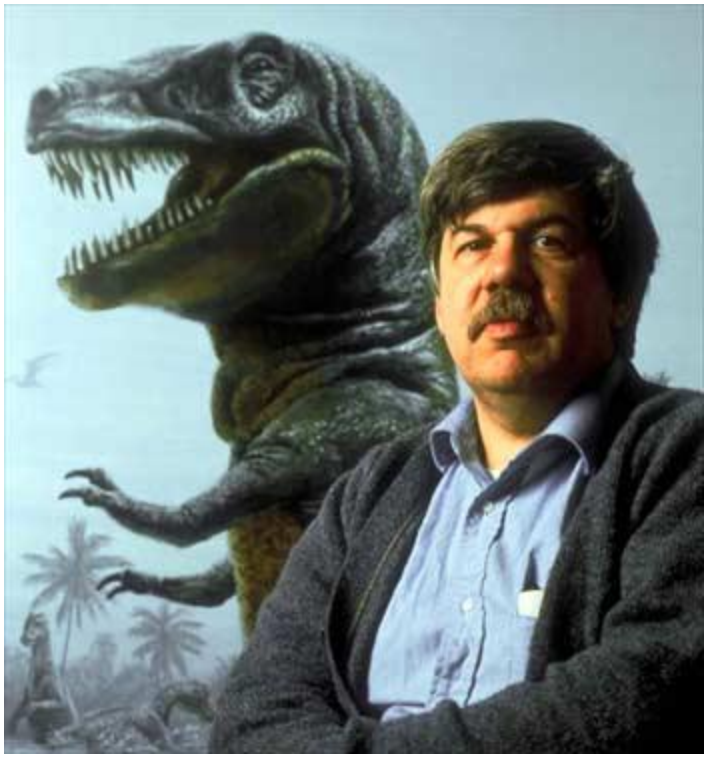
Molecular Epidemiology (First practical book 1993)
P. Eline Slagboom, biologist, Prof. of Molepi

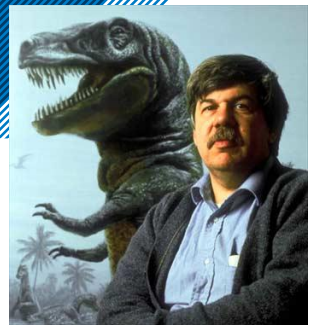


Stephen Jay Gould, Evolutionary Biologist
1982; 41 years , abdominal mesothelioma
Prognosis: median survival is 8 months

I met him in 1993, the year of my PhD.

He died 20 years after diagnosis.
What happened.
Gould: “The Median is’nt the Message”

Four lines of prognosis research to explain Stephen Jay Gould to live up to 2002



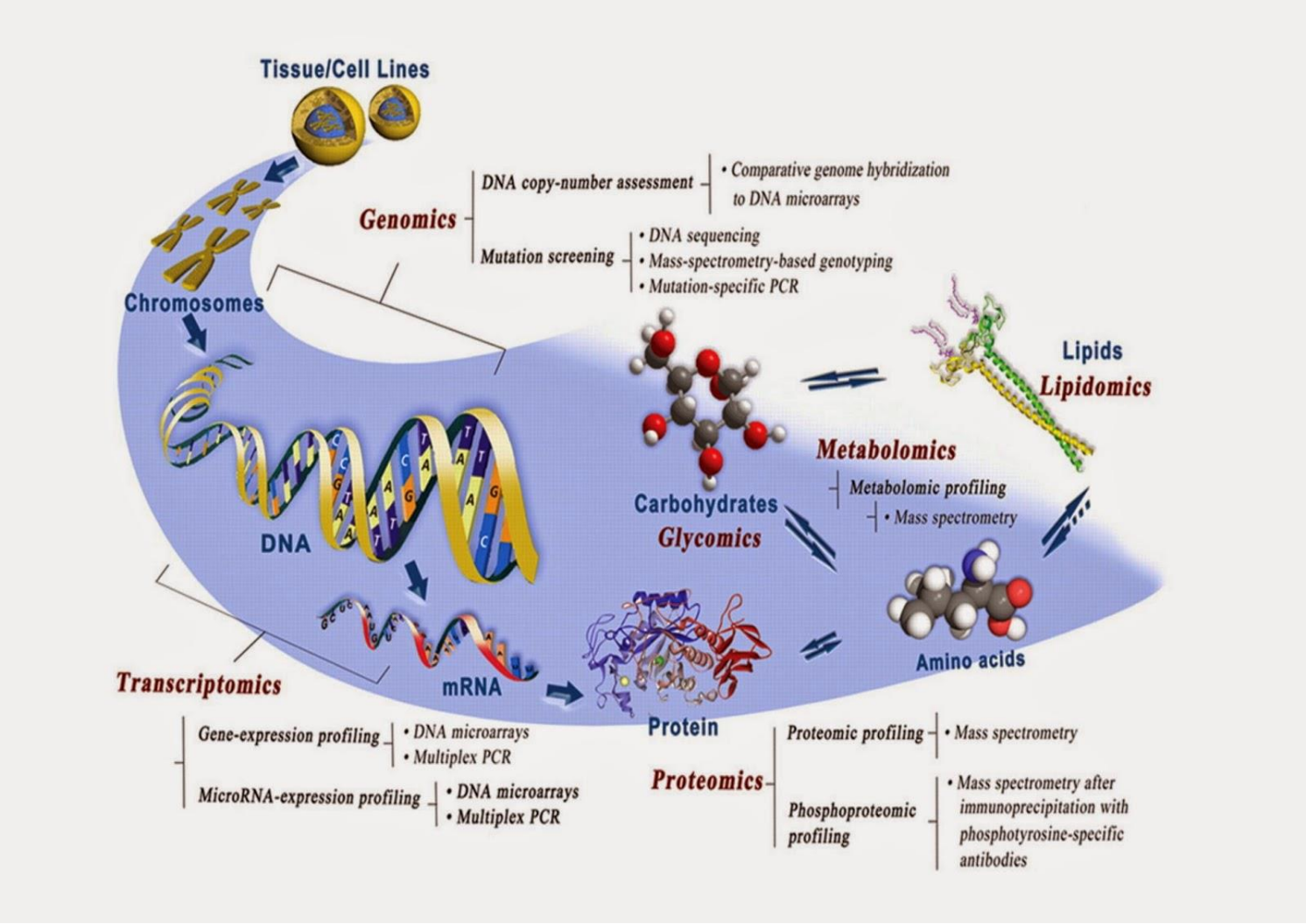
- Which hospital, country, environment, health care policy, socioeconomic status (very good) in 1982 (baseline) and 1982-2002: Fundamental prognosis research

- Which biomedical personal background: age (young) genomic (protection ?), behavior (lifestyle), psychology (stress, optimistic), immune response and cell replication capacity (biomarkers). In 1982-2002 Prognostic Factor research

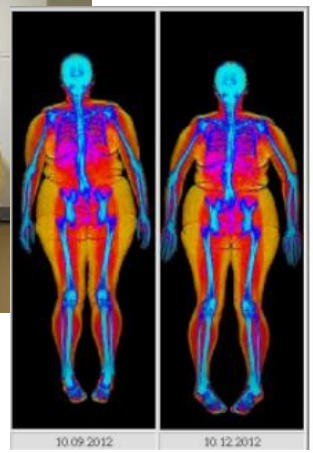
- Individual risk prediction. Could a specific model based on a combination of multiple variables have calculated Goulds individual risk of dying (instead of 8 months survival) given his starting point (baseline): Prognostic Modeling Research

- Was a particular treatment highly effective for Goulds biomedical background and characteristics and molecular profile of his mesothelioma. Stratified/Precision Medicine (experimental treatment of radiation, chemotherapy, and surgery)

Since 1990: Revolution in Molecular Technology
high throughput recording exposure, consequence (biological change), pathway analysis.
Holistic: Targeted and Untargeted (hypothesis-driven and -free)



Phenotypes in patients and populations (biobanking)



- Lifestyle, Demographic
- Morbidity, Physiological

Functional

- Handgrip strength
- Cognitive functioning (memory, attention, speed, MMSE)
- Short Physical Performance Battery Protocol (gait, balance, chair rise)
- Questionnaires (sleep, quality of life, mood, depression, MMSE, 24 hrs food recall)

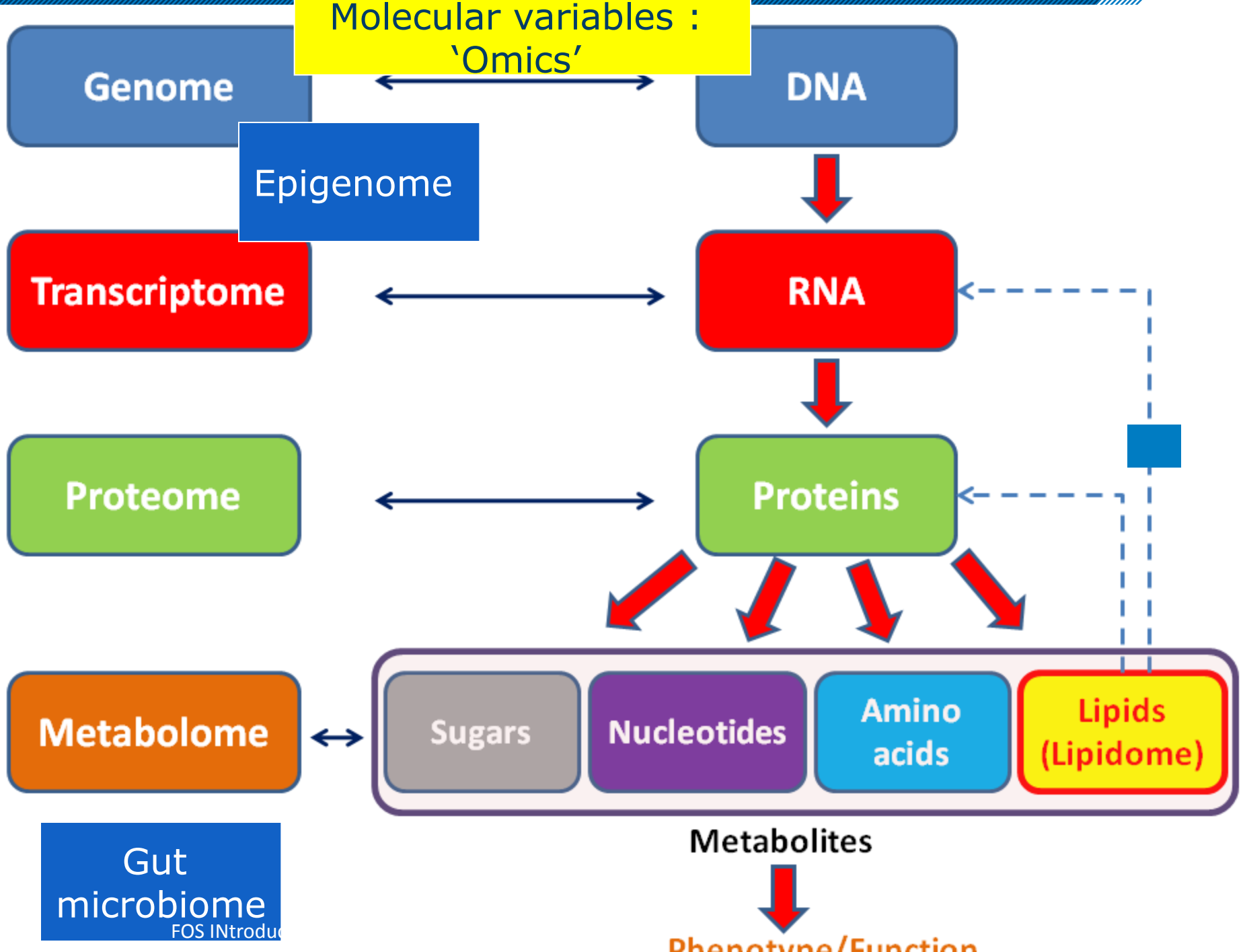
- Magnetische Resonantie Imaging (MRI)
- DEXA scans
- Wearable data
- Fluorenscence of the skin

Molecular Epidemiology

- Introduced by Kilbourne (1973), infectious diseases; Schulte and Perera (1993 Principles and Practices)
- Integrates Epidemiology, Medical Sciences and Molecular Biology
- Studies the influence on health of environmental and genetic risk factors measured by (holistic) molecular signatures
- Contributes to
  prediction/prognosis
  monitoring exposure, response to interventions
  etiological understanding (disease mechanisms)

Complex traits: effect of multiple genes and multiple environmental factors

What molecular data do we collect and in what study designs ?


Molecular variables : 'Omics'
Genome
DNA
Epigenome
Transcriptome
RNA
Proteome
Proteins
Metabolome
Sugars
Nucleotides
Amino acids
Lipids (Lipidome)
Metabolites
Gut microbiome
FOS INtrodu
Phenotype/Function

Biomaterials: biopsies at operations; visits in studies; by mail

SOURCE	INPUT	YIELD
BLOOD	200 µl	4-12 µg
BUFFYCOAT	200 µl	25-50 µg
MOUTH SWABS	1	3-10 µg
SALIVA	2 ML	110 µg
BONE	55-70 mg	5.5 µg
CARTILAGE	50-100 mg	2.4 µg

First : Study designs in human studies

http://cisncancer.org/research/images/epi_study_types-rew.jpg

Study designs

Observational studies (cohorts, patient populations)
Cross sectional (Case-Control)
Prospective

Experimental studies in humans (RCT)
Response to treatment
Recording exposure
(to food; in blood, urine, faecal samples)
In depth biological studies (i.e. cell biology)

Complex traits: effect of multiple genes and multiple environmental factors

Is the trait or an element in it heritable ?

How to calculate the genetic and environmental component of any phenotype ?

