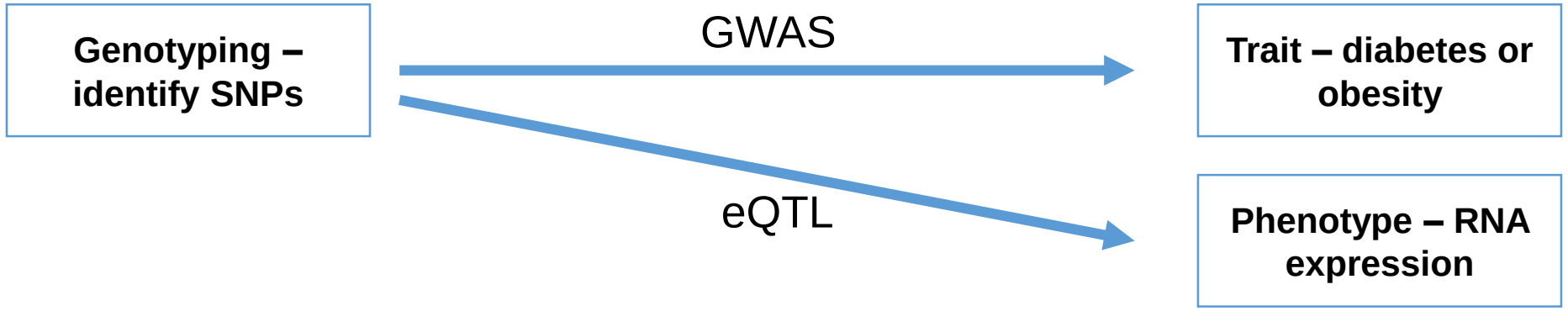
- GWAS loci implicate nearest gene associated with disease
- Deciphering the causal variant and making inferences from GWAS to physiology is still a challenge
- Few GWAS SNPs are near biological candidate genes
- Majority of SNPs lie within non-coding regions of the genome
- Genes are not linear - chromatin looping facilitates TF binding





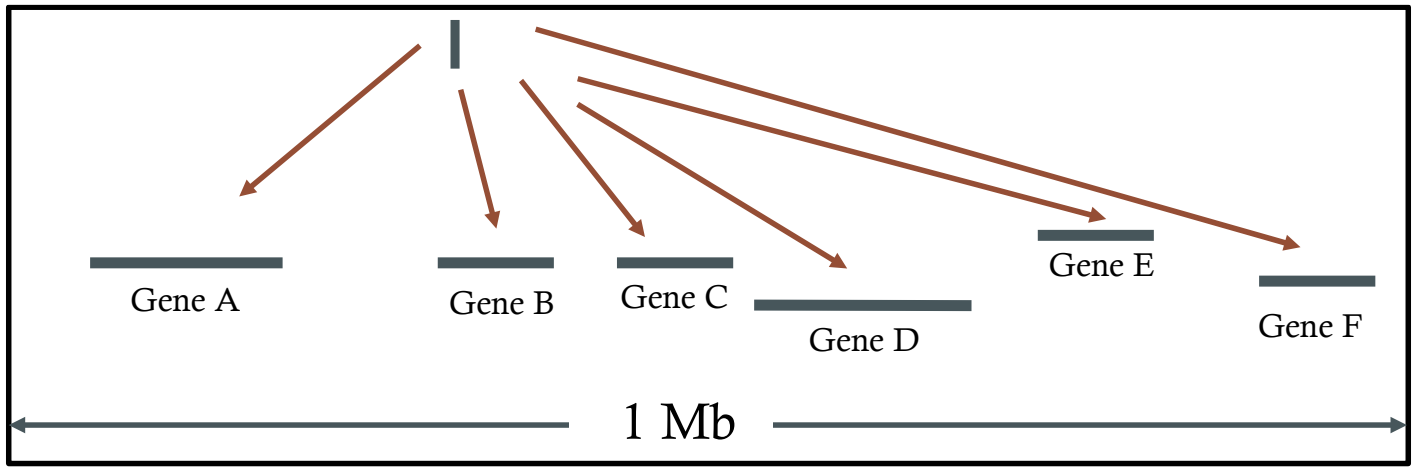
Expression Quantitative Trait Loci (eQTL)

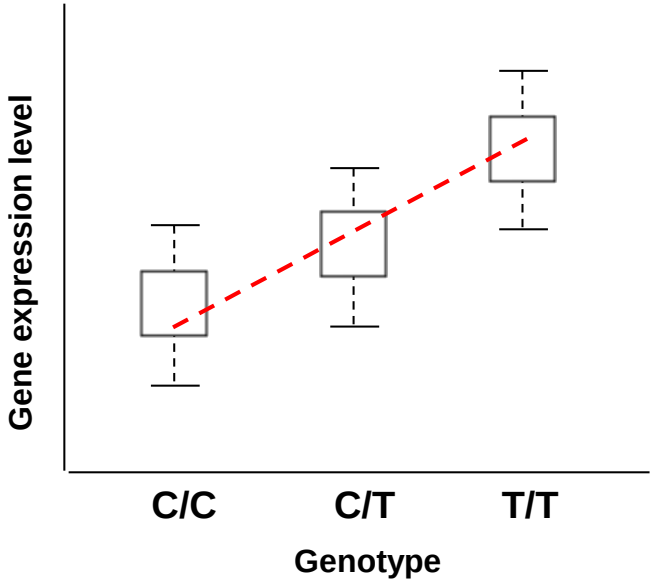
- Quantitative trait *loci*: a region of DNA associated with any measurable trait. i.e. BMI, disease
- GWAS: Trait is disease or measurable trait (e.g. BMI)
- eQTL: Trait is RNA expression level
  - Is a change in expression level at a particular gene driven by genotype?
  - Method used to speculate and investigate what these effects might do



Expression Quantitative Trait Loci (eQTL) cont.

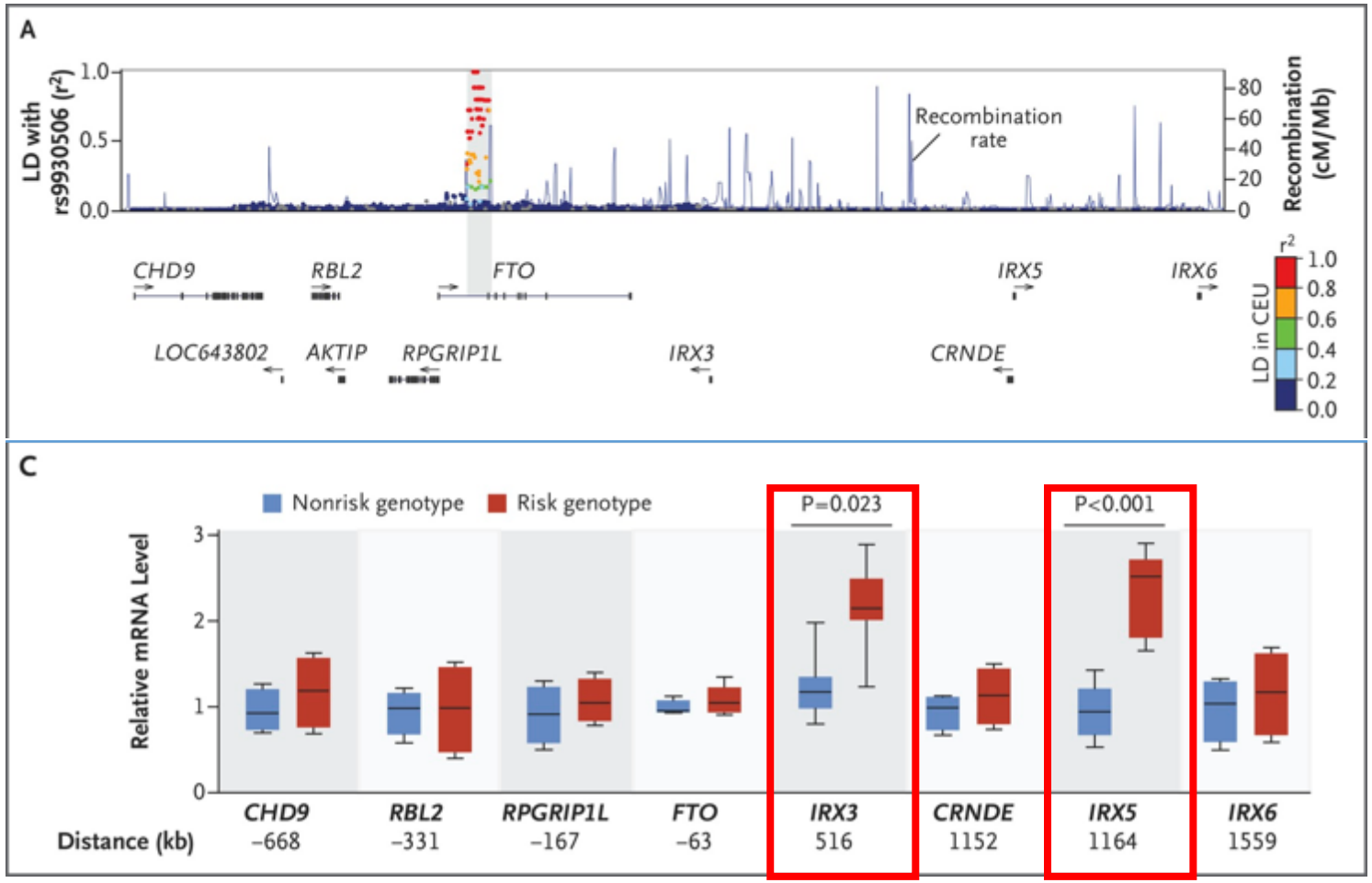
- Genotype-expression association
- eQTL data can open up new biology through reverse genetic approaches
- Aim: identify genomic locations where genotype significantly affects gene expression





GWAS gene FTO associated with BMI & obesity

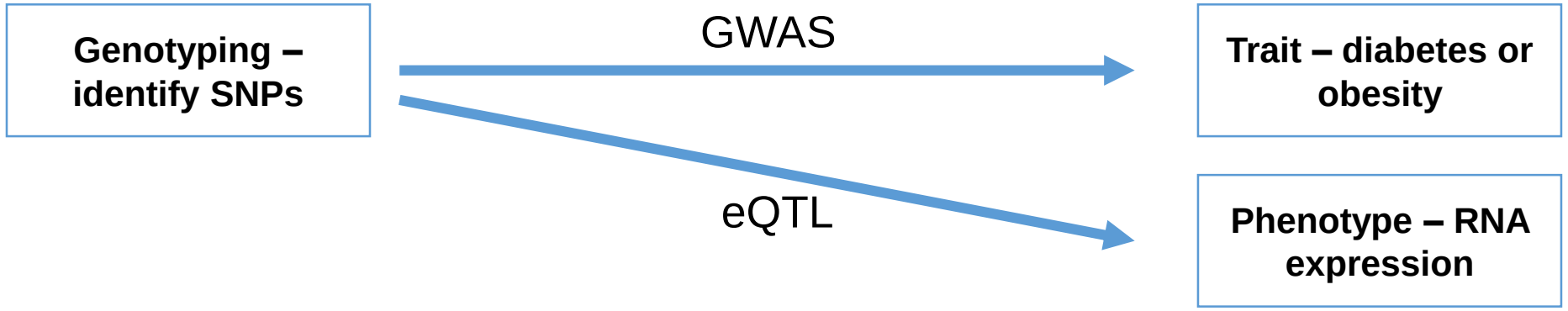
- GWAS - associated with BMI and obesity risk
- Implicate the nearest gene – FTO (gene functions in mRNA demethylase)
- Null mice – no obesity of thinness
- So, is FTO the implicated gene?



Claussnitzer et al *NEJM* 2015

Expression Quantitative Trait Loci (eQTL)

- Quantitative trait *loci*: a region of DNA associated with any measurable trait. i.e. BMI, disease
- GWAS: Trait is disease or measurable trait (e.g. BMI)
- eQTL: Trait is RNA expression level
  - Is a change in expression level at a particular gene driven by genotype?
  - Method used to speculate and investigate what these effects might do



Methods to identify common variants: genomics

- An organism's complete set of DNA is called its genome
- **Genomics** is the study of all of an individual’s genes
- Sequencing simply means determining the exact order of the bases in a strand of DNA.
- All DNA in an organism are the same, therefore we can use DNA isolated from any tissue
- At the gene level: PCR, Sanger sequencing
- At the genome level: NGS - DNA sequencing methods

Methods to identify gene expression: Transcriptomics

- **RNA microarrays**: Slide or membrane with numerous probes that represent various genes of some biological species.
  - The array type corresponds to a list of reference genes on the microarray with annotations.
  - Example: Affymetrix
- **Sequencing technique** which uses next-generation sequencing to quantity of RNA
  - RNA-sequencing is the gold standard to other technologies, such as microarray because:
  - Genome wide
  - Low background signal – mapped to the genome, therefore no issues with cross-hybridization
  - More quantifiable - Issues with microarray data in extremely high or low expression levels

Expression quantitative trait loci

- eQTL are used to interpret GWAS hits, e.g. to narrow candidates
- eQTLs can be used to identify novel genes associated with a particular trait
- If we could quantify the amount of RNA in a particular tissue or cell type, under a specific set of conditions, this might be informative (i.e. a proxy for gene expression)
- A case where different allelic states at a specific site (locus) in the genome alter a measured expression variable in a tissue / cell population under a given a set of conditions is an eQTL
- **eQTL therefore describes a variable pair (genotype-expression association)**

Important factors for eQTL discovery

1: Population
i.e. is an eQTL in Europeans informative of mechanism underlying disease risk for a disease found in African?