- Comparison of phenotype in MZ and DZ twins

  Heritability $h^2 = 2*(rMZ-rDZ)$

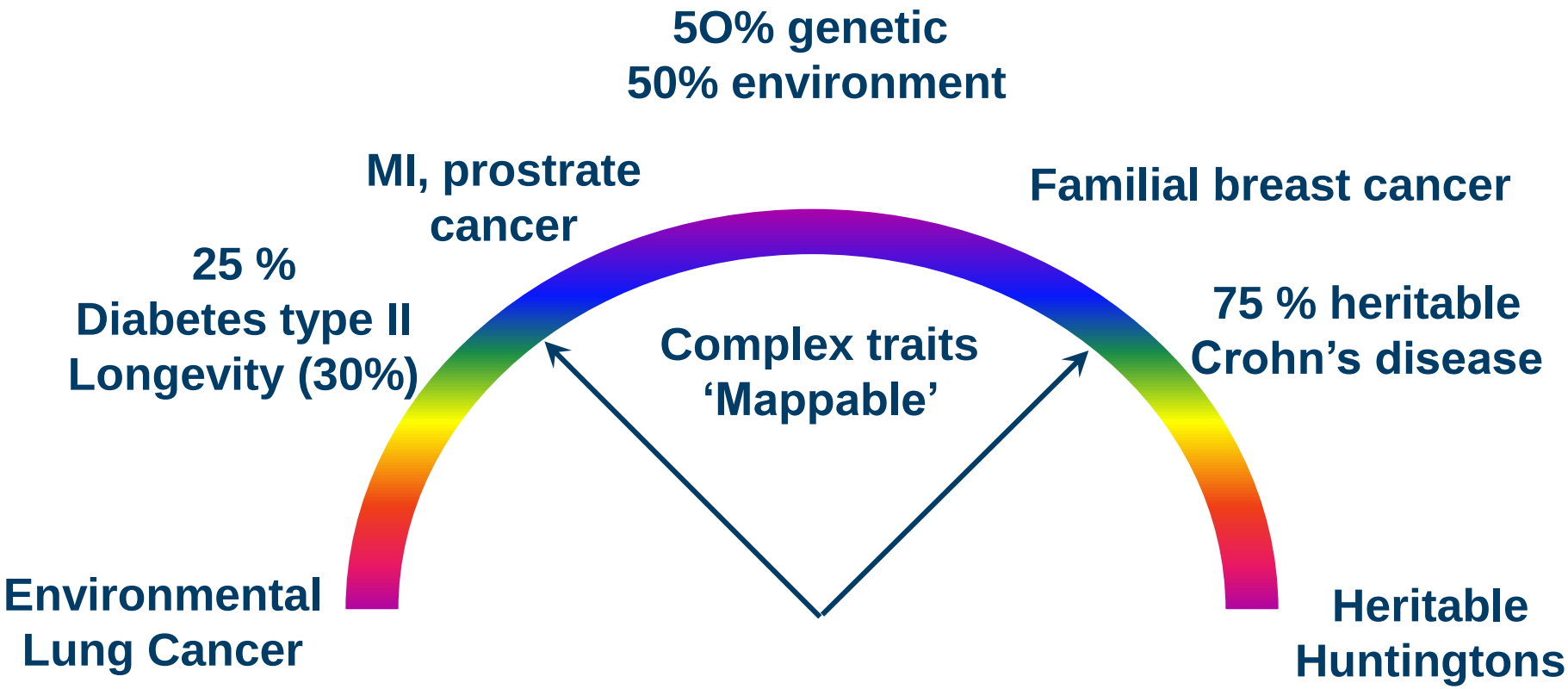
  Heritability of human longevity 33%

- Comparison of phenotype in family members

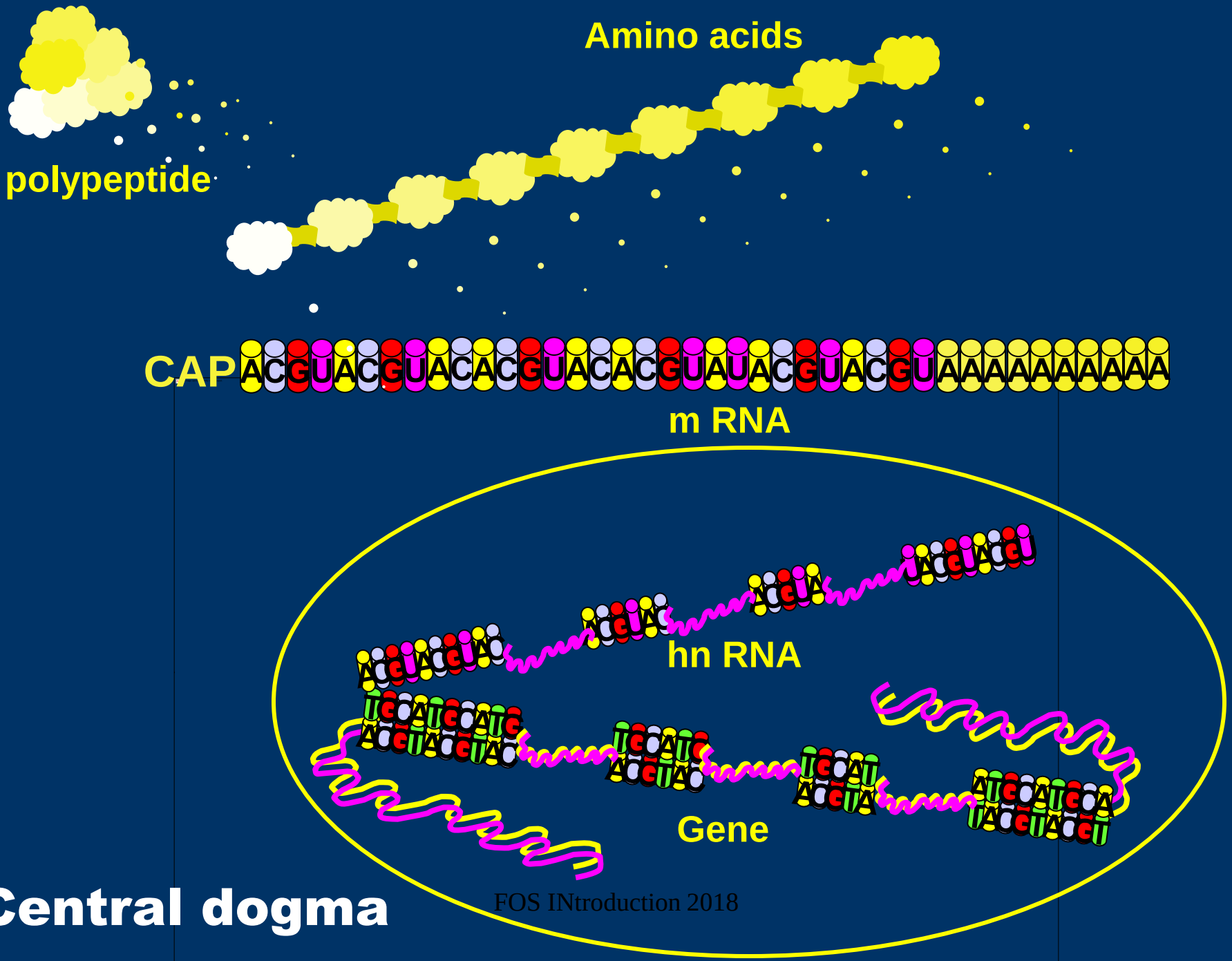
  $\lambda_{s=}$ $\frac{\text{risk for a sibling of an affected proband}}{\text{risk in the general population}}$

  cystic fibrosis $\lambda_{s=}500$ schizophrenia $\lambda_{s=}8.6$

Heritability for common human disease

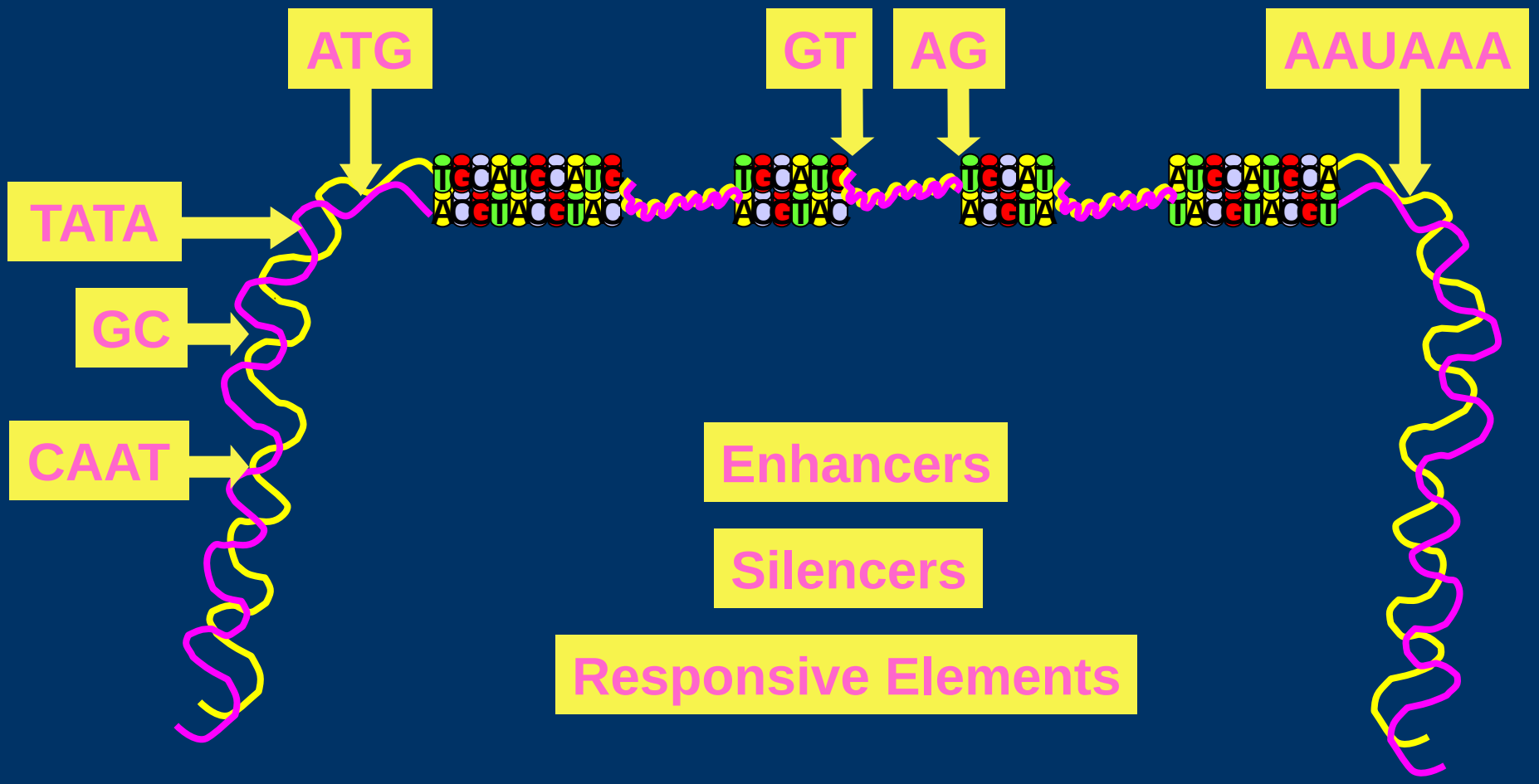


What type of genetic variation may affect Late life disease and ageing



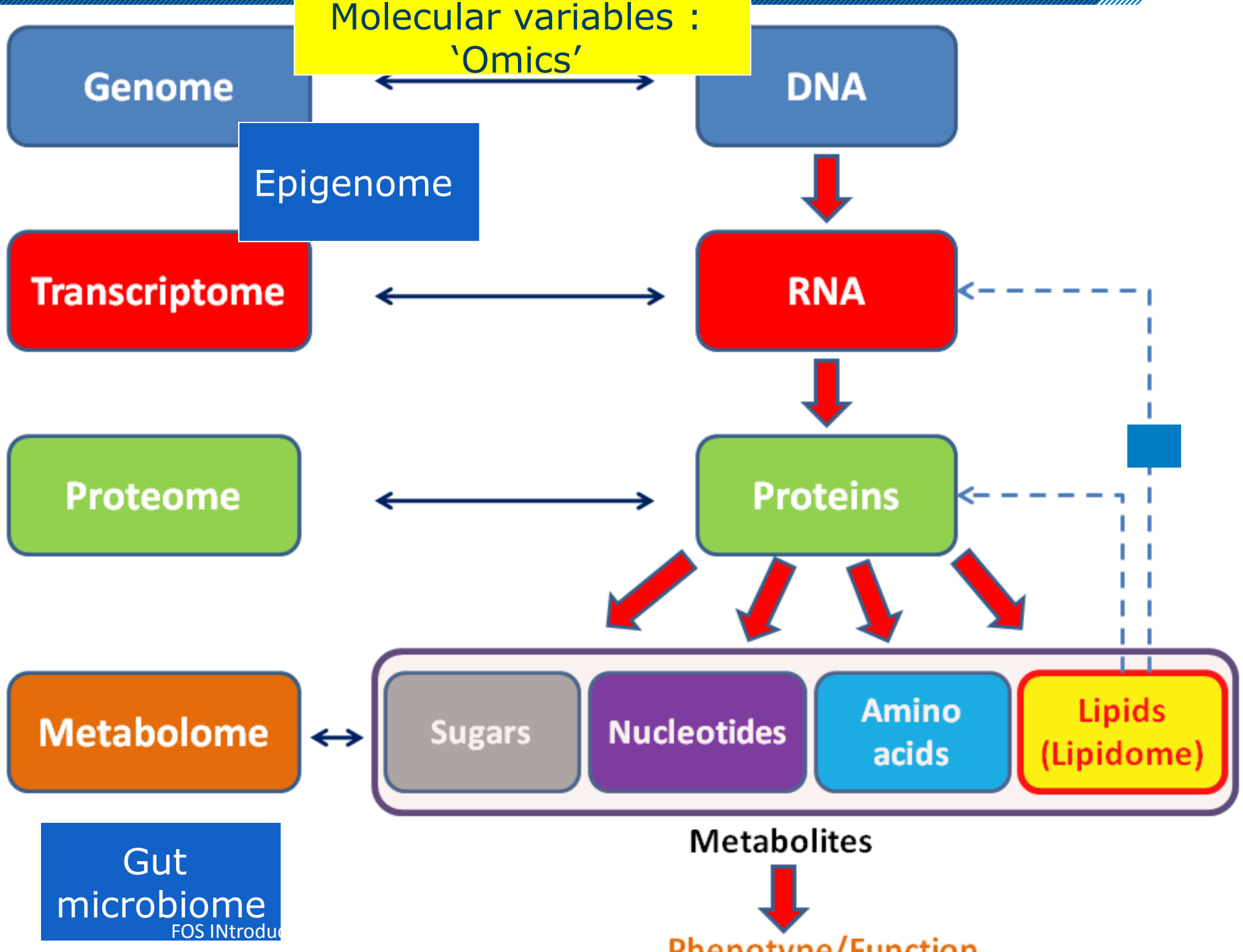
Central dogma

Signals for gene expression

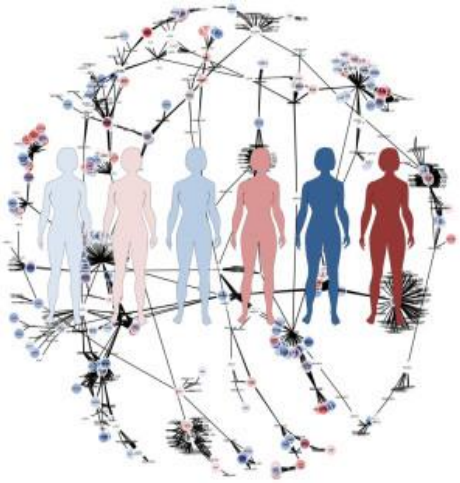


Genetic variation with mild effect: not expected in exons

If there is a genetic component in a trait , how do we find the genes involved ?


Molecular variables : 'Omics'
Genome
DNA
Epigenome
Transcriptome
RNA
Proteome
Proteins
Metabolome
Sugars
Nucleotides
Amino acids
Lipids (Lipidome)
Metabolites
Gut microbiome
FOS INtrodu
Phenotype/Function

BBMRI Biobanking consortium
Multi-level omics data



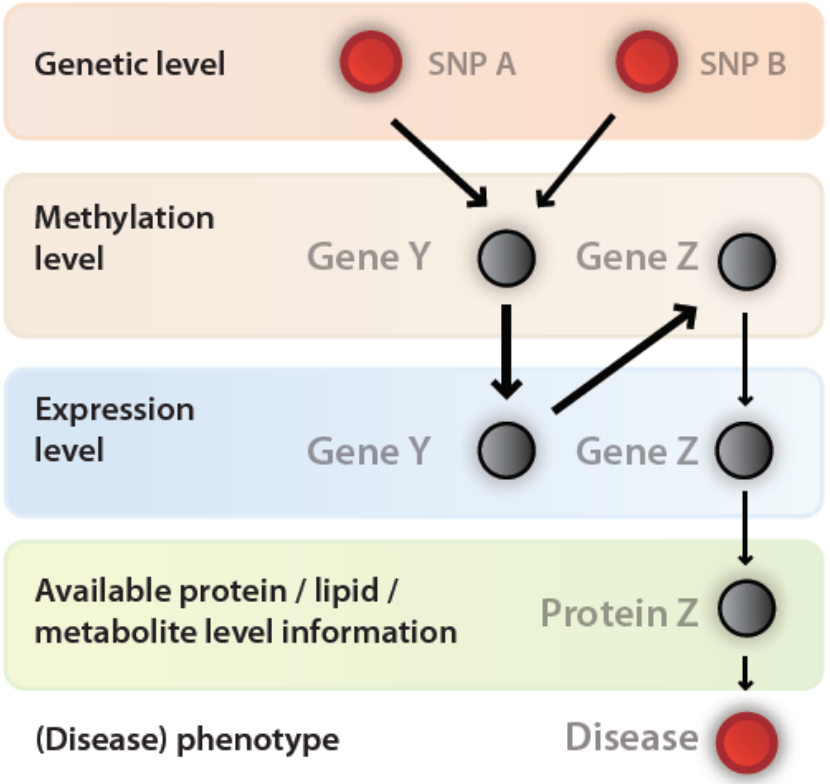
N=100,000 GWAS
N=750 Go.NlL

N=4,000

N=4,000

N=50,000

N>250,000



Go•NL
GENOMEoftheNETHERLANDS

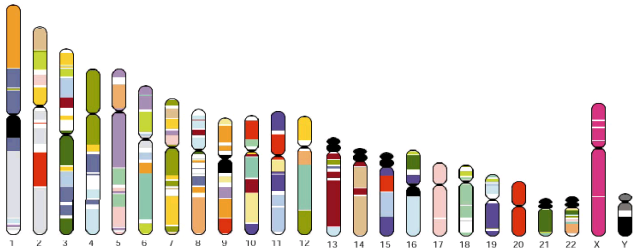
BIOS consortium

BIOS consortium

METABOLITE consortium

Central databases for the research community

Approaches to study how genetic variation contributes to disease (hypothesis driven and h-free, discovery science)



Genome wide association scan GWAS

Hypothesis generating
Genomic architecture (Explorative)
Linkage analysis (<5%)
Association analysis (>10%)

Candidate gene

'Hypothesis'-driven
Association analysis

DNA sequence

'Hypothesis'-driven
Next generation sequencing
whole genome/exome

Genetic variation

From a single polymorphism to genome wide analysis

DNA sequence variation (inherited) does not change During life disease etc.

Single Nucleotide Polymorphisms
APO E (apolioprotein E gene) chromosome 19 exon 4

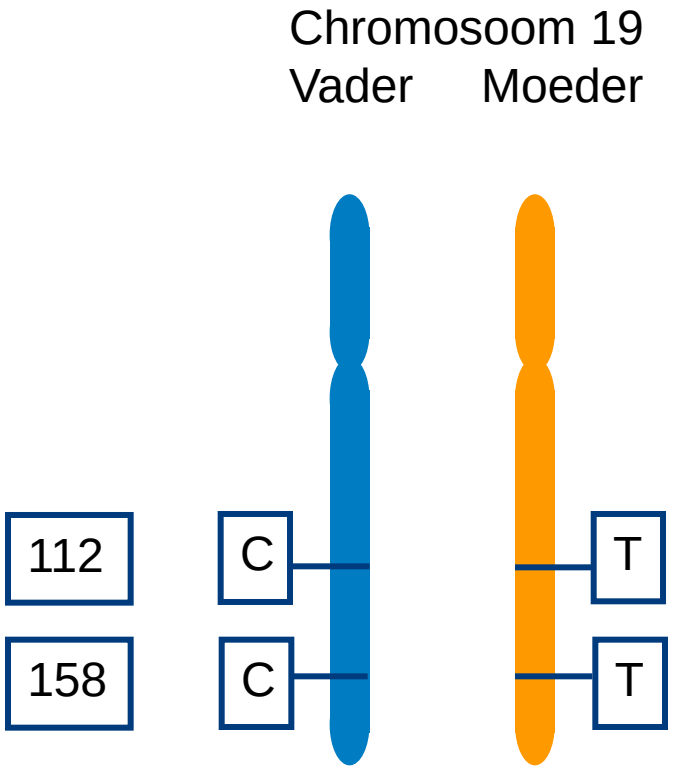
SNP = Single Nucleotide Polymorphism

Position 112 : allele C and T
Position 158 : allele C and T

Genotype 112 : CT

Haplotype 112-158 Father : C-C

Haplotype 112-158 Mother: T-T

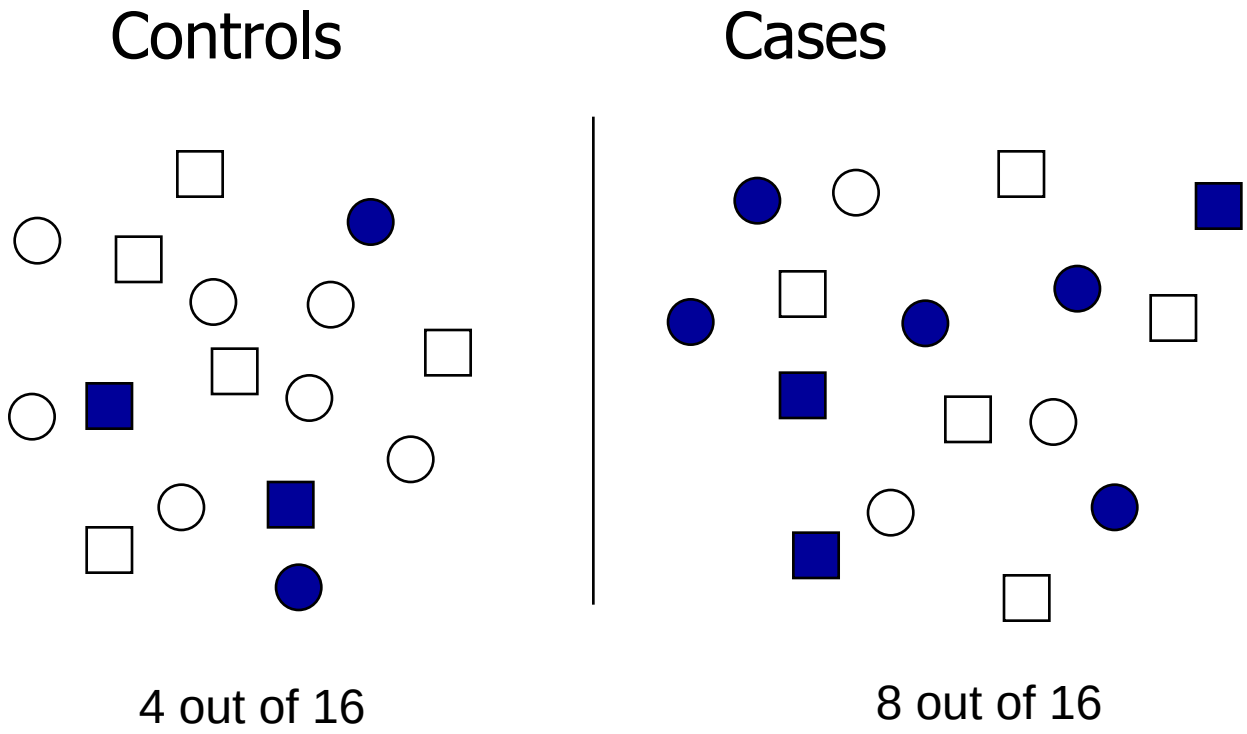


6 billion basepairs

150 million polymorphic positions In the human genome known

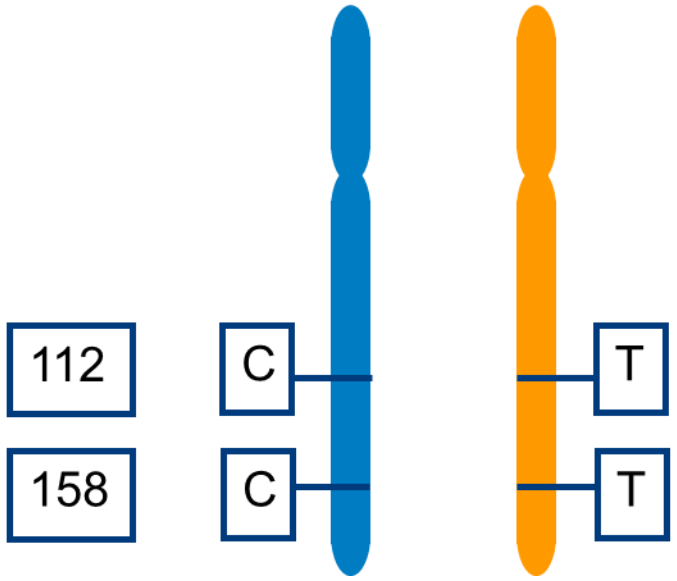
Genetic association study : Unrelated subjects

Genotype frequency in cases versus controls
Allele 2 is risk allele; count 12 genotypes



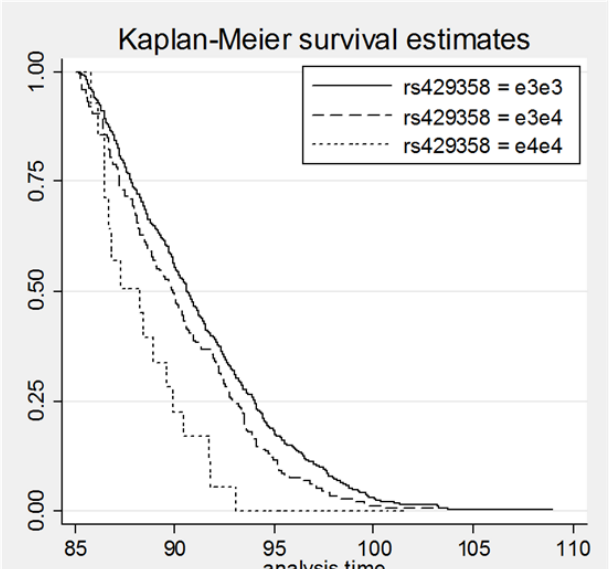
FUNCTIONAL VARIATION
APOEε2,3,4 locus

Haplotype
- APOEƐ2 : T-T
- APOEƐ3 : T-C
- APOEƐ4 : C-C



Ɛ2 : 8% is carrier ;
Associates with longevity

Ɛ4: 14% is carrier;
Associates with higher mortality risk and dementia
Homozygotes 15 times
Increased risk of dementia