2: Technology
i.e. Array and sequencing technologies

3: Cell type
i.e. is an eQTL in blood informative of mechanism underlying disease risk for a disease based in adipocytes?

Data requirements for an eQTL study

- Individuals or samples
- Genotyping data
- RNA expression/transcriptomics data

Overview of eQTL method

1. Subjects – Homogeneous population
2. DNA Genomics
3. RNA transcriptomics
4. Combine genomics + transcriptomics = eQTL analysis

Example of eQTL study to illustrate methodology

Original Article

MOLECULAR METABOLISM

Laser capture microdissection of human pancreatic islets reveals novel eQTLs associated with type 2 diabetes

Amna Khamis [1,2,11], Mickaël Canouil [2,11], Afshan Siddiq [1], Hutokshi Crouch [1], Mario Falchi [1], Manon von Bulow [3], Florian Ehehalt [4,5,6], Lorella Marselli [7], Marius Distler [4,5,6], Daniela Richter [5,6], Jürgen Weitz [4,5,6], Krister Bokvist [8], Ioannis Xenarios [9], Bernard Thorens [10], Anke M. Schulte [3], Mark Ibberson [9], Amelie Bonnefond [2], Piero Marchetti [7], Michele Solimena [5,6], Philippe Froguel [1,2,*]

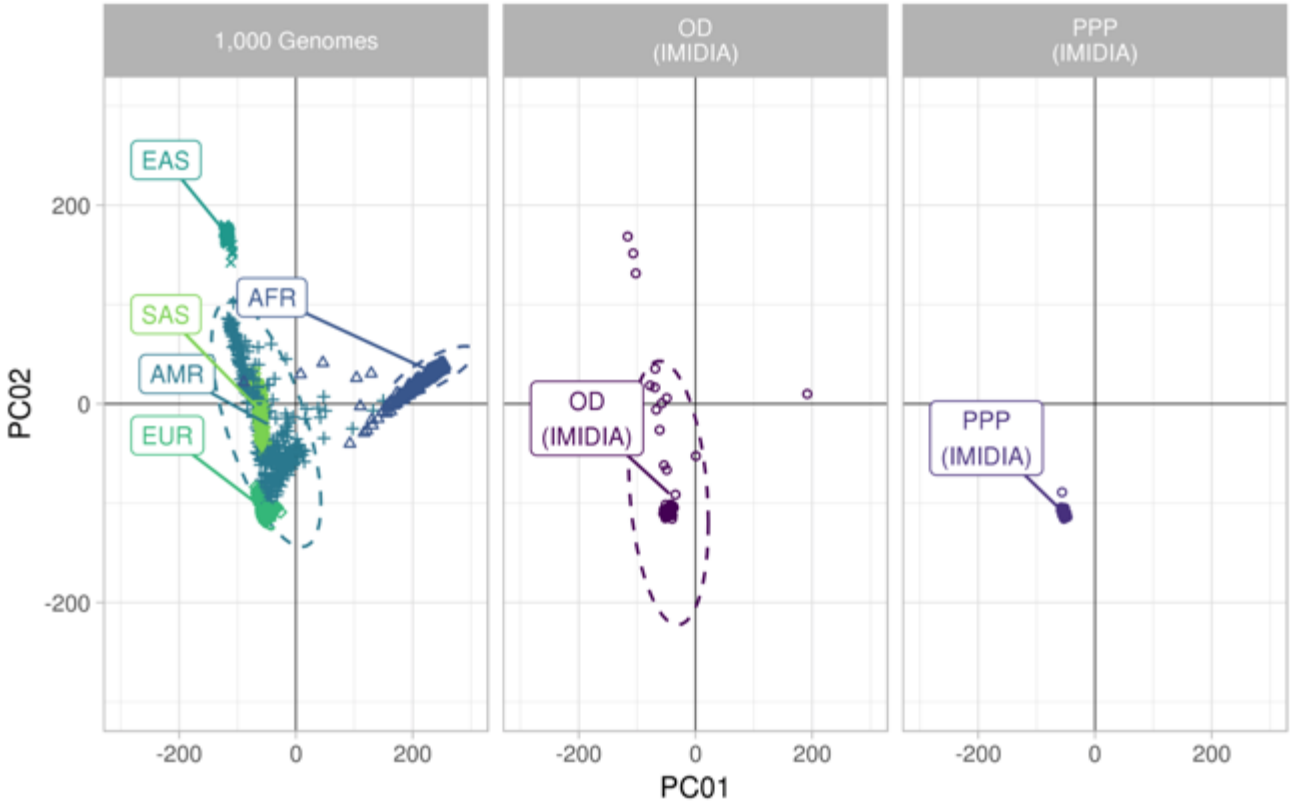
Type 2 diabetes

- Type 2 diabetes (T2D) is a metabolic disorder that is characterised by high blood glucose due to insulin resistance or deficient beta-cell function
- GWAS results so far:
  - >300 *loci* associated with type 2 diabetes and associated traits
  - Deciphering the causal variant and making inferences from GWAS to physiology is still a challenge
- Aim: identify eQTLs in the samples collected from pancreatic islets

1 Subjects – Homogeneous population

- Identify population outliers

- Two populations:
- **OD:** Organ donors from Italy
- **PPP:** Pancreatic pancreatectomy patients from Germany



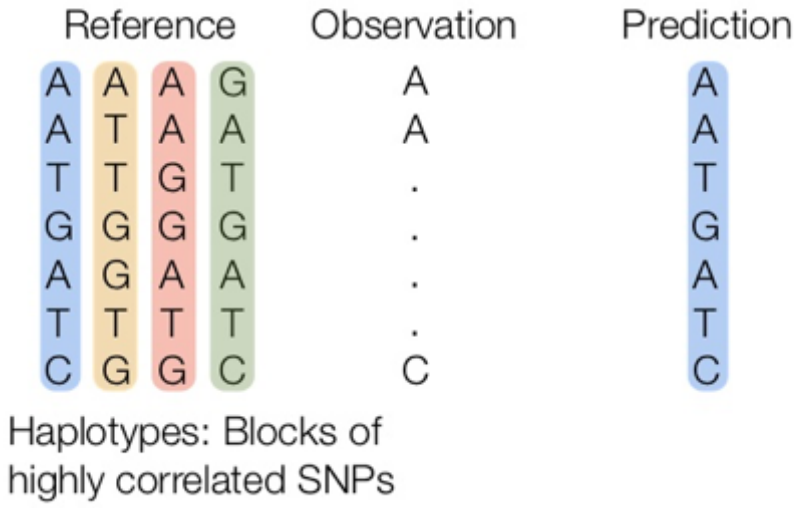
2

DNA Genomics

- DNA: Genotyping of blood - analysed total of > 8M SNPs

2.5M Omniarray Beadchip - Illumina
1,233,520 SNPs

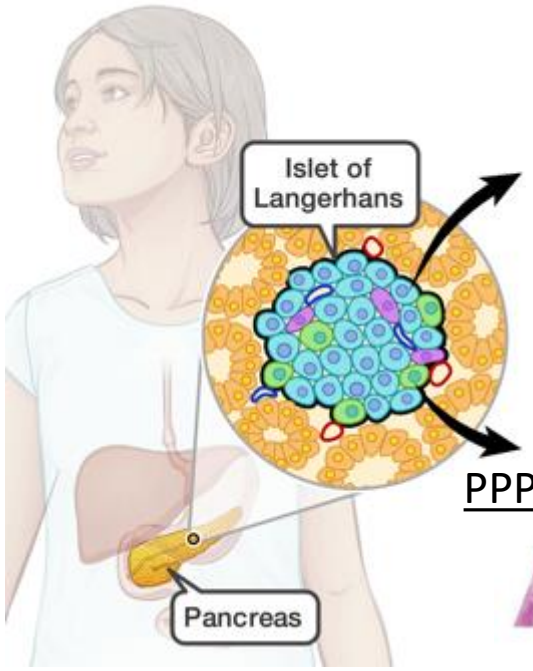
Imputation: the statistical prediction of unobserved genotypes
7,574,416 SNPs



Imputation methods take advantage of linkage disequilibrium
Individuals that are identical at a subset of genetic variants will likely be identical in between those variants

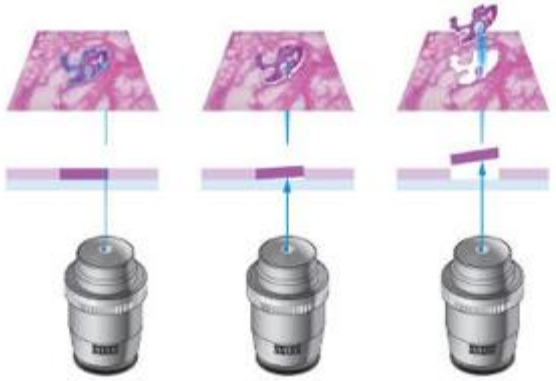
3 RNA transcriptomics

- RNA: Whole RNA analysis of pancreatic islets using two different approaches



OD: Enzymatic digestion
Standard protocol for collecting islets

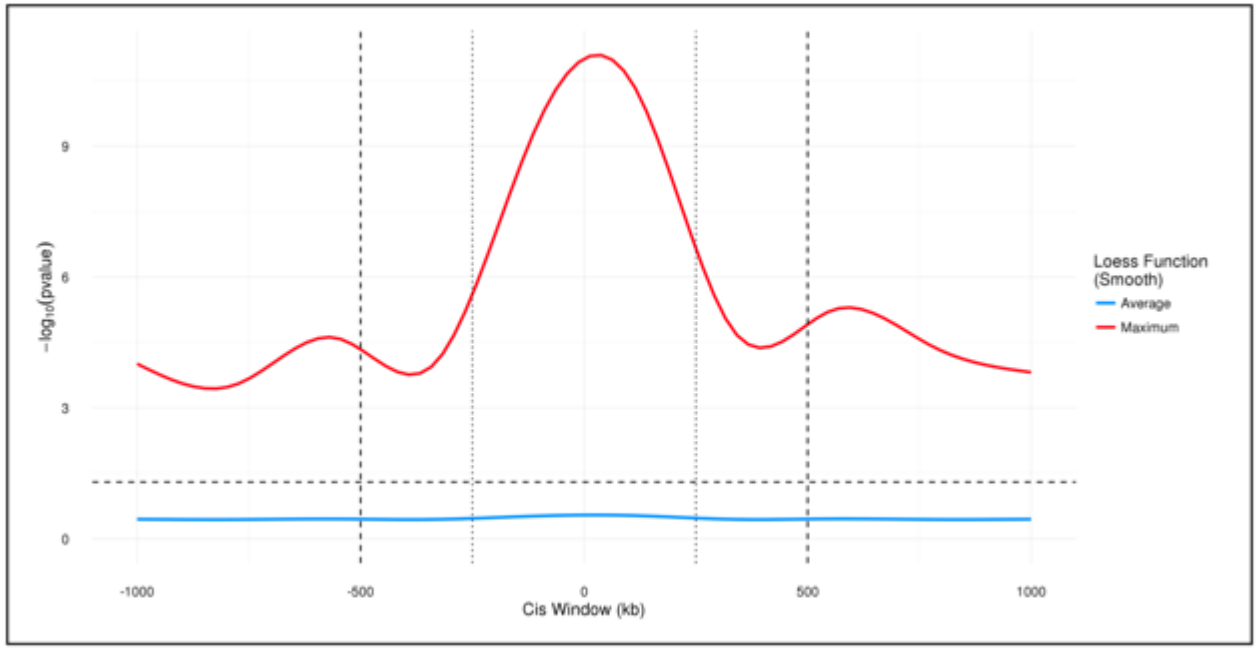
PPP: Laser Capture Microdissection



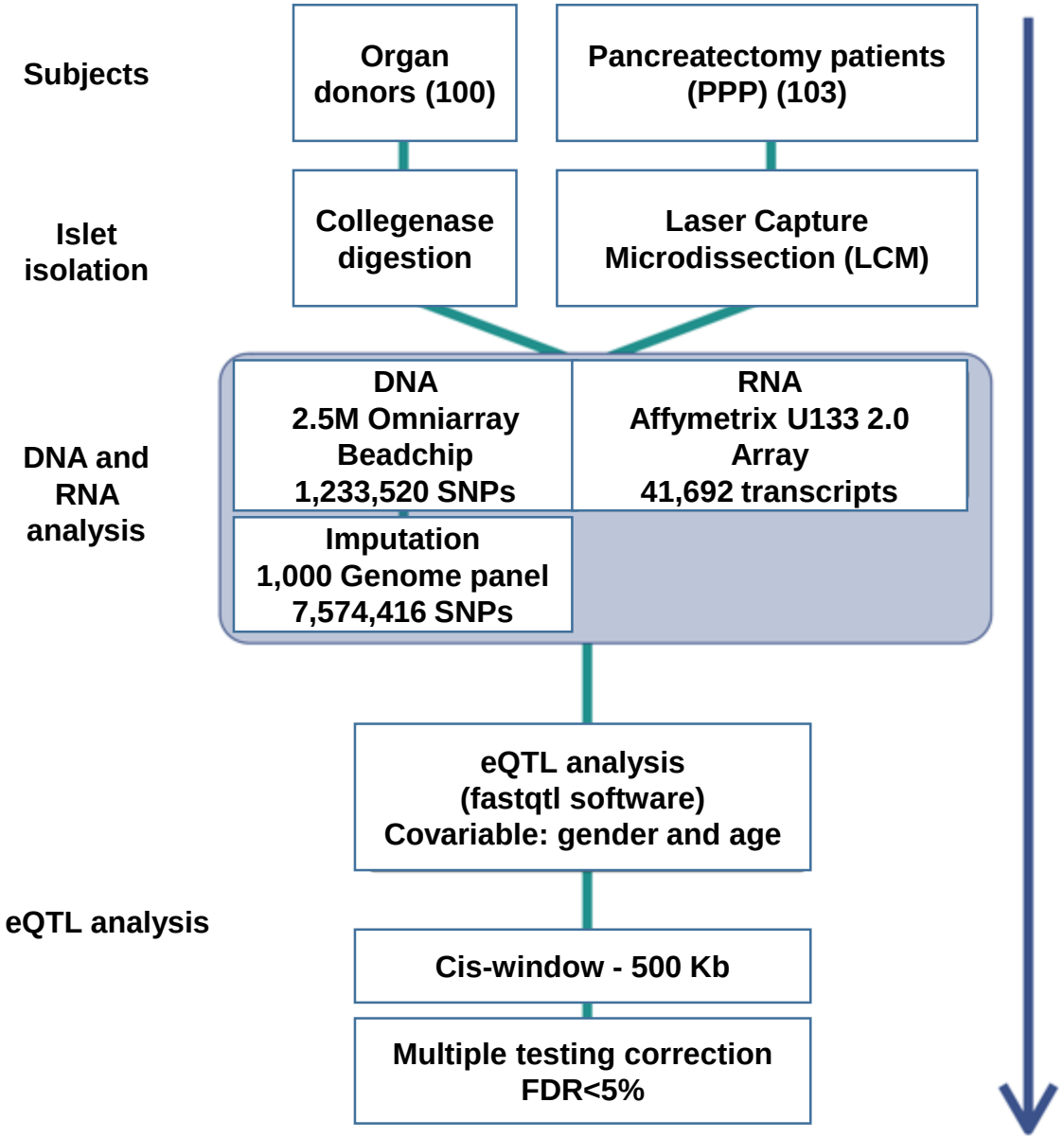
Affymetrix U133 2.0 Array
41,692 transcripts