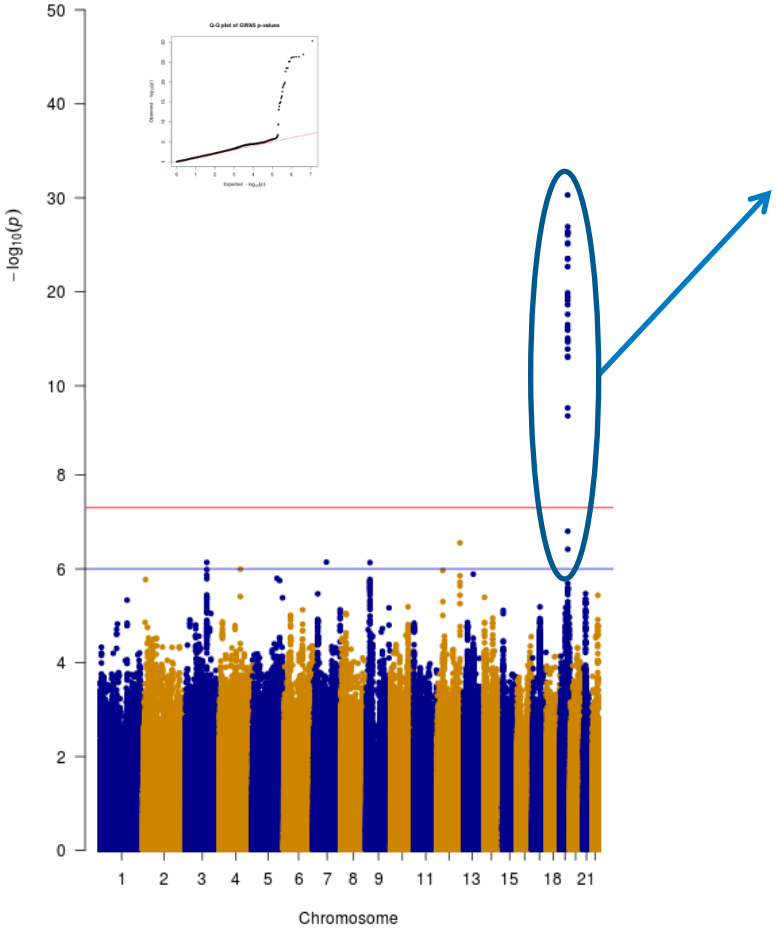
APOE in Leiden 85+ Study
15 years to follow up

Genome wide association study (GWAS) : search for genomic positions involved in the trait; then find the causal variant

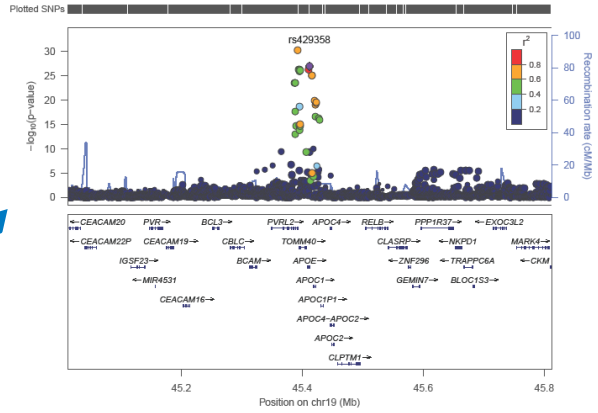
Added value of genome wide association studies (GWAS; scan million markers); also problems ?

Genome Wide Scan

Longevity : Finding the APOE locus in a GWAS



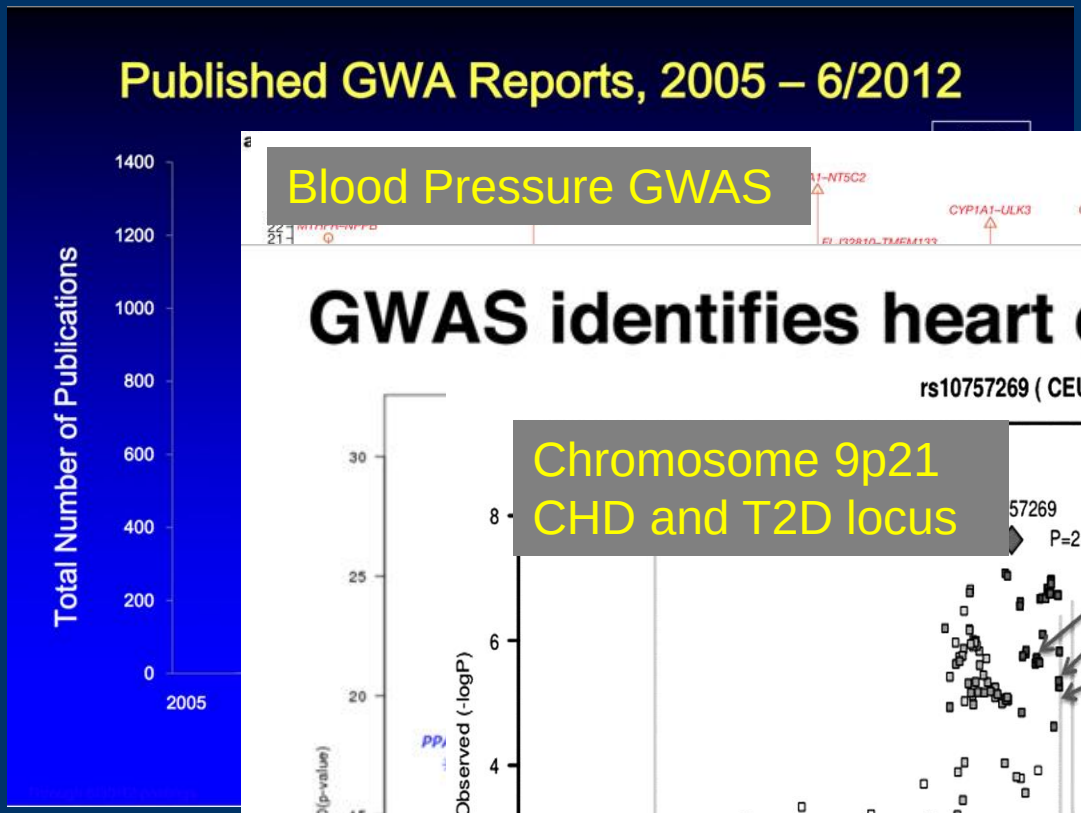
Manhattan plot



Interpretation of high dimensional data; Multiple testing

P 0.05 / no of tests (million) (Bonferroni correction

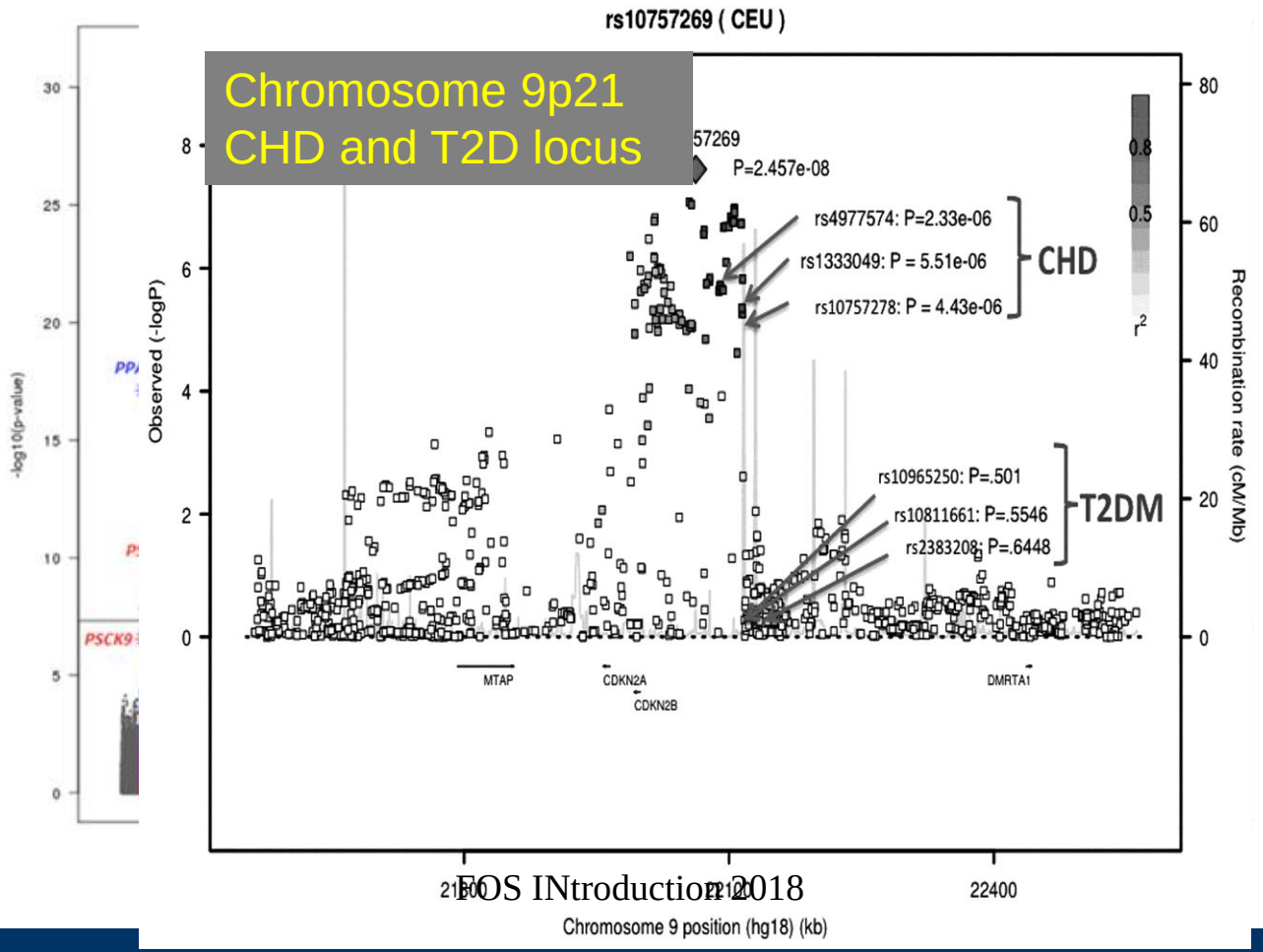
GWAS to find the location of disease susceptibility genes



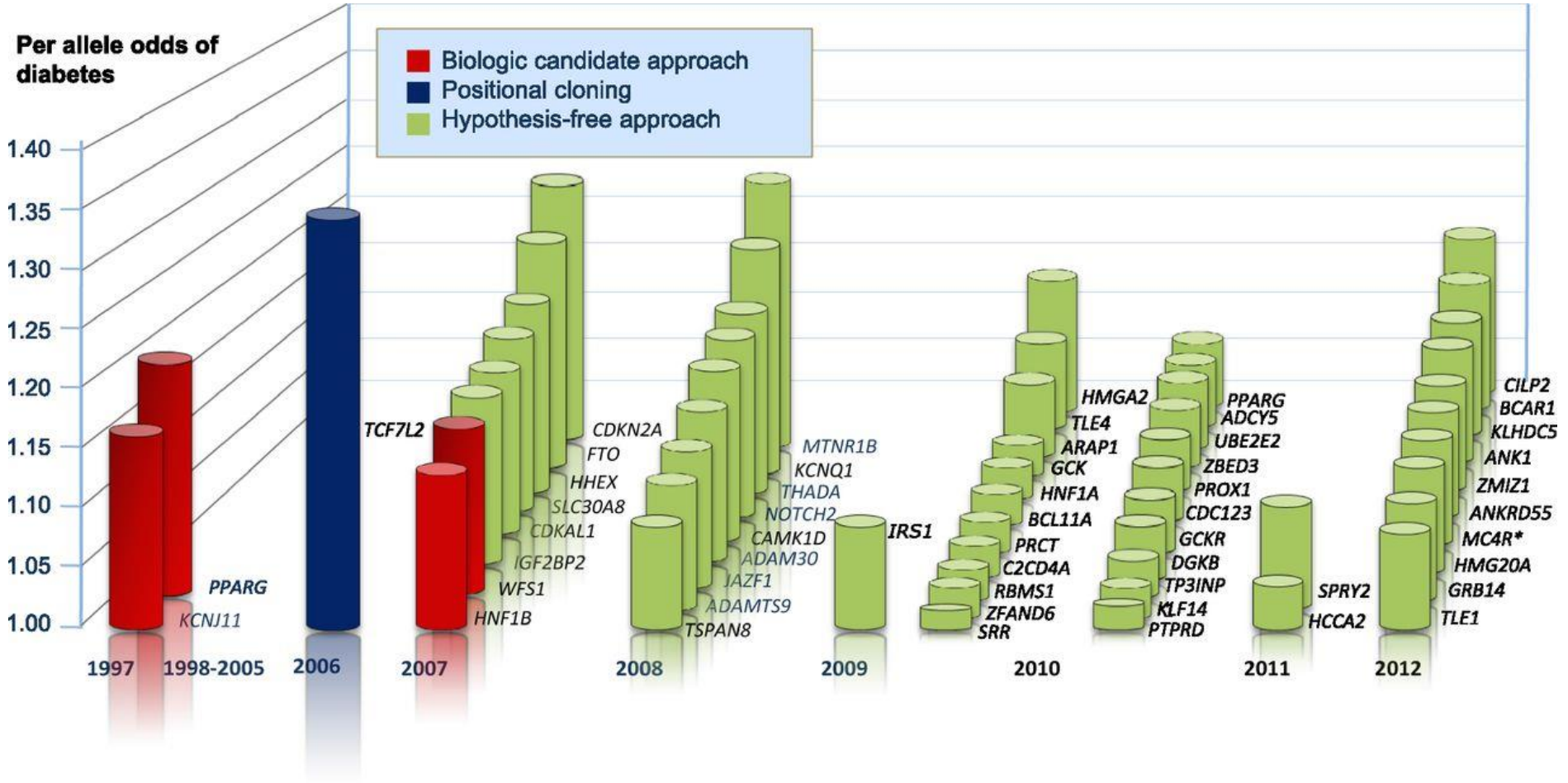
Blood Pressure GWAS

GWAS identifies heart disease factors

Chromosome 9p21
CHD and T2D locus



Type 2 Diabetes loci identified with different strategies
Take joint effect → Polygenetic Risk Score (PGRS or GRS)



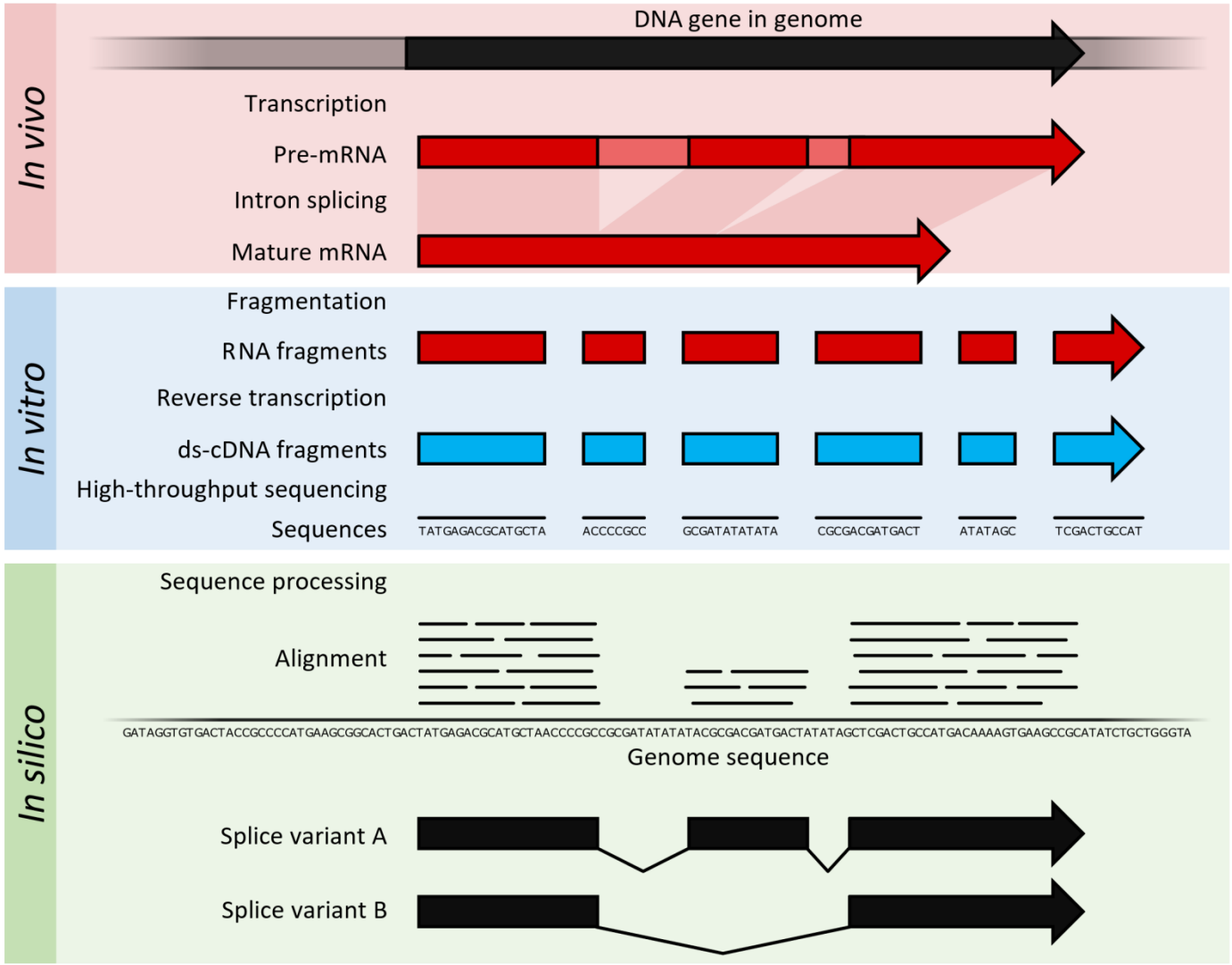
Gene expression variation (the transcriptome)

Expression of a single gene,
To 20.000 coding genes (coding for proteins)
(2% of the genome)

Non coding genes: 22 K
Transcripts : 98 K

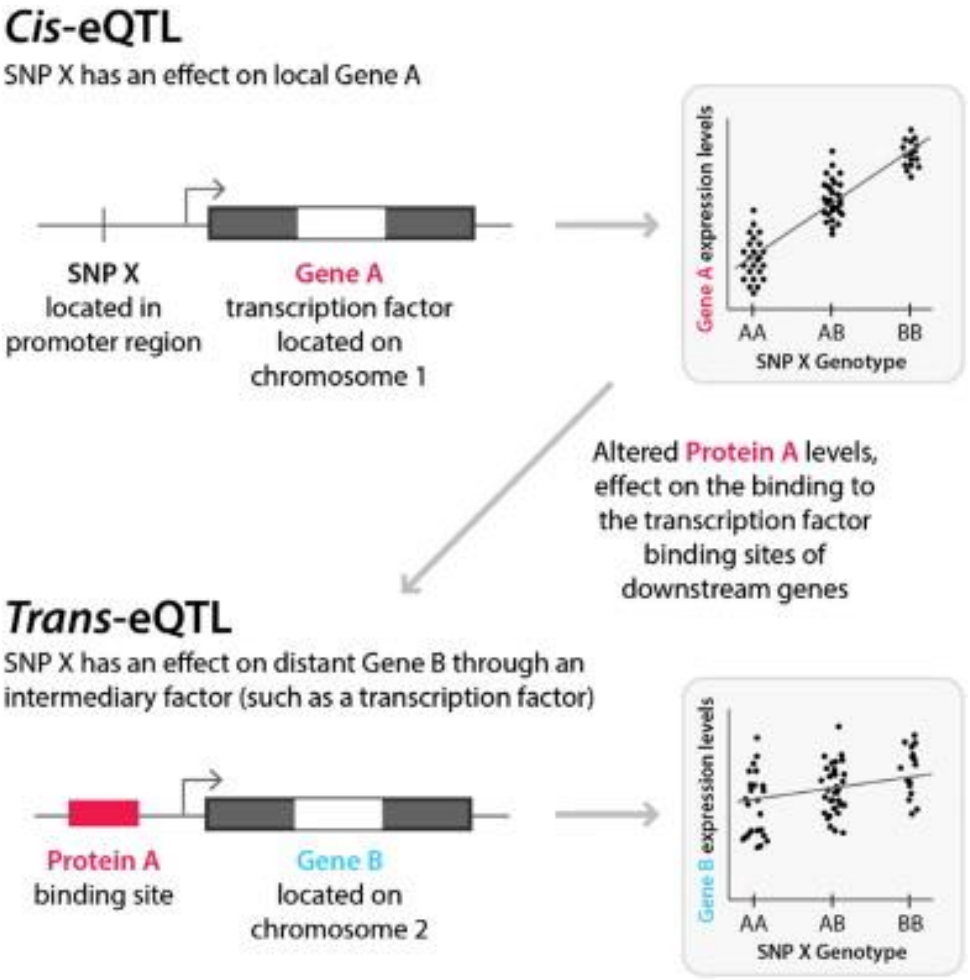
Quantitative data

Gene expression.
Quantitative data (20.000 genes, different expression levels

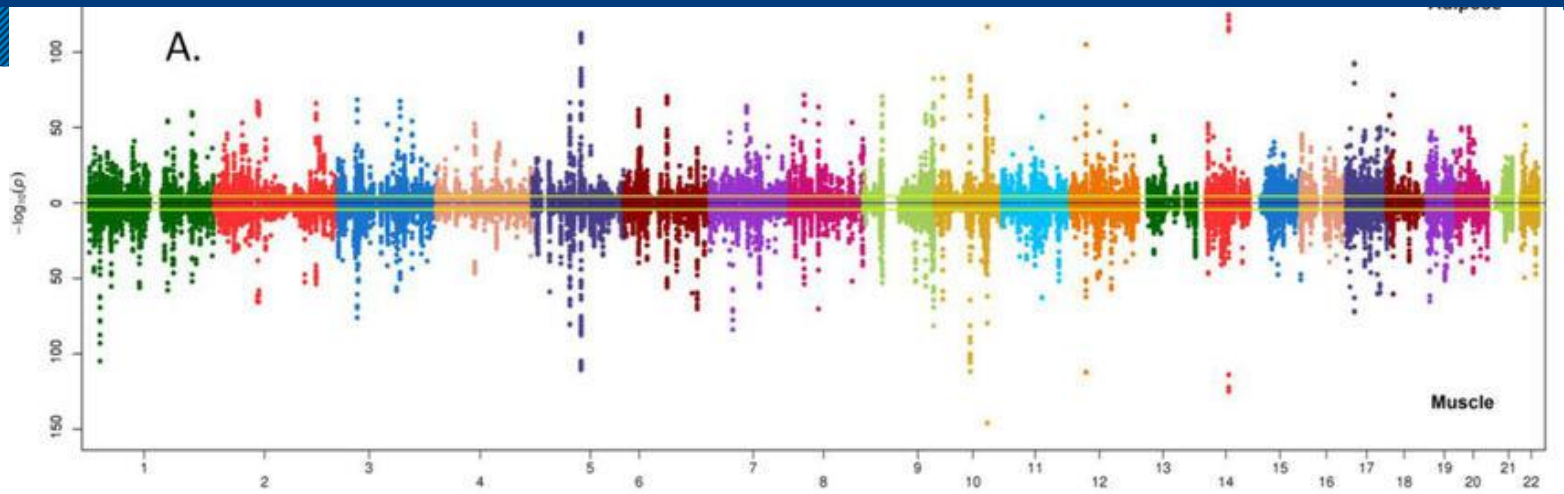


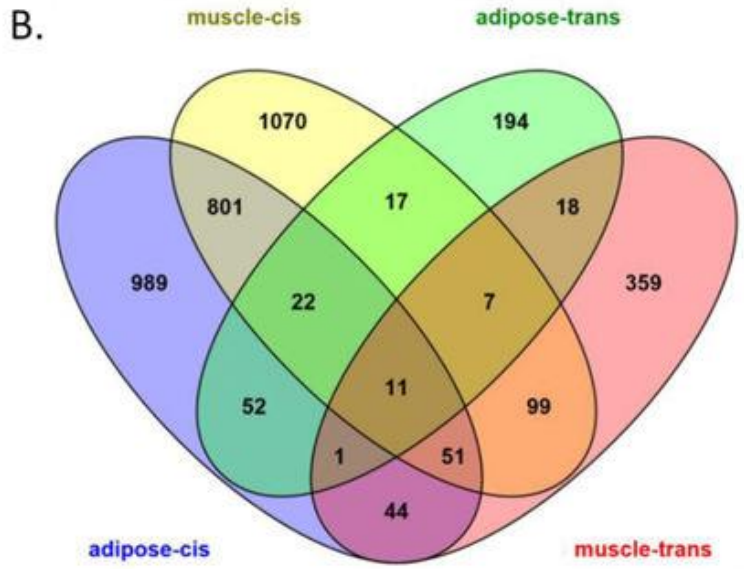
Genetic variation X quantitative variable (Quantitative Trait Loci (QTL) Quantitative data (20.000 genes, expression levels)

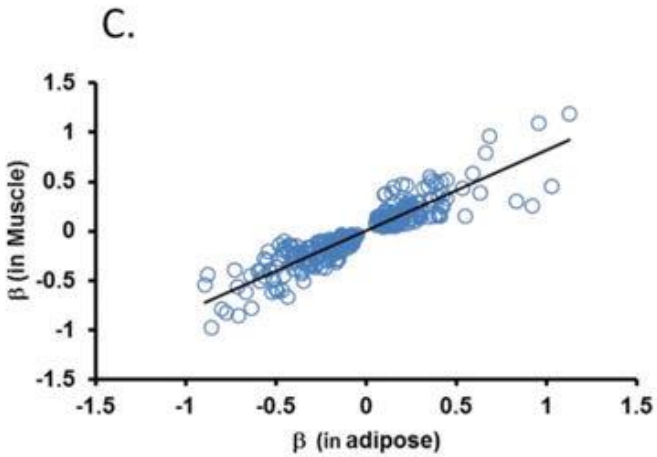
Expression QTLs (eQTL)



Scan the genome for SNPs/alleles influencing gene expression



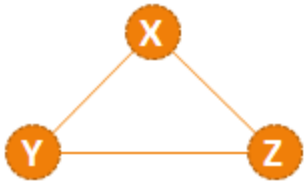




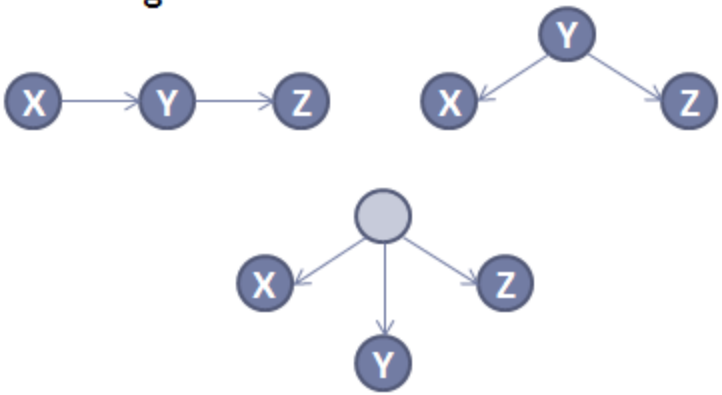
Expression QTLs (eQTL) in muscle and fat

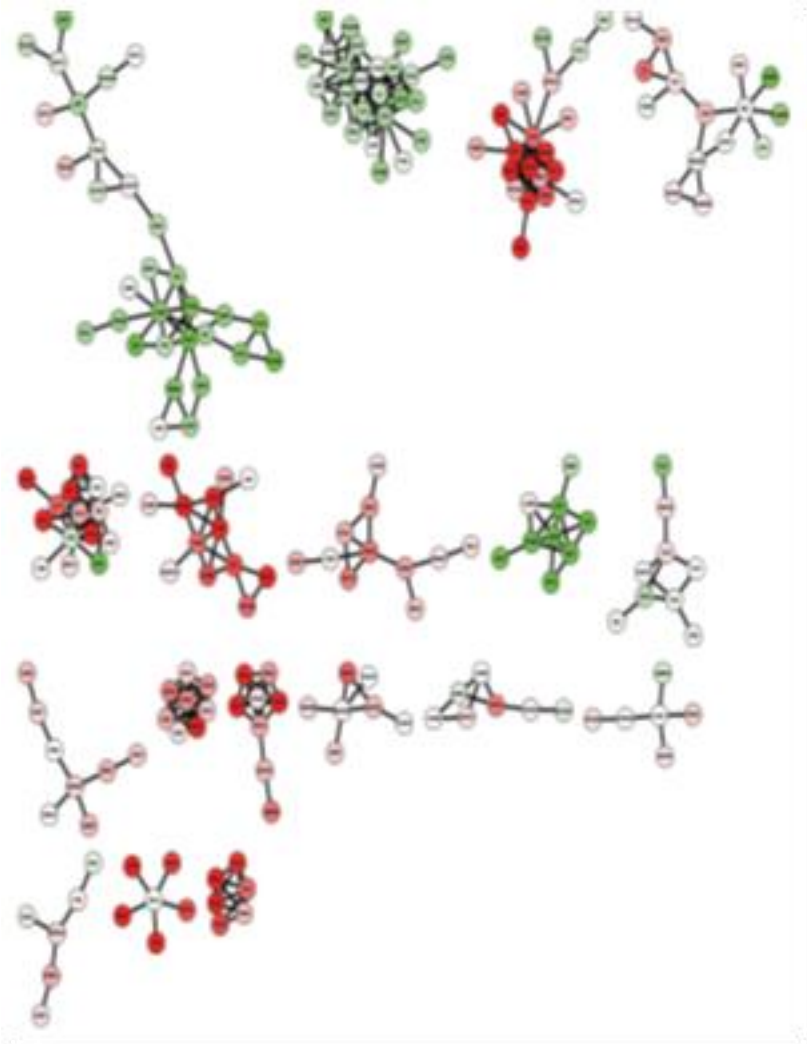
Gene expression.
Correlation in the data. Biology. Co-expression networks.