4 Combine genomics + transcriptomics = eQTL analysis

- Cis-eQTL study
- Single SNP additive model and was to run cis eQTL analysis within a window of 500 Kb around each transcript (maximum distance at which gene-SNP pair is considered local)
- Correct for technical and biological biases. i.e. population, sex, age
- Multiple testing: Bonferroni correction – 5% considered significant



Summary of method

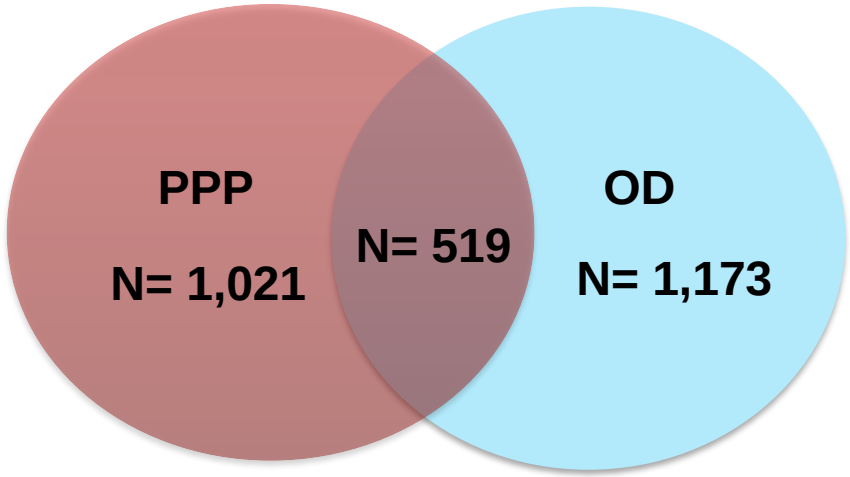


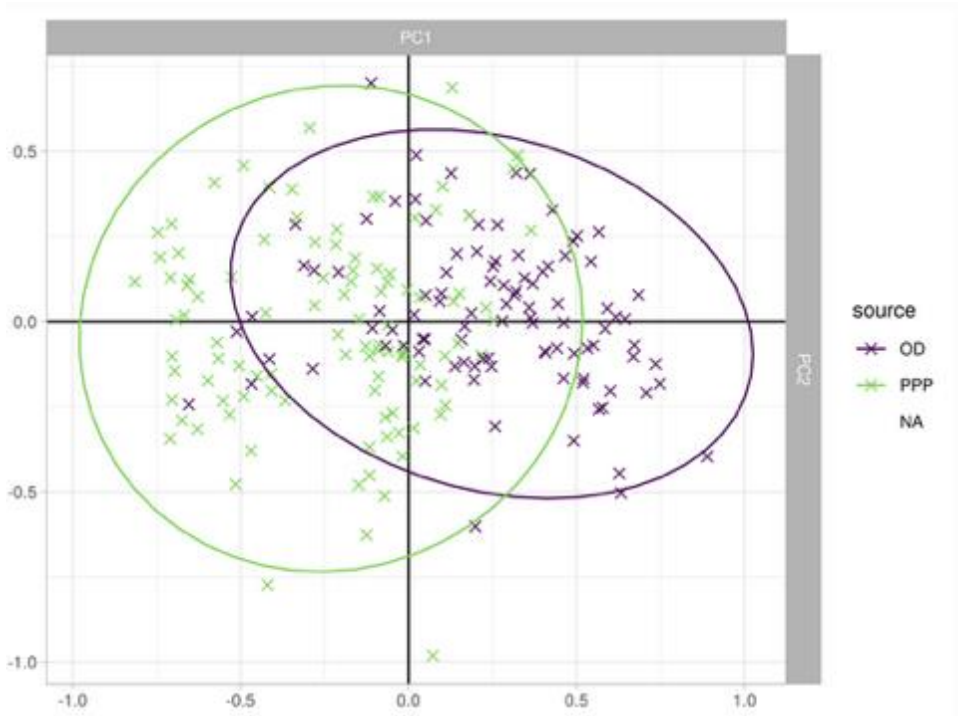
Results

Summary of identified eQTLs

- Genotype-expression pairs
- Differences between two cohorts highlight the importance of tissue/cell type & methods of extraction

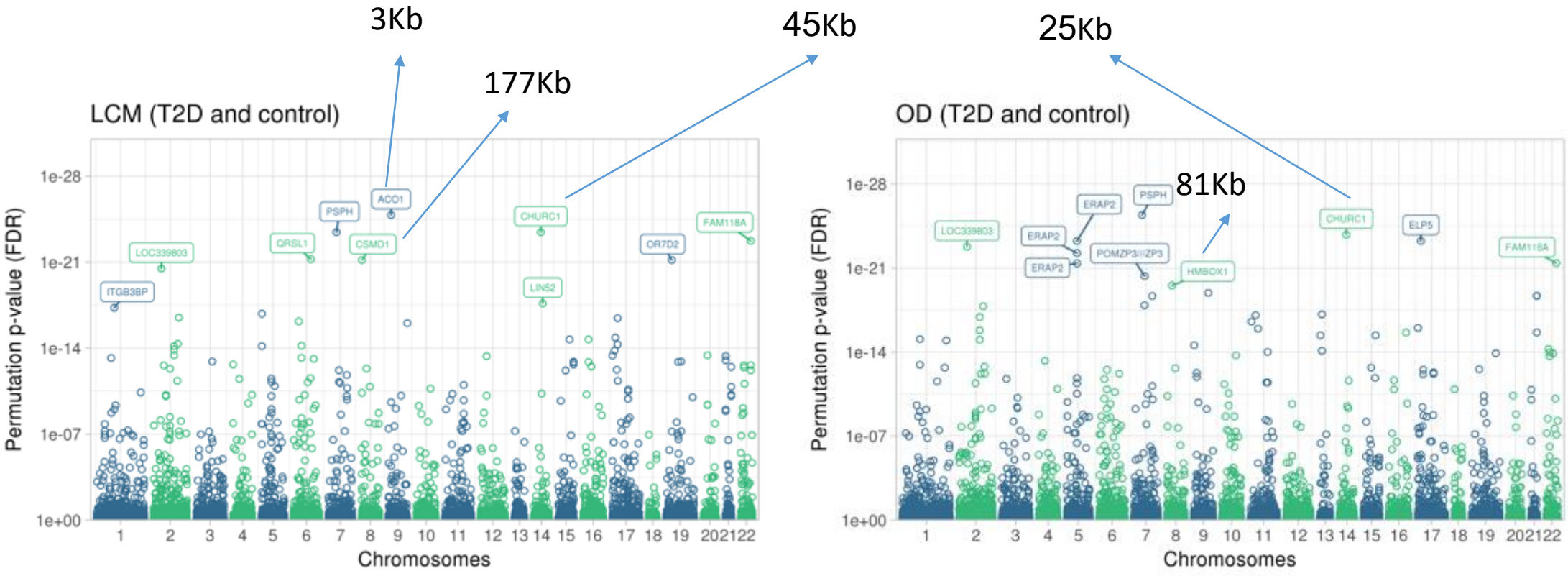
Shared genes



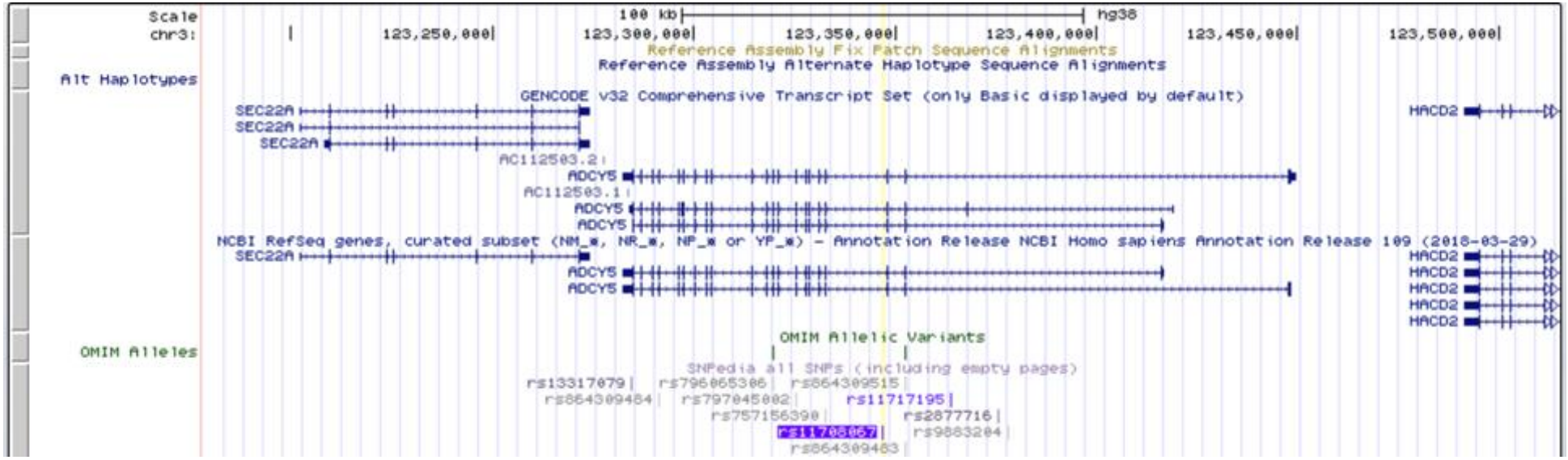


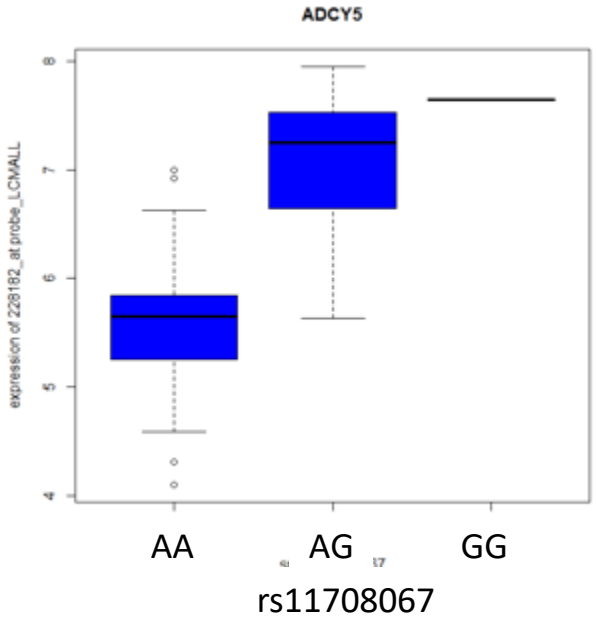
Most significant eQTLs in OD and PPP

- Cis-eQTLs – coincides with the location of the underlying gene (within 500 kb)



Example: ADCY5 – member of adenylate cyclase family





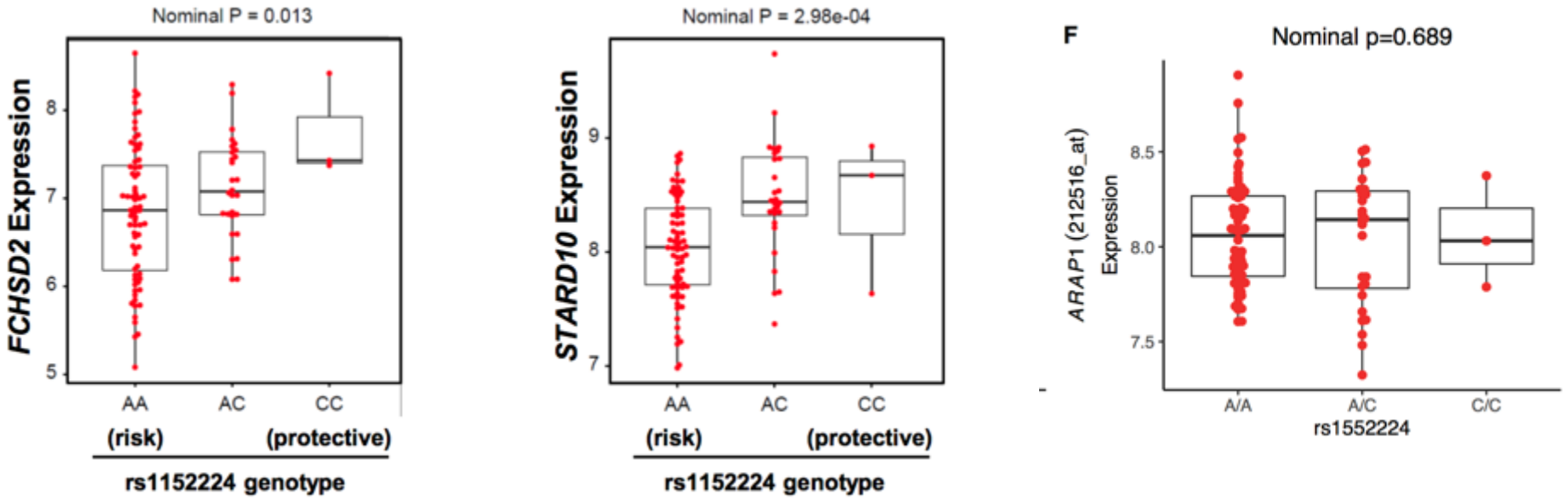
- rs11708067 A: Associated with T2D, fasting glucose and 2-hour glucose
- Previous report, from a small candidate gene study Hodson et al 2014, of a negative correlation between risk allele count and *ADCY5* expression levels.
- In human islets, ADCY5 is thought to couple glucose stimulation to insulin secretion, and this coupling is disrupted upon gene knockdown.