Gene Co-expression



Gene Regulation

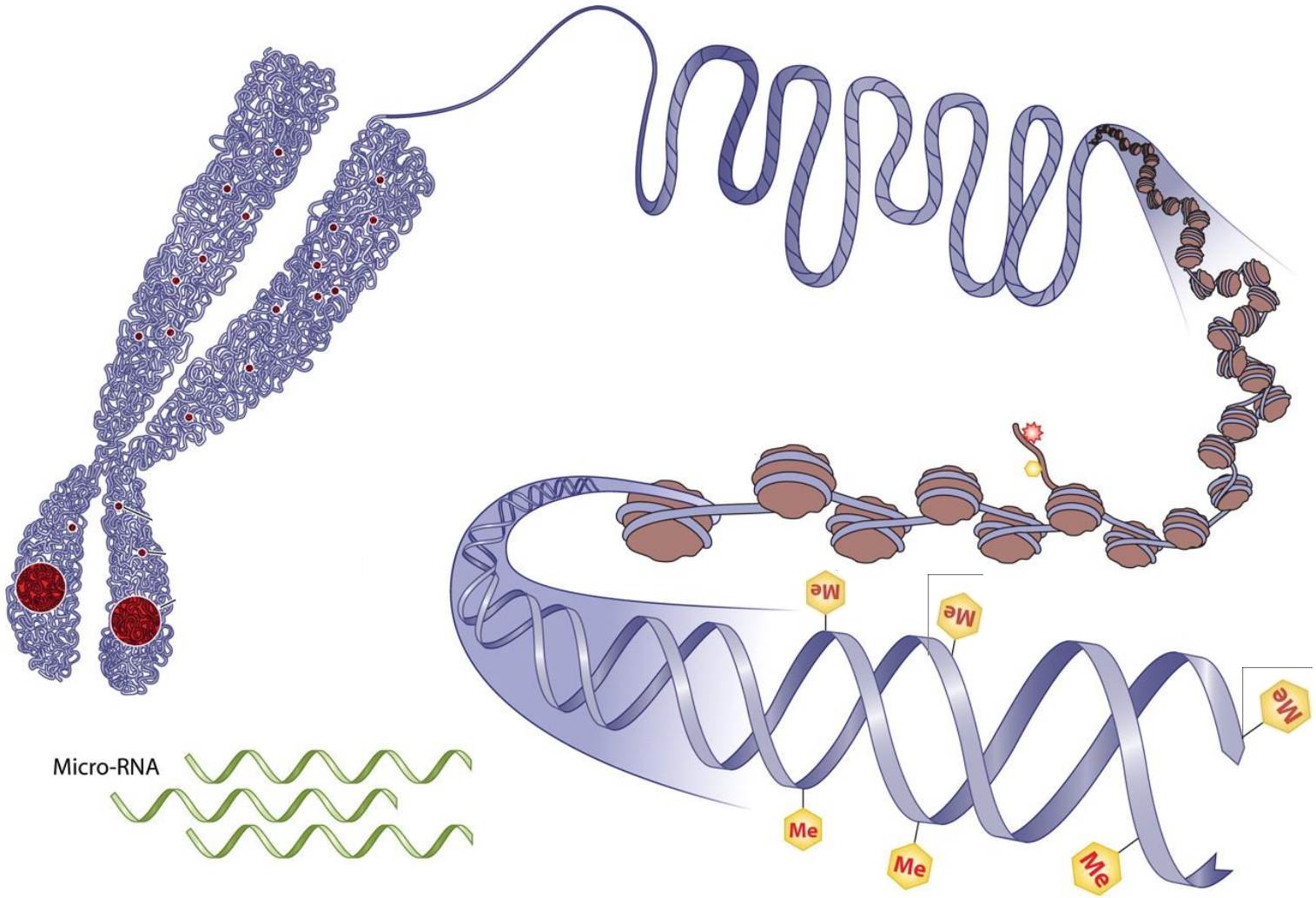




Effects of the environment: early in development, adulthood and late in life

Added value of Epigenome

Epigenome



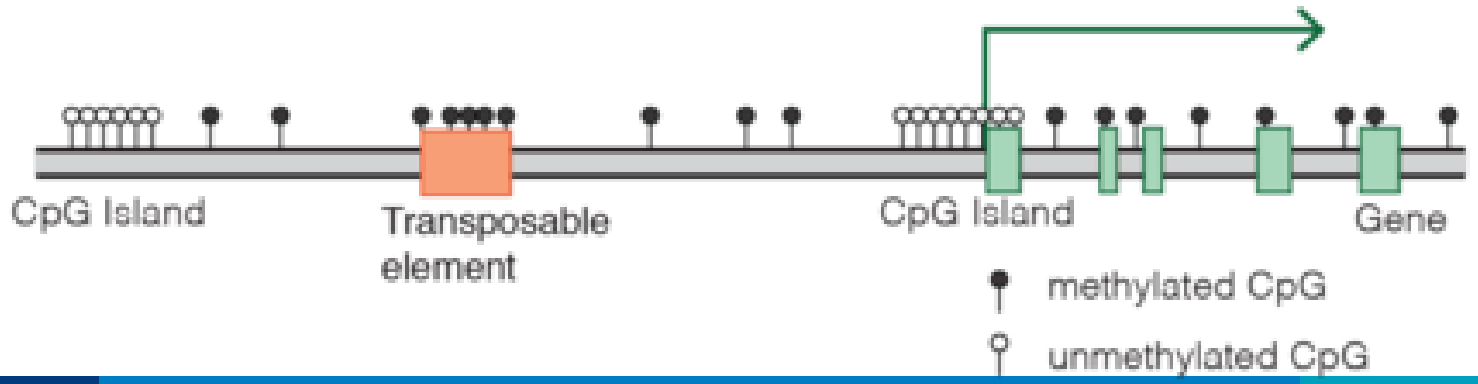
Epigenome

Most population studies into DNA methylation or non coding RNA

Smaller studies into chromatin configuration

Quantitative data
Genome wide DNA methylation:
27.000 to 85.000 CpG sites now

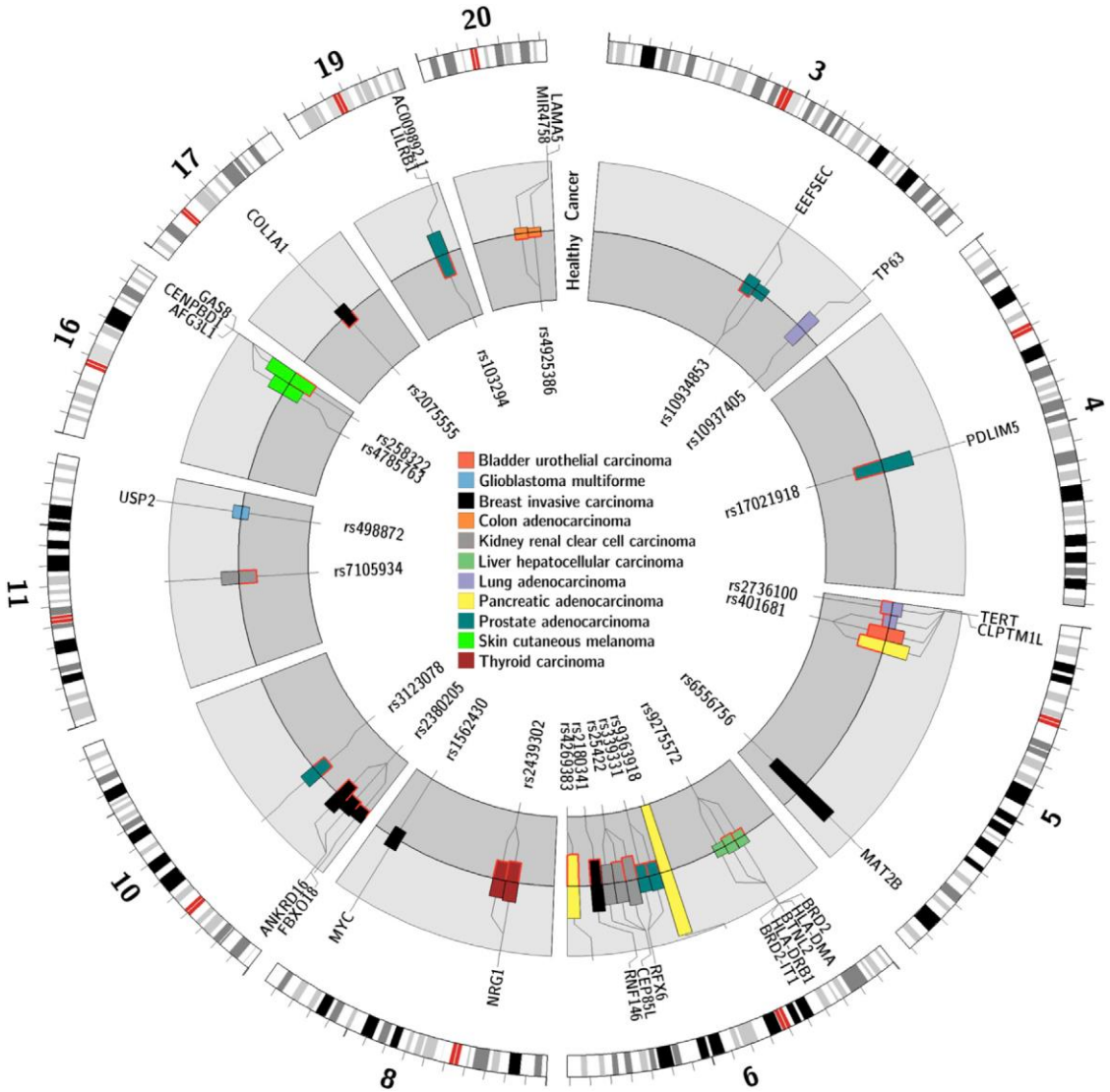
Typical mammalian DNA methylation landscape

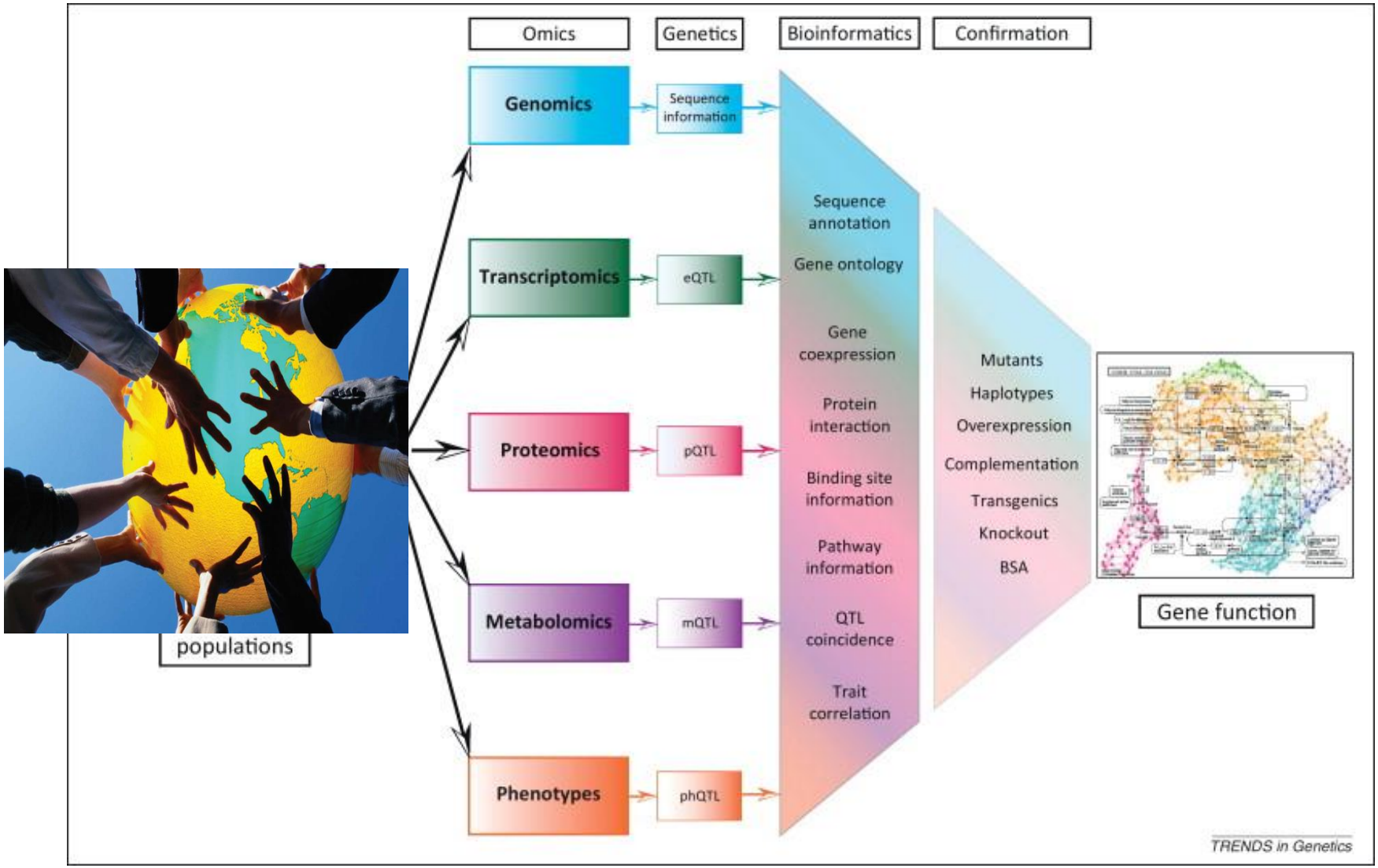


Why is high dimensional molecular data complex ?

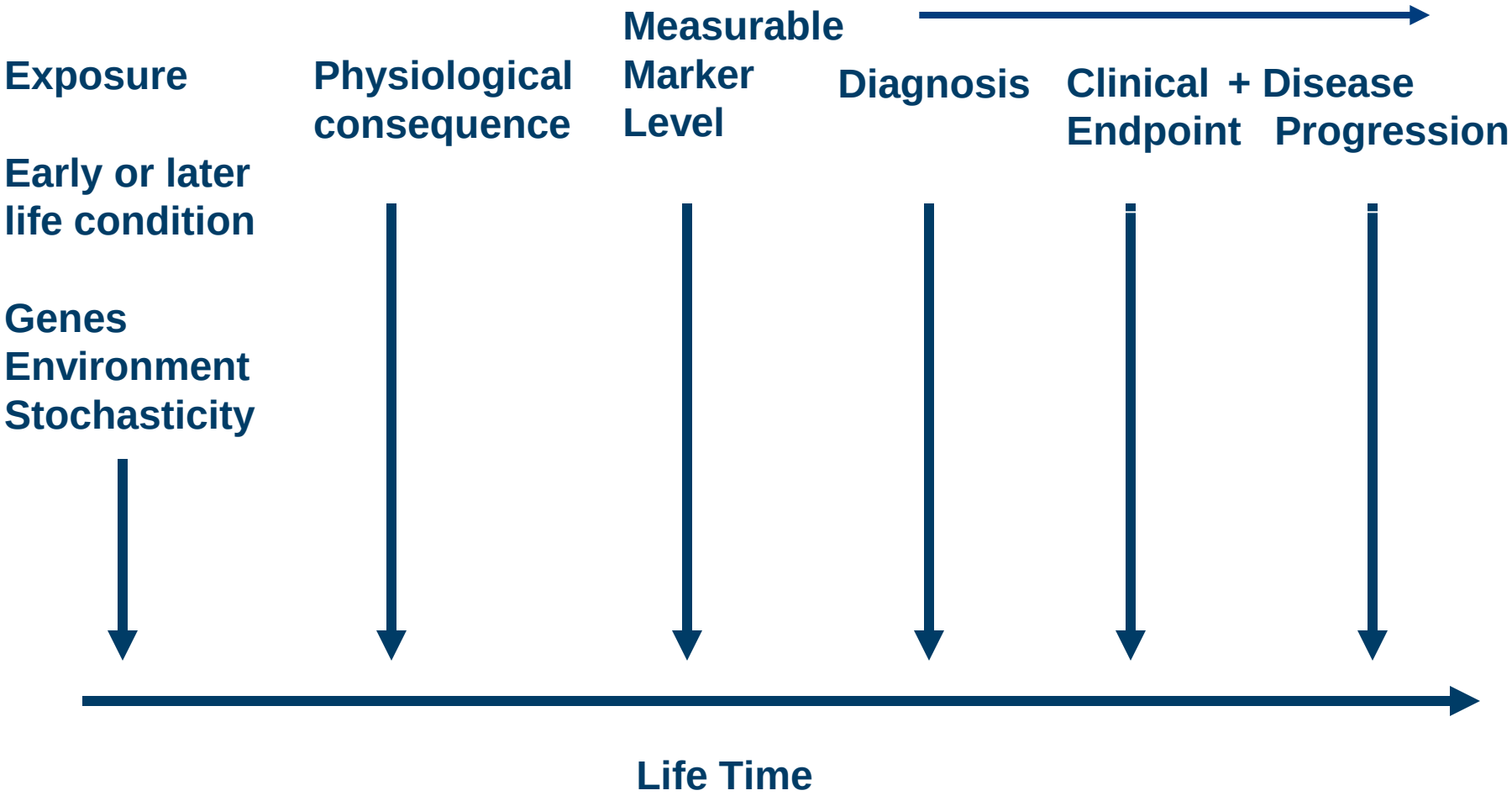
1. Multiple testing
2. Variables not independent
3. Correlation in the data
4. Distribution
5. Visualization of results (associations)

Genetic variation x DNA methylation
DNA methylation QTLs of cancer risk SNPs




Omics
Genetics
Bioinformatics
Confirmation
Genomics
Transcriptomics
Proteomics
Metabolomics
Phenotypes
Sequence information
eQTL
pQTL
mQTL
phQTL
Sequence annotation
Gene ontology
Gene coexpression
Protein interaction
Binding site information
Pathway information
QTL coincidence
Trait correlation
Mutants
Haplotypes
Overexpression
Complementation
Transgenics
Knockout
BSA
populations
Gene function
TRENDS in Genetics

Exposure Events in lifetime perspective

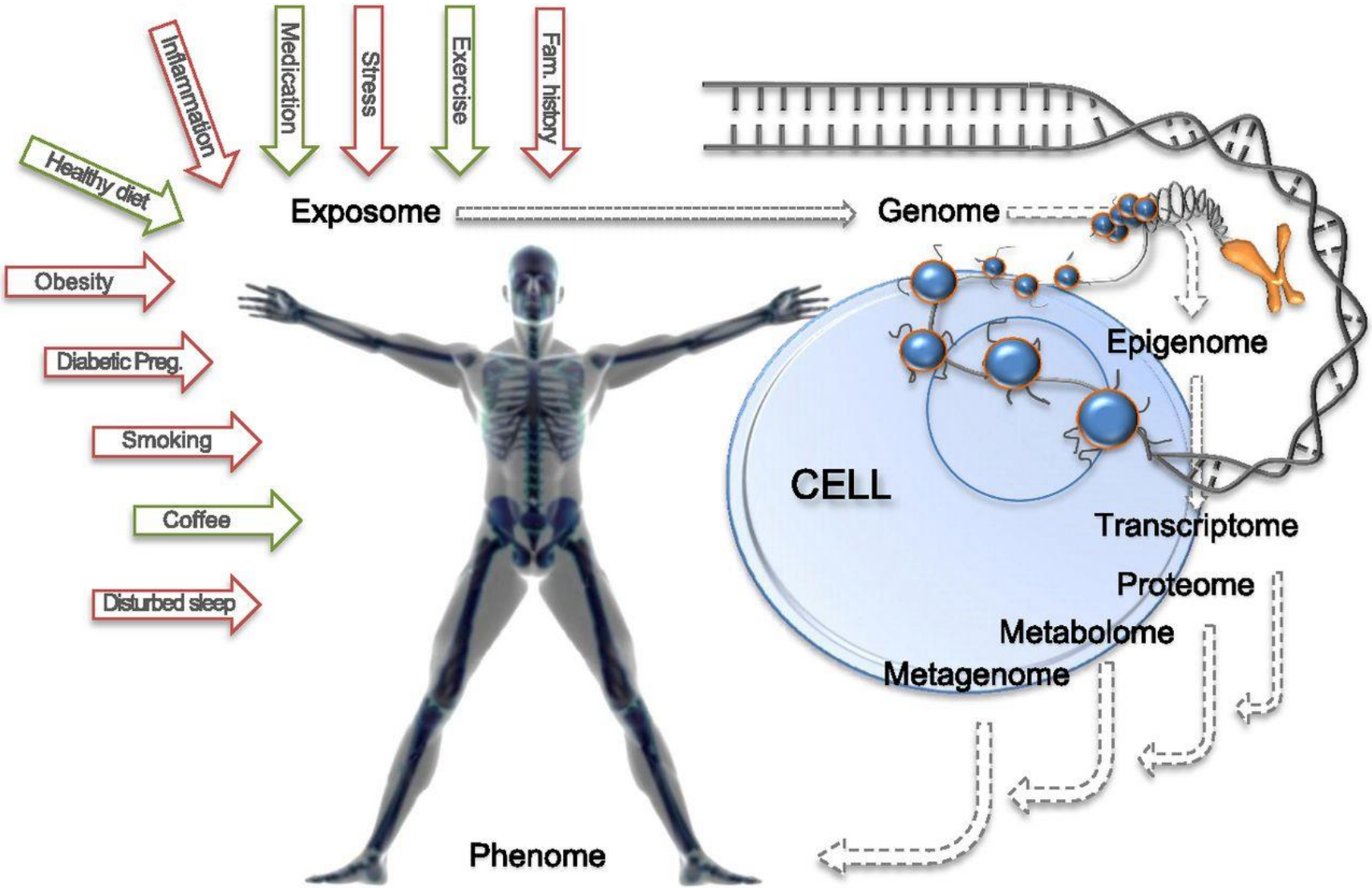


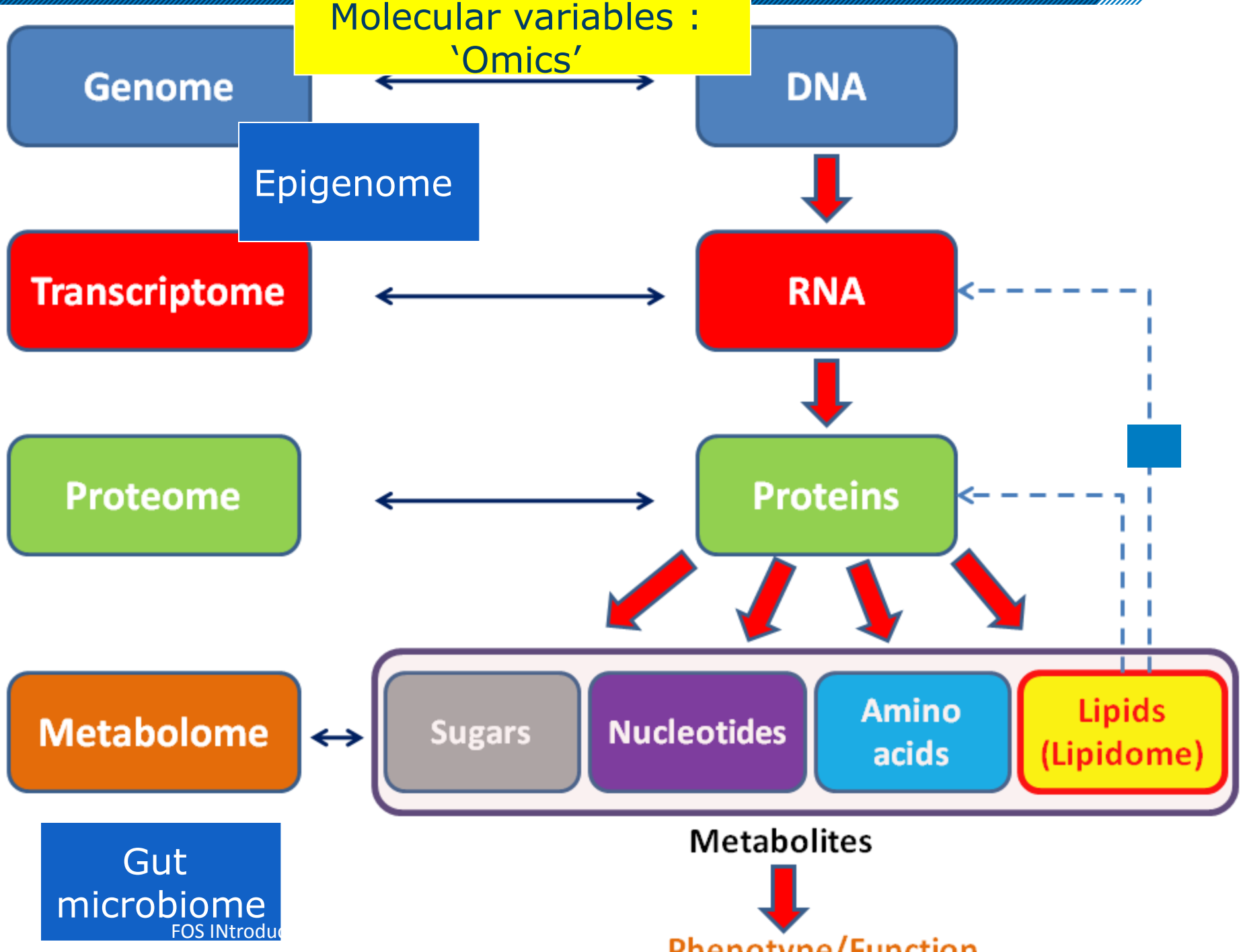
Biomarkers

- Relation of Exposure /determinant and outcome

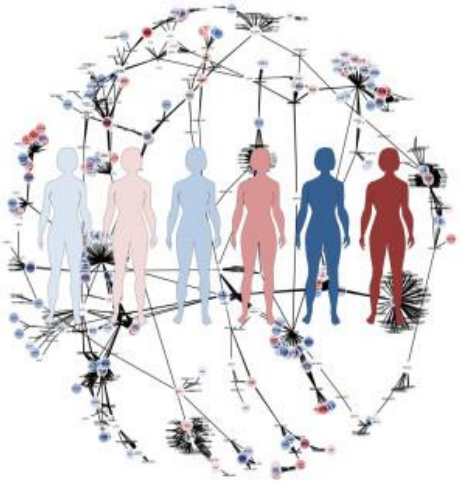
- Exposure: environment (early, late, diet, lifestyle, chemicals, geography), host (genetic background, age), health change over time (disease, biological ageing process)

- Biomarkers: a substance or biological structure that can be measured in the human body and may influence, explain or predict the incidence or outcome of disease


Inflammation
Medication
Stress
Exercise
Fam. history
Healthy diet
Exposome
Genome
Obesity
Diabetic Preg.
Smoking
Coffee
Disturbed sleep
Epigenome
CELL
Transcriptome
Proteome
Metabolome
Metagenome
Phenome