GWAS T2D and associated trait loci that co-localise with eQTL

GWAS Catalog						
			PPP			
GWAS gene	**rs ID**	**eQTL SNP**	**eQTL distance**	**eQTL FDR**	**eQTL beta**	**eQTL gene**
FREM3	rs13134327	rs5015757	-174187	9.2E-15	-0.96	LOC101927636
UBE2Z	rs12453394	rs318092	1381	1.2E-14	-0.93	*UBE2Z*
HLA-DQA1	rs9271774	rs9271770	48612	6.7E-11	1.2	LOC100996809
			OD			
SSR1	rs9505118	rs3087986	1363	2.6E-12	0.35	*SSR1*
UBE2Z	rs12453394	rs3744608	3813	4.1E-12	-0.77	*UBE2Z*
BRAF	rs9648716	rs28529157	81058	2E-08	-0.45	*BRAF*
Mahajan et al., 2018						
			PPP			
GWAS gene	**rs ID**	**eQTL SNP**	**eQTL distance**	**eQTL FDR**	**eQTL beta**	**eQTL gene**
TTLL6	rs2032844	rs11657371	-145547	1.8E-05	-0.64	*UBE2Z*
MACF1	rs2296172	rs61779279	287263	0.0010	-0.2	*MACF1*
MLX	rs665268	rs646123	-114000	0.0014	0.24	*CNTNAP1*
			OD			
KIF9	rs2276853	rs2276854	-47481	0.0009	-0.32	*KLHL18*
CENTD2	rs56200889	rs12575364	-56695	0.0009	0.29	*STARD10*
TTLL6	rs2032844	rs11657371	-145547	0.0014	-0.41	*UBE2Z*

GWAS SNP in ARAP1 (CENTD2) is associated with T2D

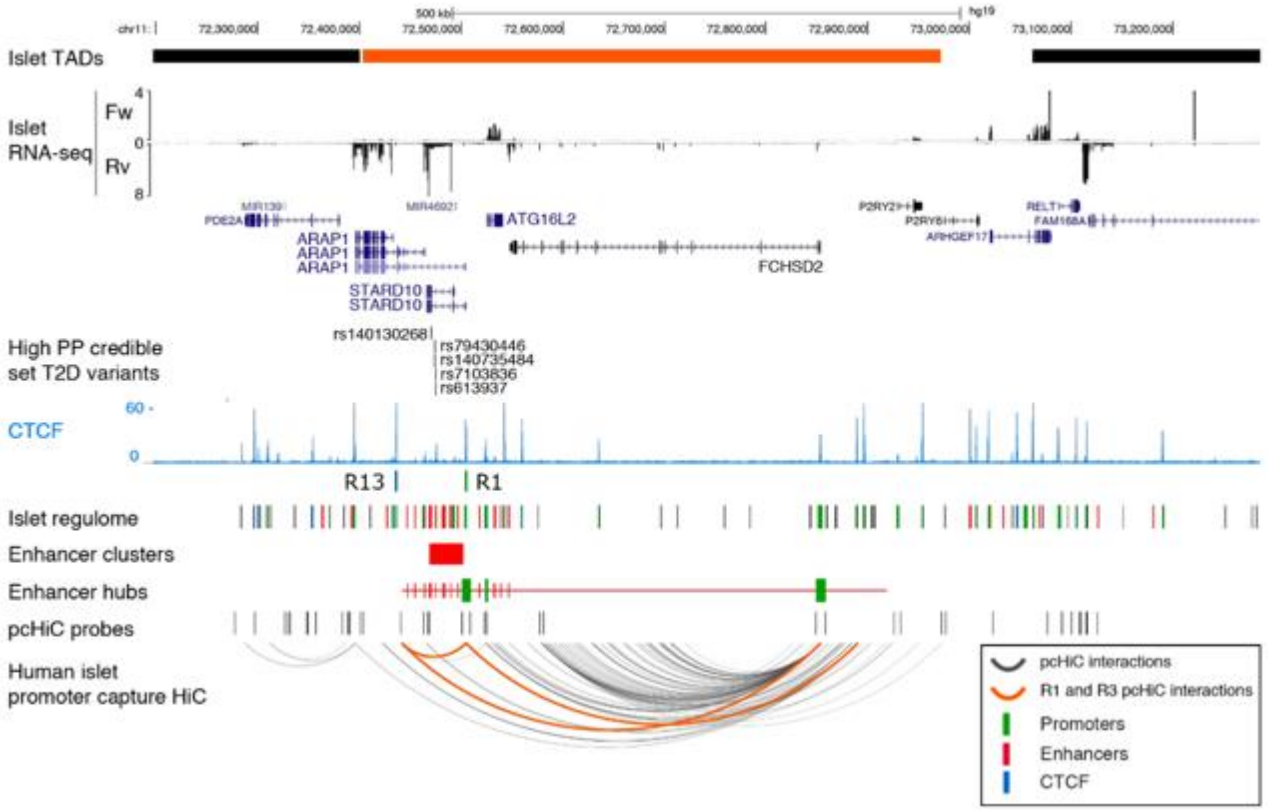
- GWAS results: Genetic variants near ARAP1 (CENTD2) and STARD10 influence T2D risk
- eQTL results: SNP associated with change in neighboring STARD10 and FCHSD2 mRNA - no change in ARAP1 mRNA levels



Khamis et al., 2019

Implicated genes: *STARD10* and *FCHSD2*

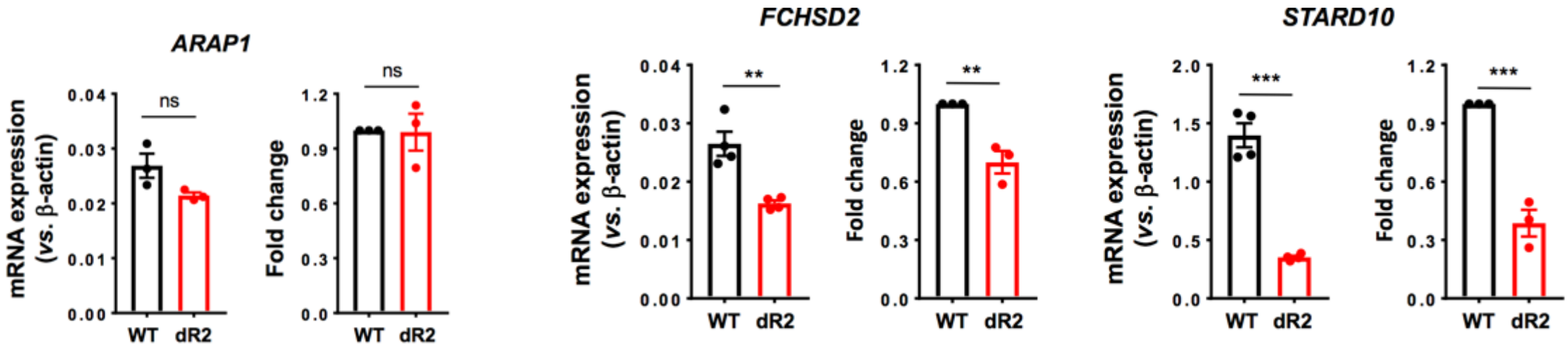
- Chromatin conformation capture: an enhancer cluster in the STARD10 T2D locus
- Region physically interacts with CTCF- binding regions and with an enhancer possessing strong transcriptional activity.
- Analysis of human islet 3D chromatin interaction maps identified FCHSD2 and STARD10 as targets of the enhancer cluster



Hu et al., 2020, submitted

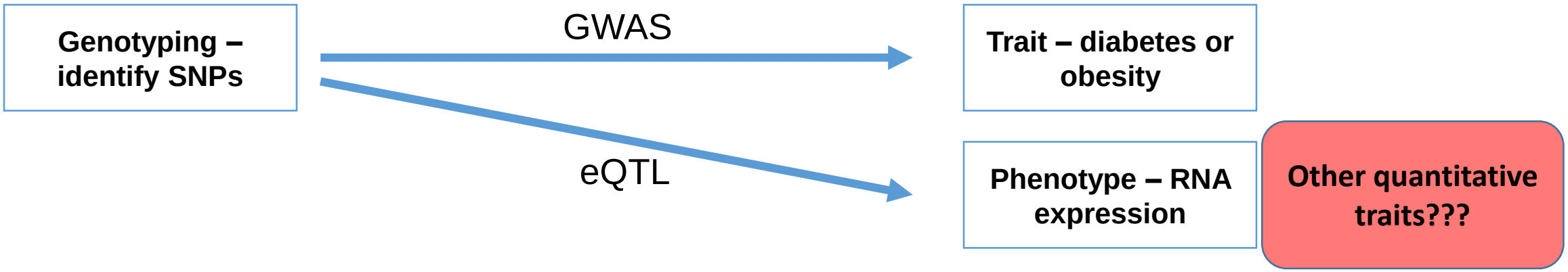
Deletion of enhancer region

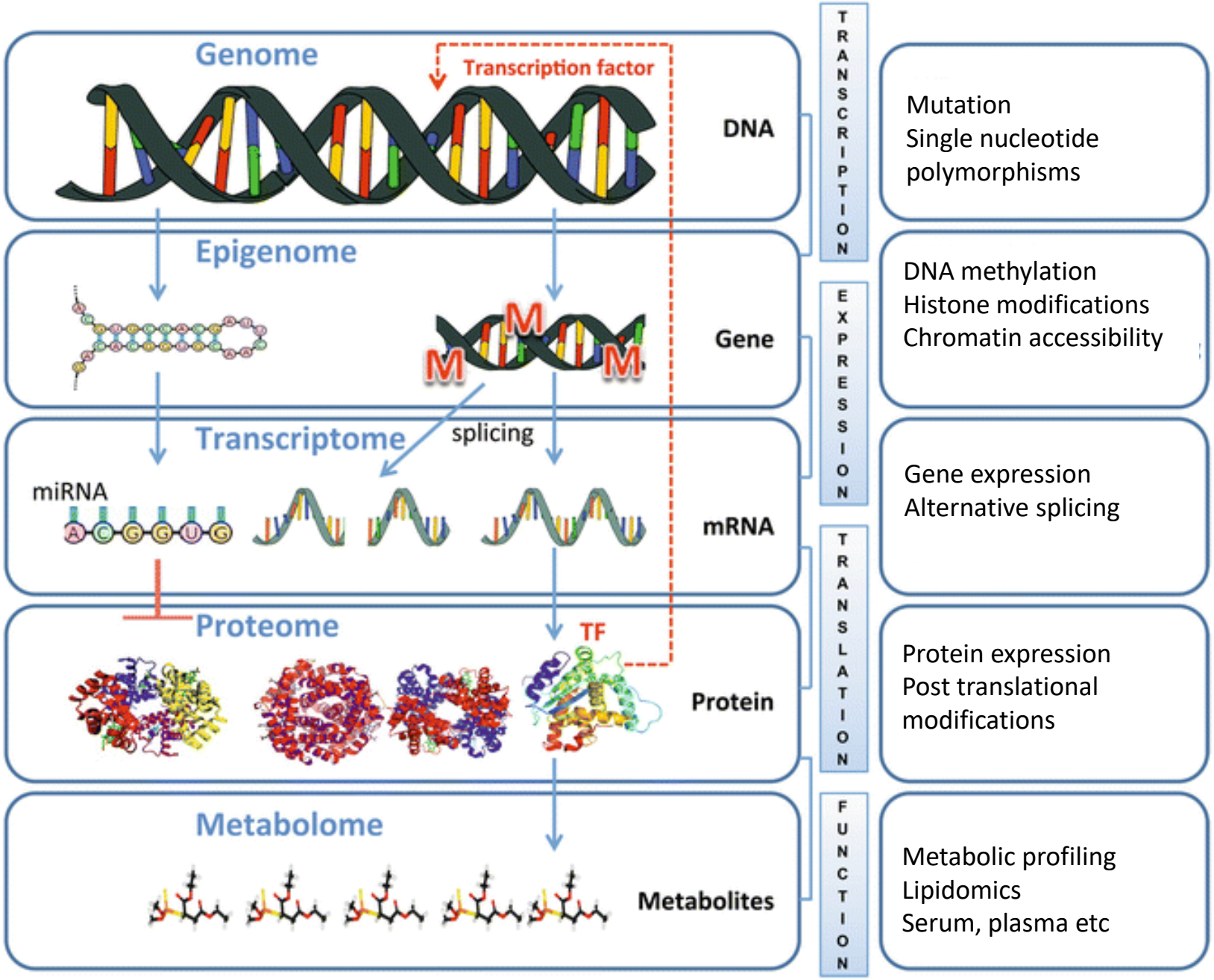
- Deletion of the variant region, or an associated enhancer (R2), from EndoC- βH1 cells using CRISPR-Cas9
- EndoC- βH1: insulin-secreting beta cell line
- Reduction in STARD10 and FCHSD2 – not ARAP1
- Confirmation of eQTL results



Hu et al., 2020, submitted

Expression Quantitative Trait Loci (eQTL)

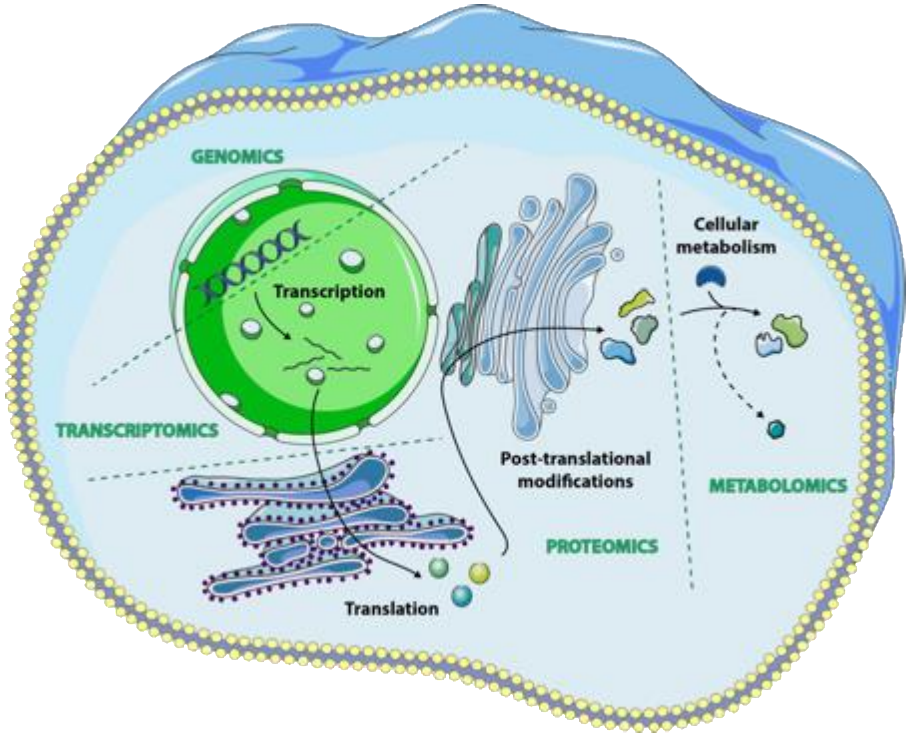



Genome
Transcription factor
DNA
Epigenome
M
M
M
Gene
Transcriptome
splicing
miRNA
A C G G U G
mRNA
Proteome
TF
Protein
Metabolome
Metabolites
TRANSCRIPTION
EXPRESSION
TRANSLATION
FUNCTION
Mutation
Single nucleotide
polymorphisms
DNA methylation
Histone modifications
Chromatin accessibility
Gene expression
Alternative splicing
Protein expression
Post translational
modifications
Metabolic profiling
Lipidomics
Serum, plasma etc

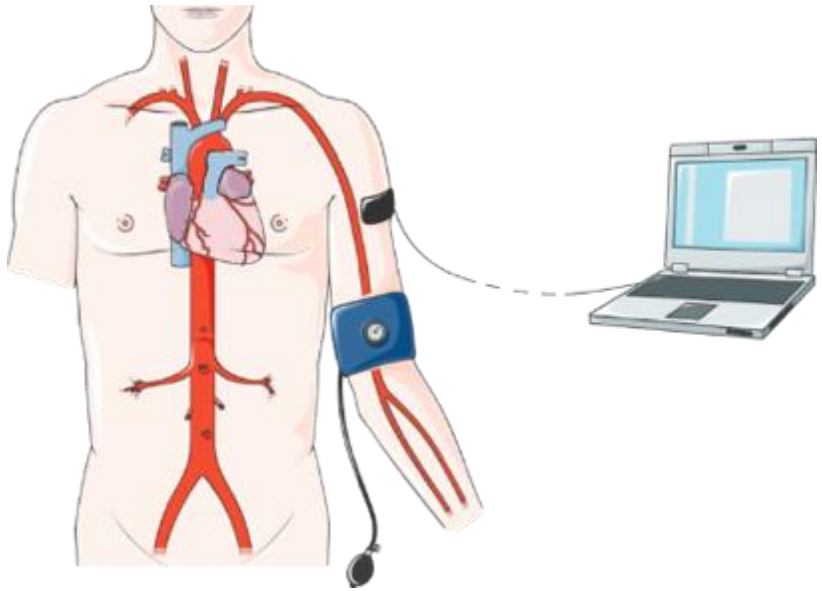
Different OMICS

eQTL

Genome → Epigenome → Transcriptome → Phenome



Cell measurements



Biomarker measurements

Epigenetics

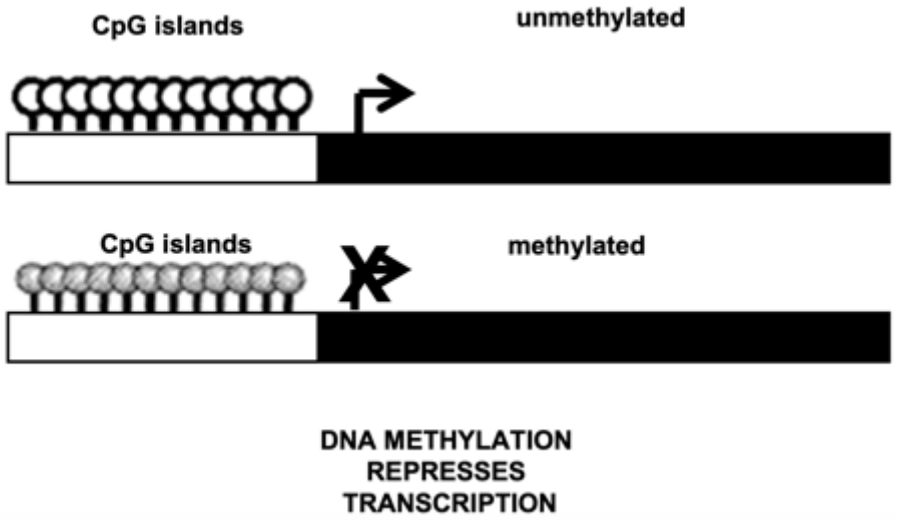
- Unlike genetic changes, epigenetic changes are reversible and do not change your DNA sequence, but they can change how your body reads a DNA sequence.
- Gene expression refers to how often or when proteins are created from the instructions within your genes. While genetic changes can alter which protein is made, epigenetic changes affect gene expression to turn genes “on” and “off.” Since your environment and behaviors, such as diet and exercise, can result in epigenetic changes, it is easy to see the connection between your genes and your behaviors and environment.

Types of epigenetic modifications:

- 1: DNA Methylation
- 2: Histone modification
- 3: Non-coding RNA

1: DNA methylation

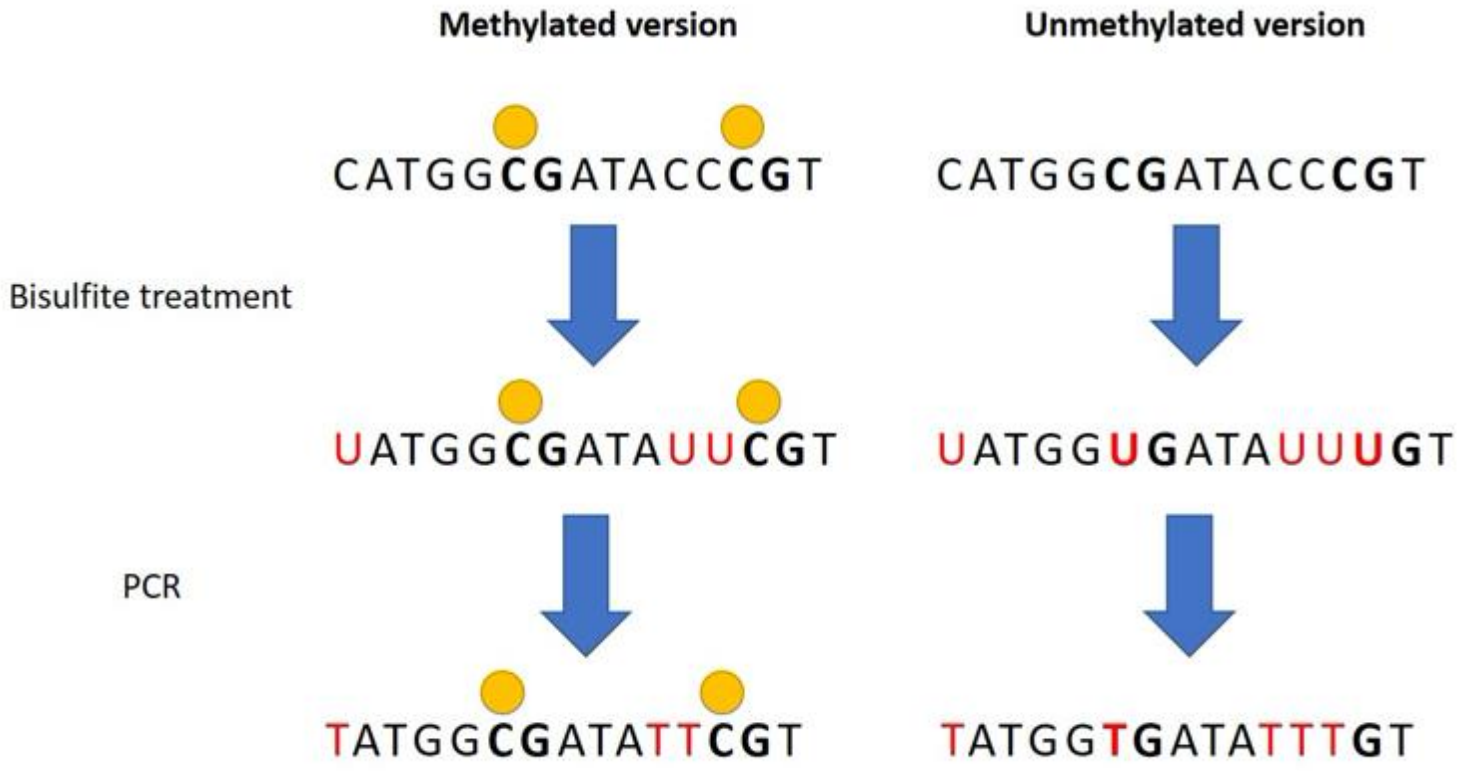
- Epigenetics: the study of DNA heritable changes through the modification of the genome, such as histone modifications, non-coding RNA transcription and DNA methylation, which regulate crucial cellular processes, such as differentiation and gene expression.
- DNA methylation patterns are crucial as methylation of CpG sites within regulatory regions, such as transcription start sites or promoter, often lead to the silencing of gene expression.



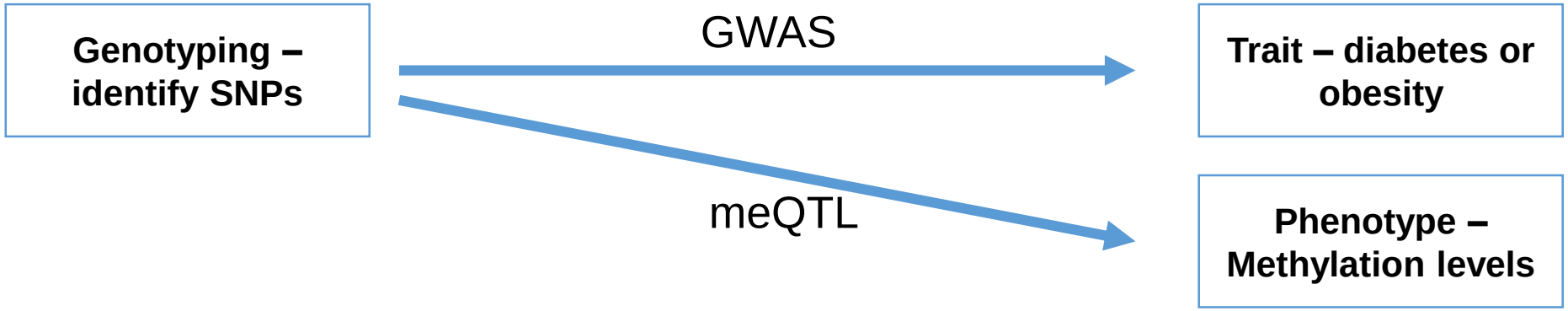
Detecting DNA methylation

- Bisulfite sequencing: sequence the methylation pattern of DNA
- Bisulfite treatment converts cytosine to uracil, but not 5-methylcytosine allowing for the identification of methylated C in the CpGs
- New cost-effective methodological advances has contributed to an increasing interest within the field
- The development of Illumina's Infinium arrays made it possible to analyse hundreds of thousands of CpG sites in a single array
- New technologies: Bisulfite-free methylation sequencing (enzymatic)

Bisulfite treatment: deamination of unmethylated cytosine residues to uracil



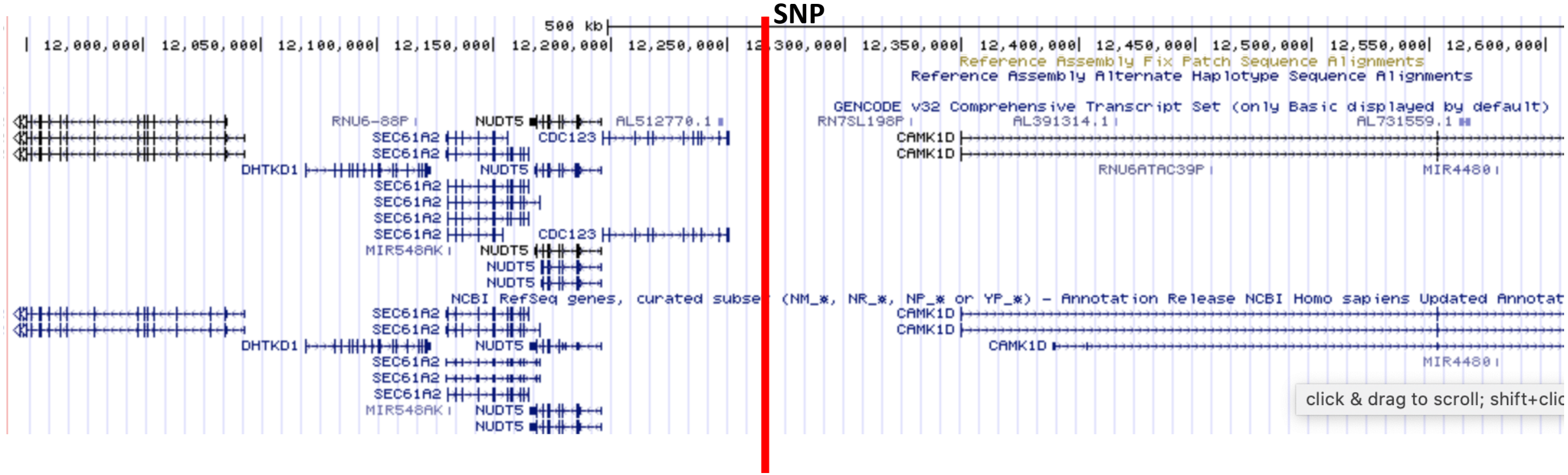
Methylation Quantitative Trait Loci (meQTL)



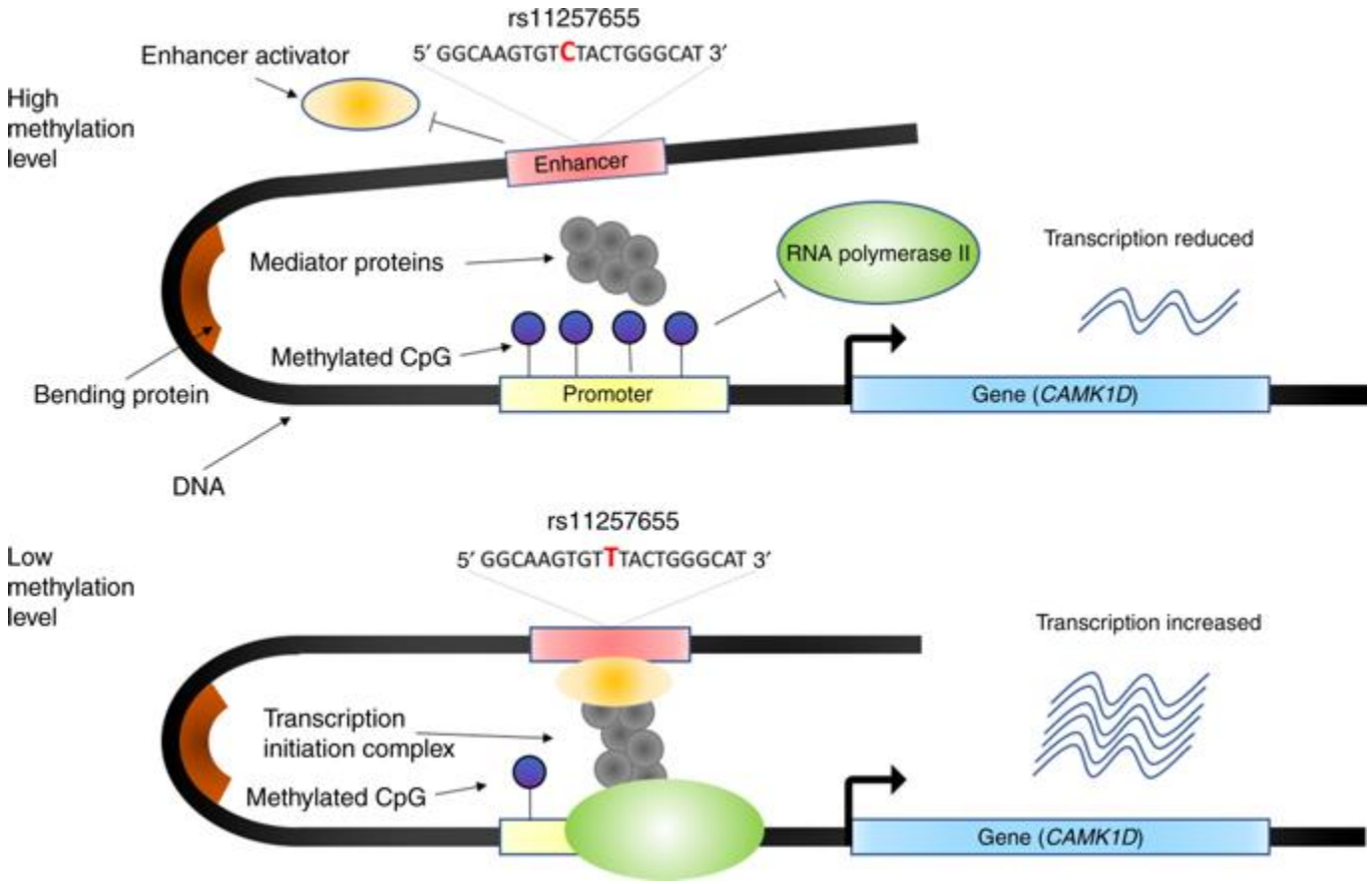
Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes

Angli Xue, Yang Wu, Zhihong Zhu, Futao Zhang, Kathryn E. Kemper, Zhili Zheng, Loic Yengo, Luke R. Lloyd-Jones, Julia Sidorenko, Yeda Wu, eQTLGen Consortium, Allan F. McRae, Peter M. Visscher, Jian Zeng & Jian Yang

Nature Communications **9**, Article number: 2941 (2018) | Cite this article



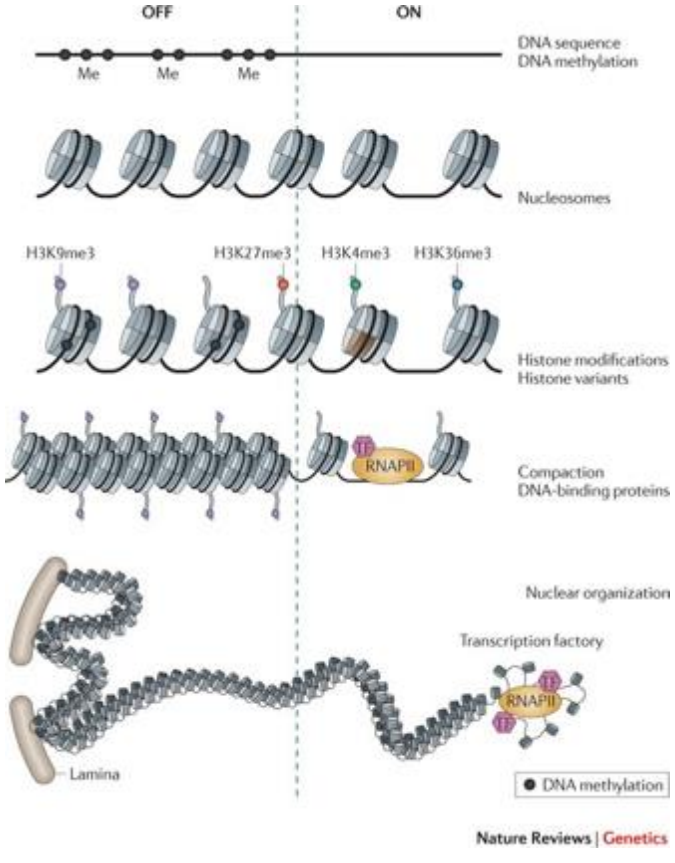
Model of the genetic mechanism at *CAMK1D* for T2D risk



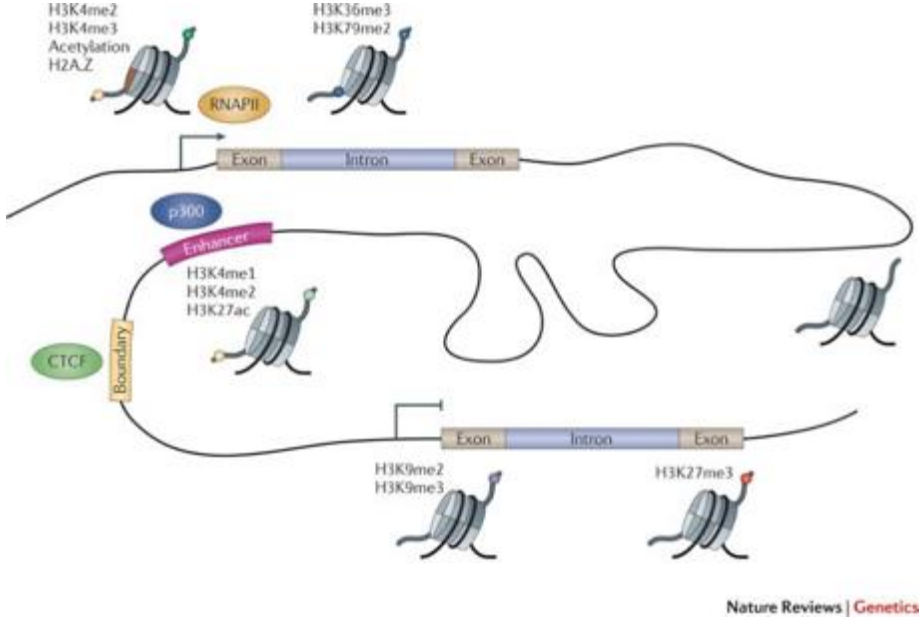
- T allele of rs11257655 increases *CAMK1D* expression by reducing the methylation level at cg03575602.
- In the presence of the T allele at rs11257655, *FOXA1*/*FOXA2* and other transcription factors bind to the enhancer region & form a protein complex
- This leads to a decrease in the DNA methylation level of the promoter region of *CAMK1D* and an increase in the expression of *CAMK1D*

2: Histone modifications

DNA wraps around proteins called histones. DNA wrapped tightly around histones cannot be accessed by proteins that "read" the gene. Some genes are wrapped around histones and are turned "off" while some genes are not wrapped around histones and are turned "on."

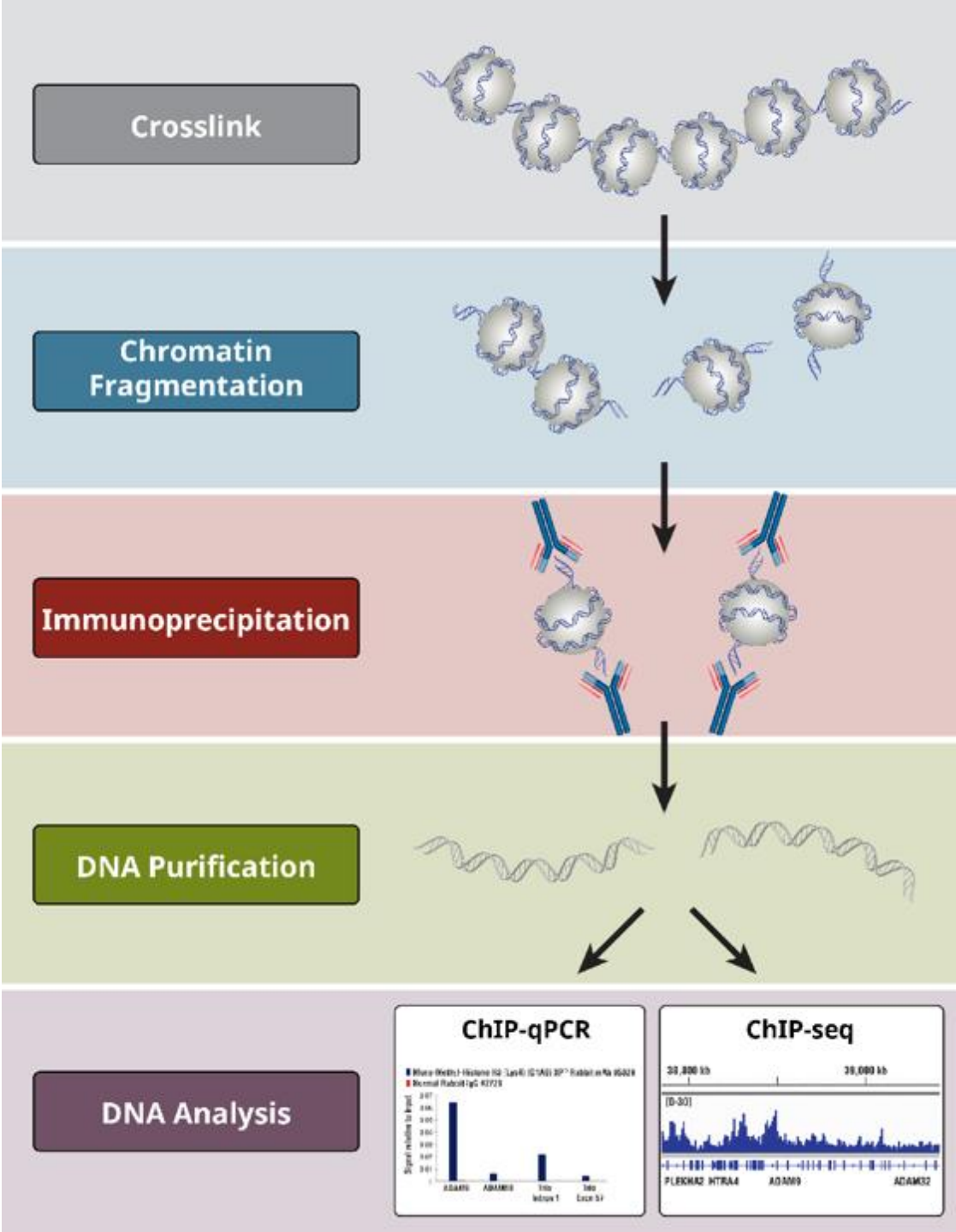


Histone modification	Function	Location
H3K4me1	Activation	Enhancers
H3K4me3	Activation	Promoters
H3K36me3	Activation	Gene bodies
H3K9Ac	Activation	Enhancers, promoters
H3K27Ac	Activation	Enhancers, promoters



Methods to detect histone modifications

- **ChIP-sequencing**, also known as **ChIP-seq:** method used to analyse protein interactions with DNA.
- **ChIP-seq** combines chromatin immunoprecipitation (**ChIP**) with DNA **sequencing** to identify the binding sites of DNA-associated proteins.



Epigenetic modifications + QTL

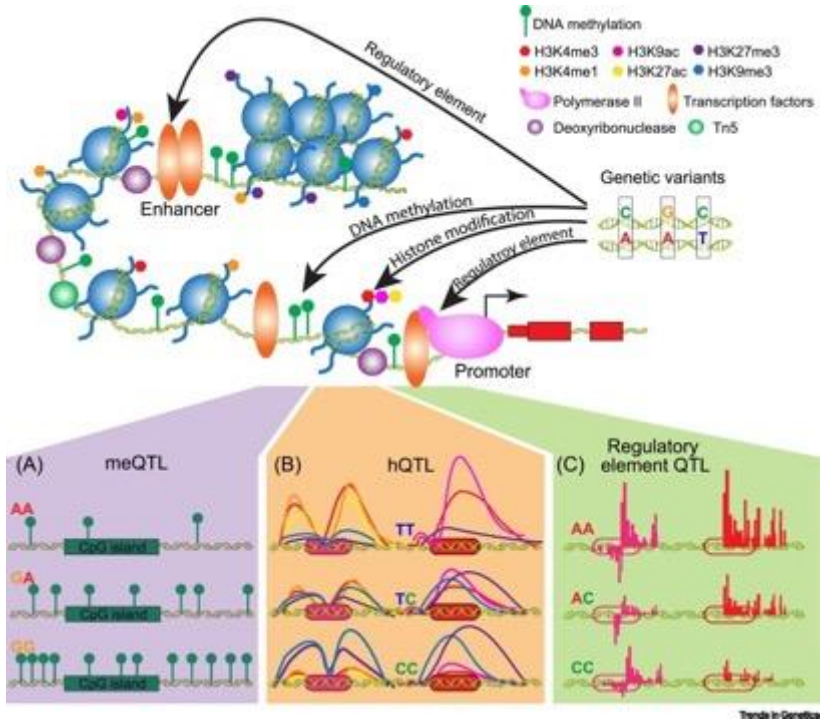
Genotyping – identify SNPs

GWAS

Trait – diabetes or obesity

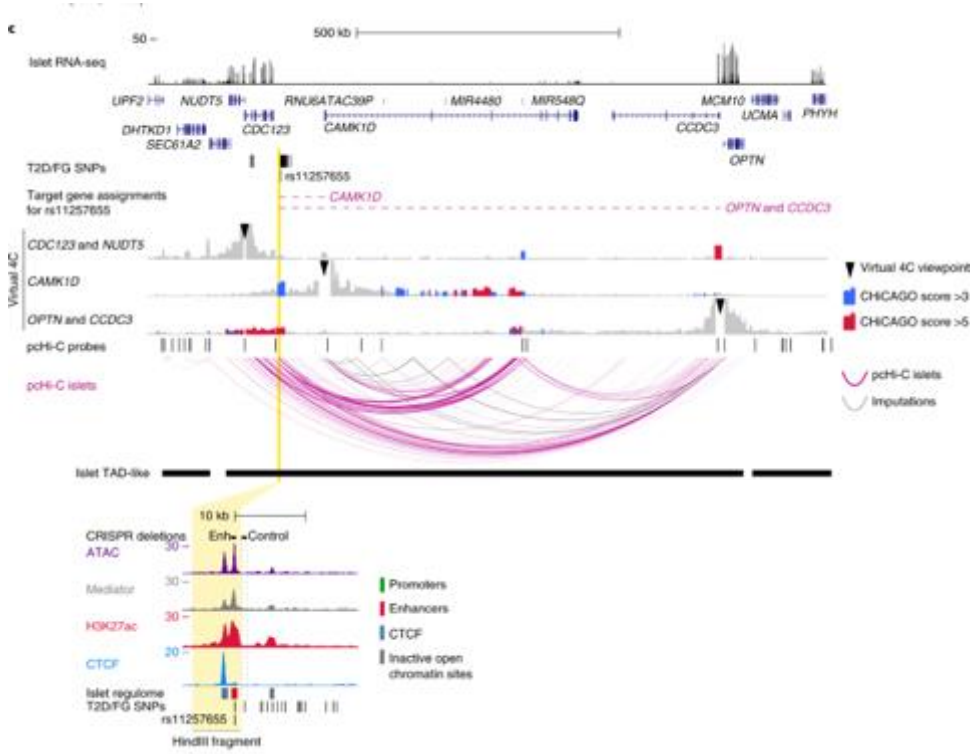
hQTL

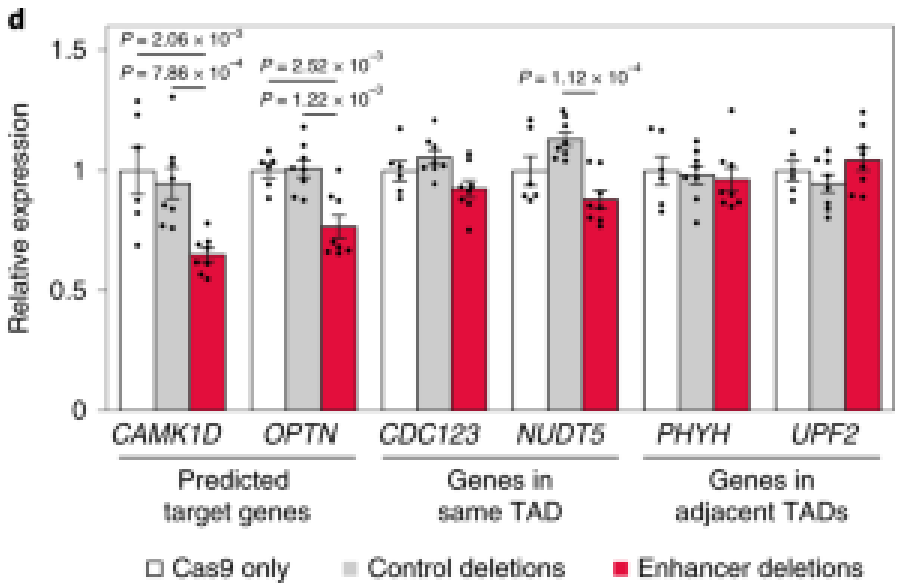
Phenotype – Histone marks



Miguel-Escalada et al., 2019: Nature Genetics

The enhancer showed moderate-confidence interactions with *CAMK1D*, but, more surprisingly, showed high-confidence pcHi-C interactions with a more distant gene, *OPTN* . Accordingly, deletion of this enhancer (but not an adjacent region), or silencing with KRAB-dCas9, led to selectively decreased expression of both *OPTN* and *CAMK1D*, whereas targeted activation of the enhancer stimulated their expression



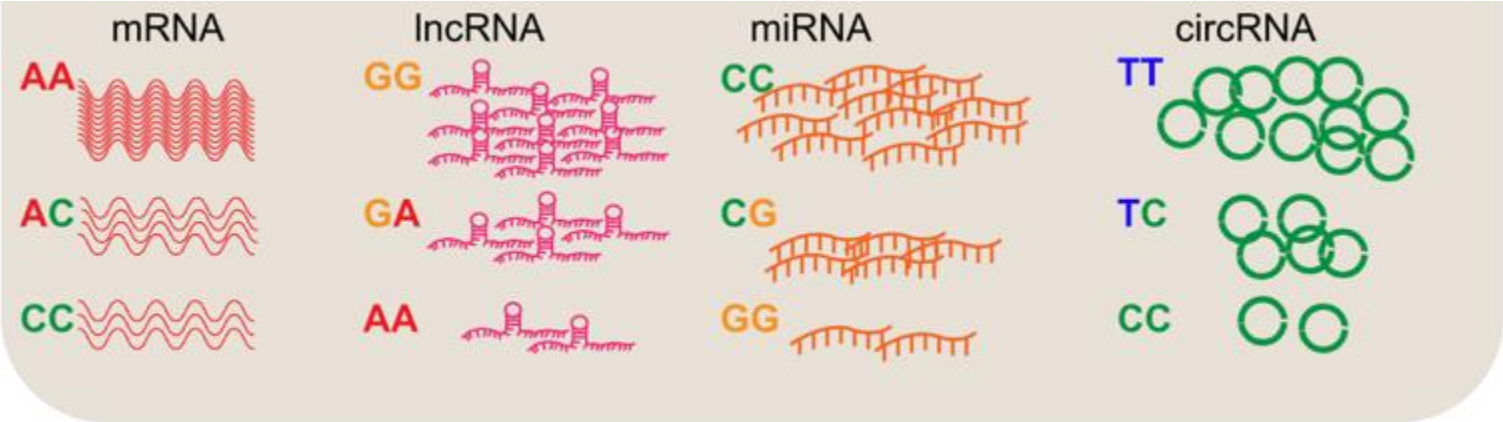


3: Non-coding RNA

DNA is used as instructions for making coding and non-coding RNA. Coding RNA is used to make proteins. Non-coding RNA helps control gene expression by attaching to coding RNA, along with certain proteins, to break down the coding RNA so that it cannot be used to make proteins.

mRNA main function: protein synthesis

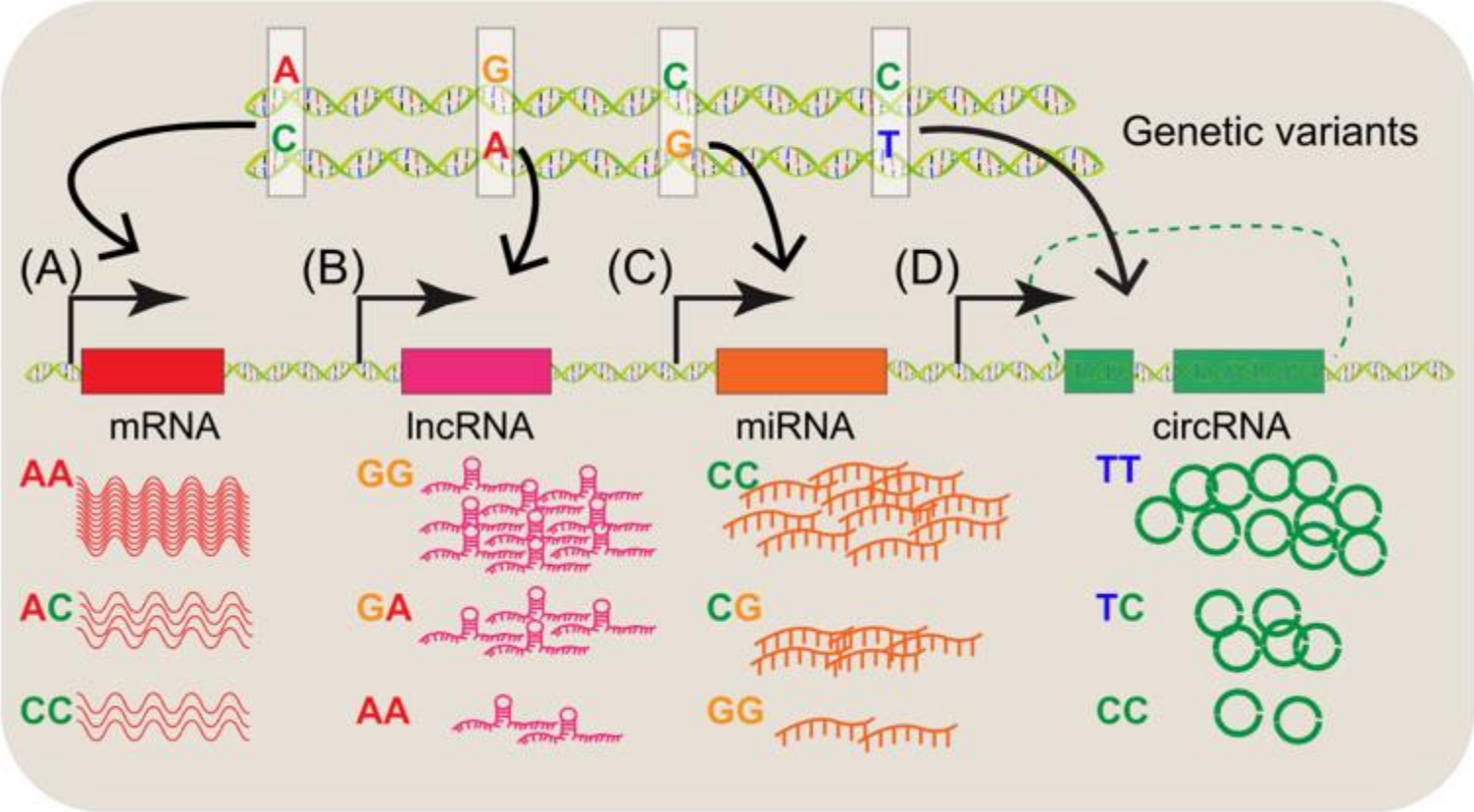
miRNA: functions as a guide by base-pairing with target mRNA to negatively regulate its expression.



lncRNA: is to serve as a molecular signal to regulate transcription in response to various stimuli.

circRNAs: action through miRNA sponge to regulate target gene expression by inhibiting miRNA activity.

Non-coding RNA + QTL

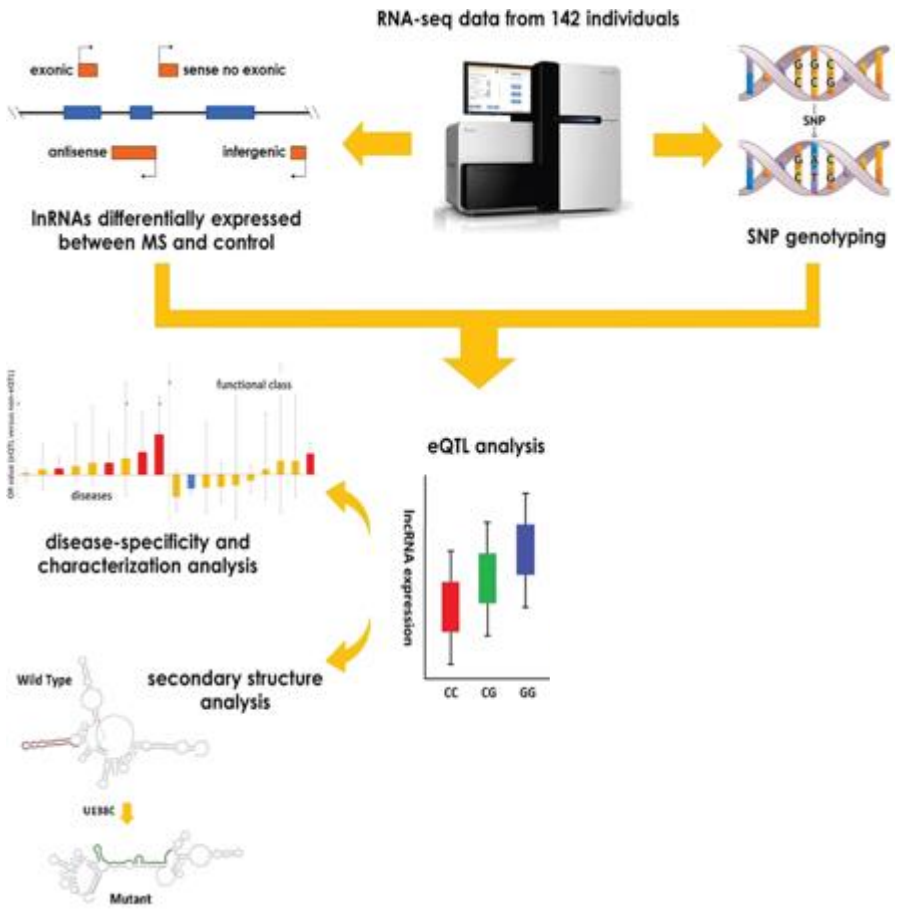


Next generation sequencing to detect RNAs

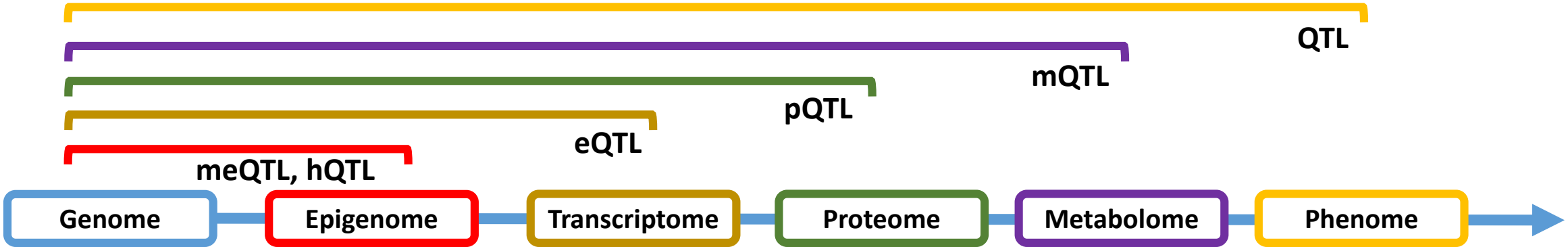
RNA sequencing method	Description and benefits
Total RNA Whole transcriptome	Whole transcriptome analysis to examine coding and noncoding RNA simultaneously; suitable for novel discovery. More throughput intensive to achieve high enough coverage for discovery. Potential inefficiencies and bias due to different sequencing lengths.
mRNA sequencing	Poly(A) selection to sequence all messenger RNA for gene expression analysis; able to identify novel and known content
smRNA sequencing	Isolation of small RNA to focus study on noncoding RNA to identify novel and known content such as microRNA (miRNA)
Targeted RNA sequencing	Sequencing specific transcripts of interest to focus efforts and lower cost to analyze specific genes of interest. Can be used for many sample types, including degraded samples from FFPE.

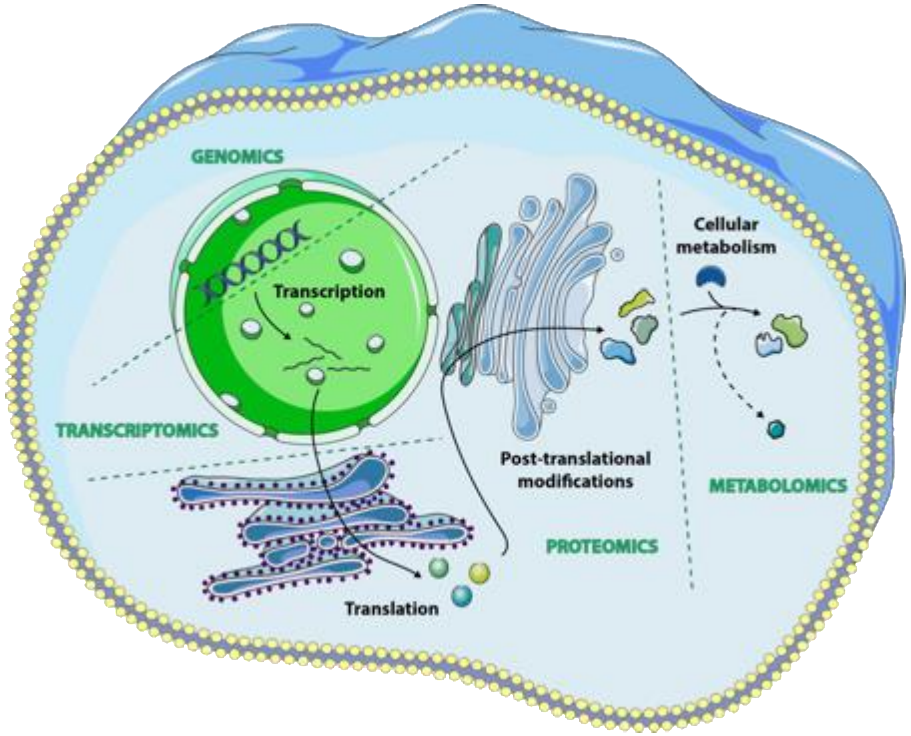
Genome-wide identification and analysis of the eQTL lncRNAs in multiple sclerosis based on RNA-seq data – Han et al., 2019

- In this study, a bioinformatics strategy was applied to obtain lncRNA expression and SNP genotype data simultaneously from 142 samples (51 MS patients and 91 controls) based on RNA-seq data, and an expression quantitative trait loci (eQTL) analysis was conducted.
- 517 lncRNAs were affected by SNPs. T
- The secondary structure was altered in 17.6% of all lncRNAs in MS.

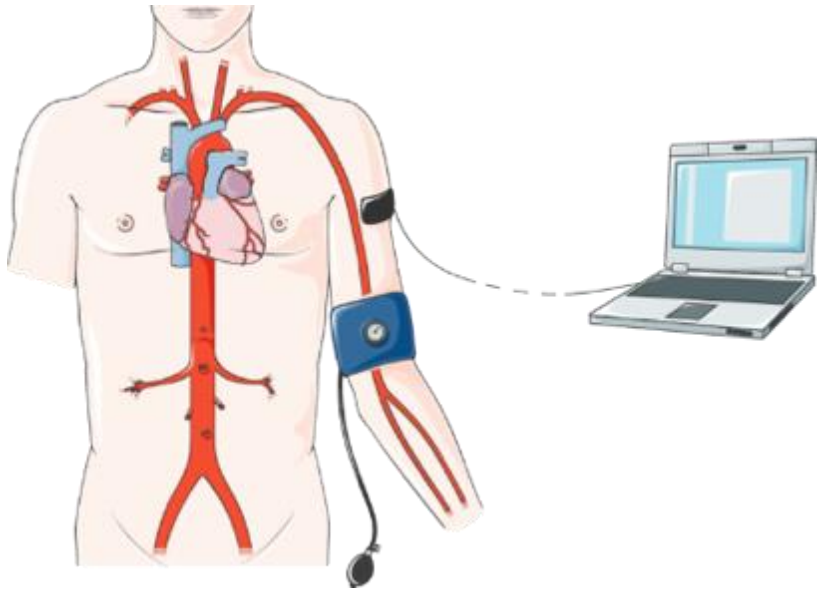


Summary (1)





Cell measurements

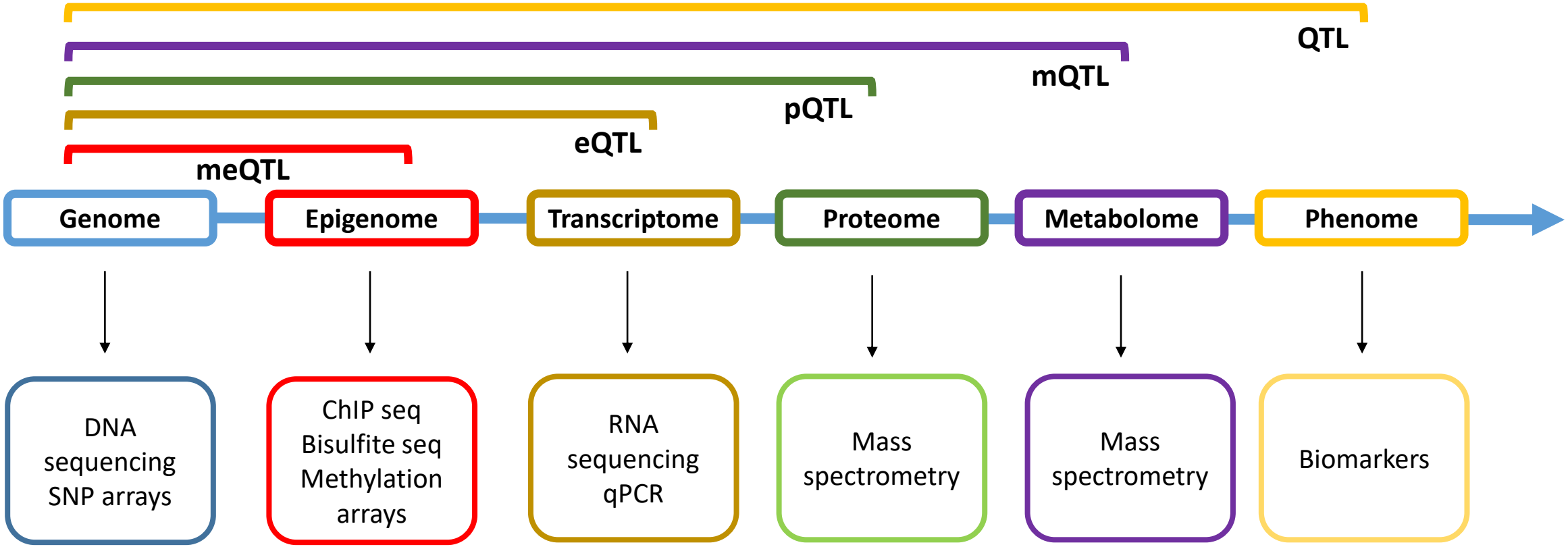


Biomarker measurements

Summary (1)

- OMICS alone are very valuable in identifying genes and loci associated with disease
- However, understanding the function of these genetic variants has been difficult
- Combining OMICS has been a valuable tool in identifying the causal loci associated with disease
- Multiple types of ‘omics data generated from high-throughput technologies enable the discovery of novel types of QTL, spanning the epigenome, transcriptome, and proteome to the metabolome, to link the genotype and phenotype.
- Integrative analysis of multidimensional data provides alternative strategies to understand the functional effects of genetic variants.

Summary (2)


QTL
mQTL
pQTL
eQTL
meQTL
Genome
Epigenome
Transcriptome
Proteome
Metabolome
Phenome
DNA sequencing SNP arrays
ChIP seq Bisulfite seq Methylation arrays
RNA sequencing qPCR
Mass spectrometry
Mass spectrometry
Biomarkers

Any questions?

amna.khamis@cnrs.fr