Molecular variables : 'Omics'
Genome
DNA
Epigenome
Transcriptome
RNA
Proteome
Proteins
Metabolome
Sugars
Nucleotides
Amino acids
Lipids (Lipidome)
Metabolites
Gut microbiome
FOS INtrodu
Phenotype/Function

BBMRI Biobanking consortium
Multi-level omics data



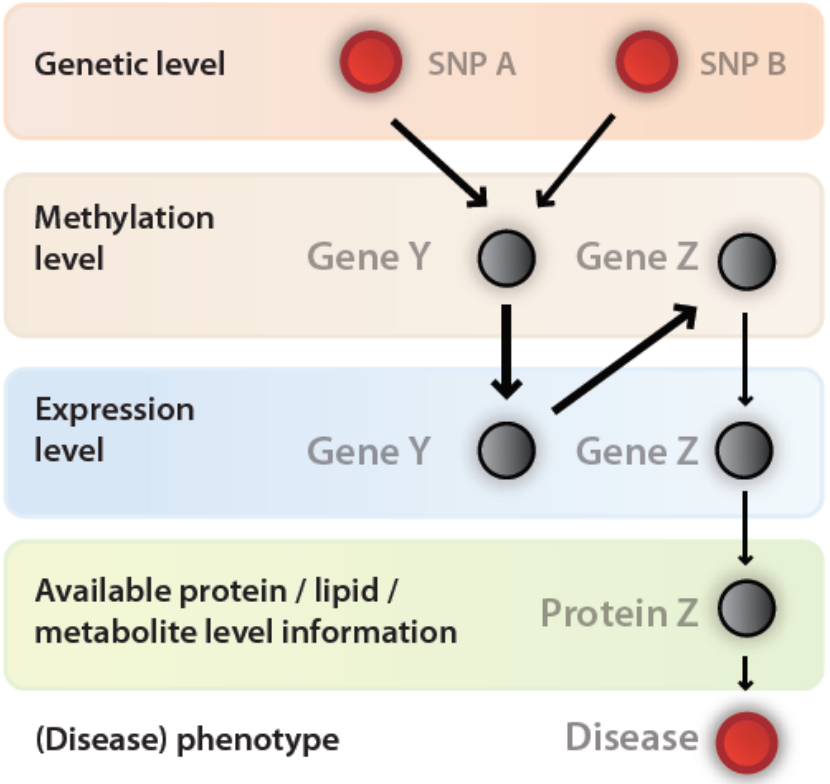
N=100,000 GWAS
N=750 Go.NlL

N=4,000

N=4,000

N=50,000

N>250,000



Go•NL
GENOMEoftheNETHERLANDS

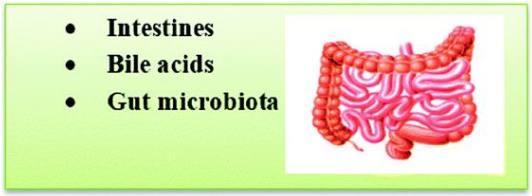
BIOS consortium

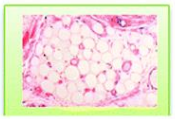
BIOS consortium

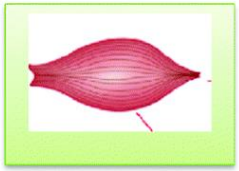
METABOLITE consortium

Central databases for the research community

Which tissue functions do the 1H NMR metabolites represent ?

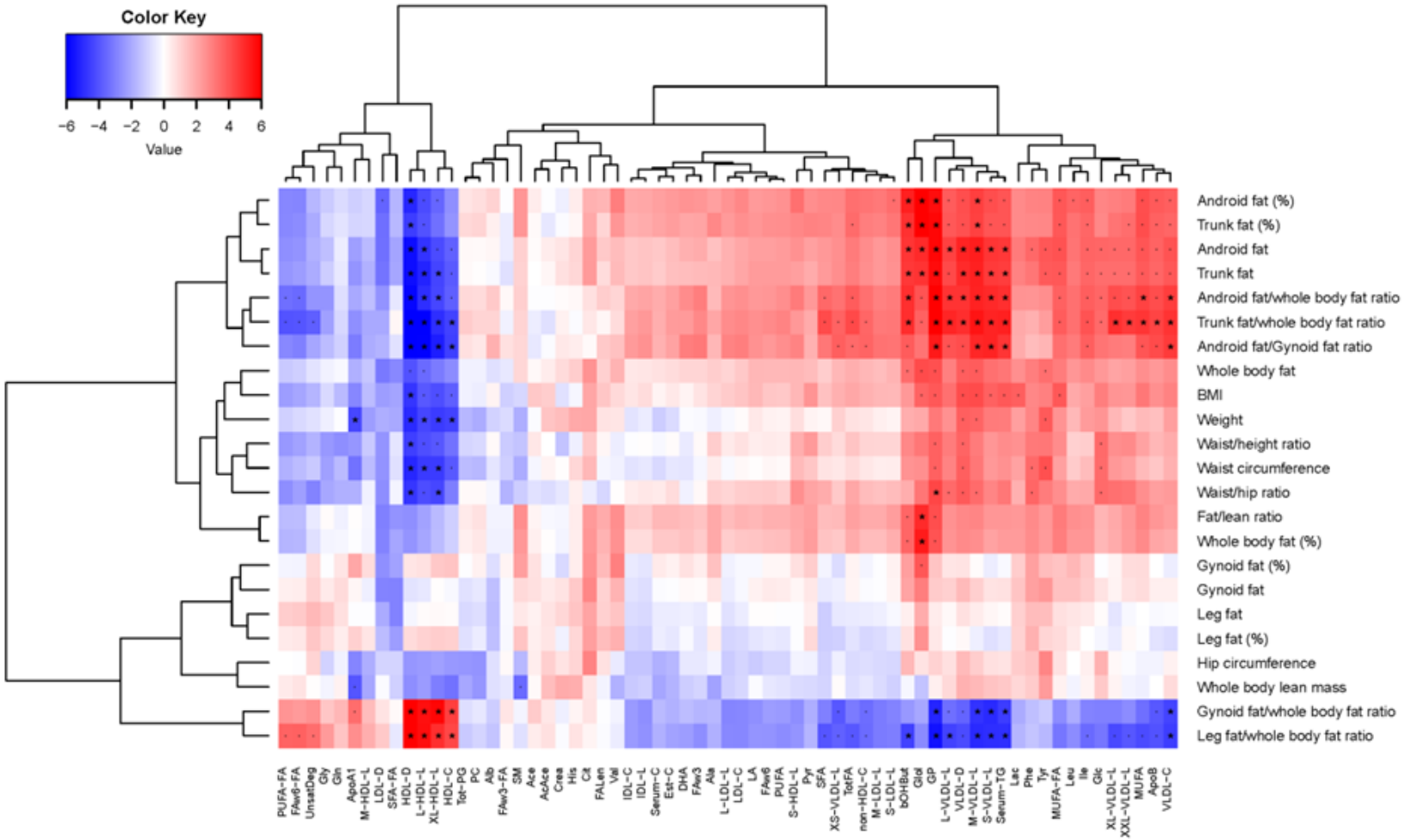




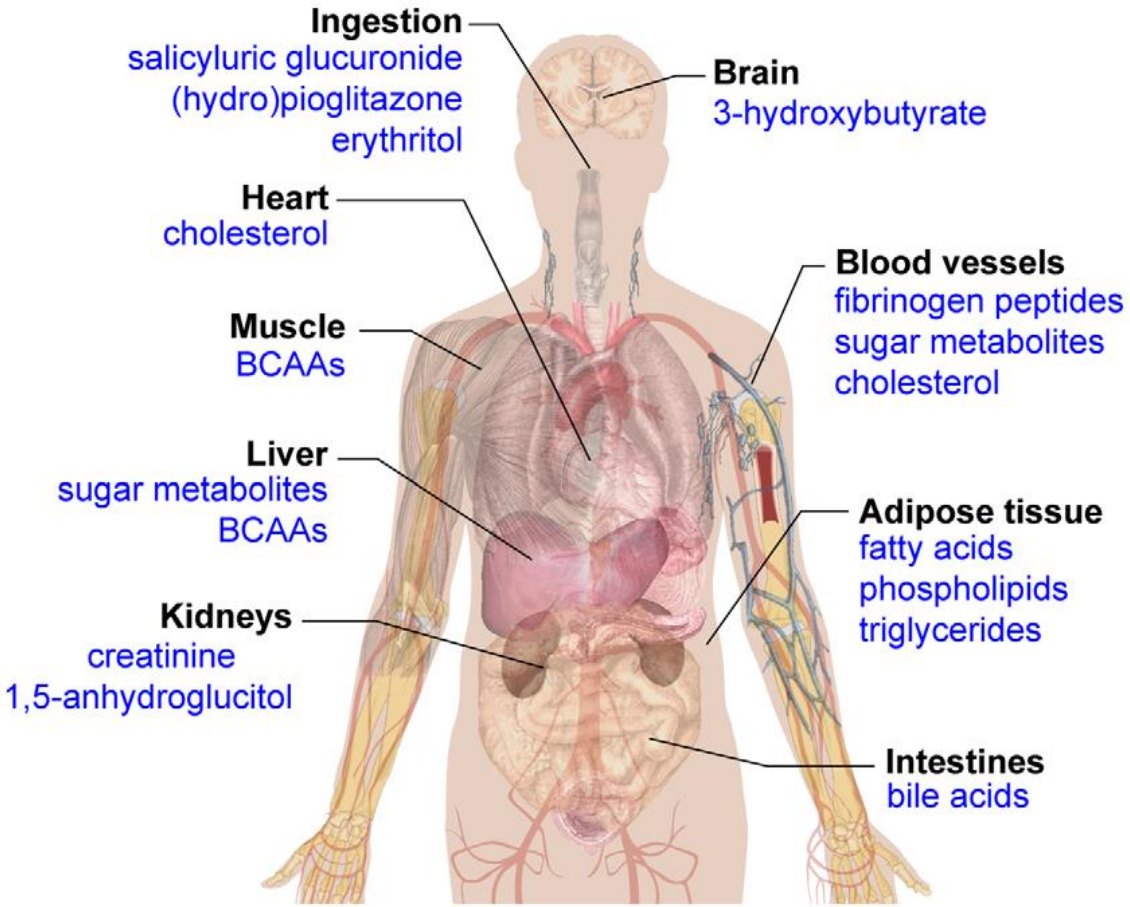


Betaine
DMG
TMAO
Choline
Phosphotidycholine
Methylamine
Dimethylamine
Phospholipids
Glycine
Creatine
Creatinine
1, 5-AG
Glycogenesis
Glycogenolysis
Glucose
Glycogen
Gluconeogenesis
Glycolysis
Carnitines
Serine
Alanine
Pyruvate
Lactate
Triglycerides
Fat Storage
B-oxidation
Fatty Acids
Acetyl-coA
Branched Chain Amino Acid (BCAA)
Lipogenesis
Acetate
β-hydroxybutyrate
Ketone Bodies
Acetoacetate
Acetone
3-HB
Oxaloacetate
Citrate
Malate
Isocitrate
TCA Cycle
Fumarate
2-Oxoglutarate
Succinate
Succinyl-coA

Correlated traits (different parameters in fat distribution) visualization (heat map)



Type II Diabetes prediction By 1H NMR metabolites



Suhre et al. PLoS ONE |November 2010 | Volume 5 | Issue 11

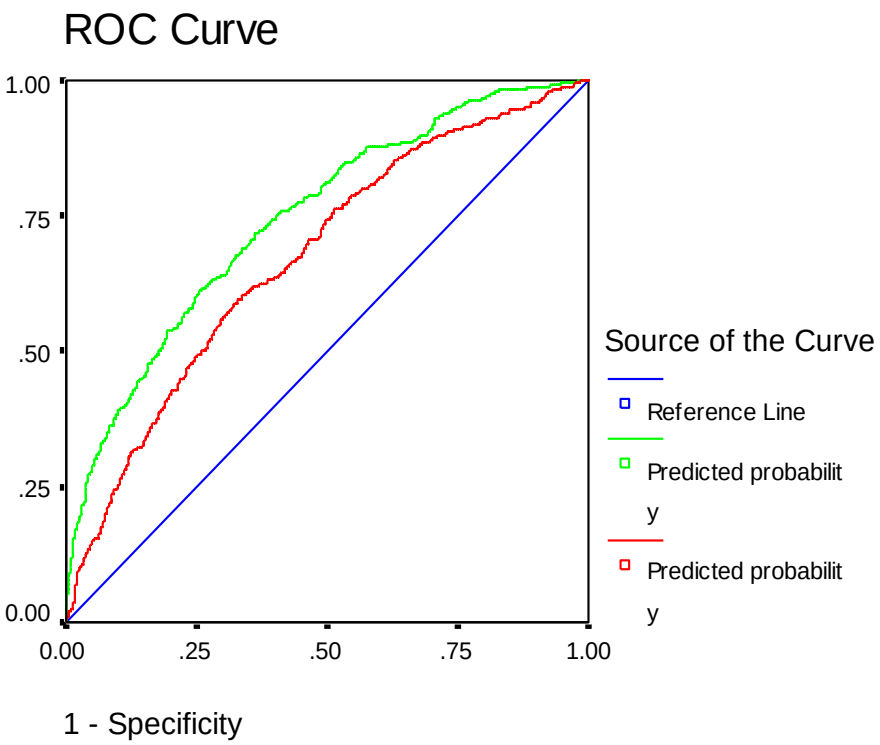
Metabolomics in Research projects

Prediction

Generate a predictor: factor identification and model development :

1. **Exploration of associations between metabolites and diverse endpoints.**
2. **Cross sectional → Prospective/longitudinal follow-up studies**
3. **Univariate (single metabolites) , multivariate**
4. **Replication in independent studies**
5. **Meta-analysis in multiple studies, create predictors (for example of mortality risk) and compare to existing predictors**

Receiver Operator Characteristic (ROC) curves to compare novel and traditional predictors (example 10 y MI risk)



Blue: 50%-50%

→ AUC=0.5

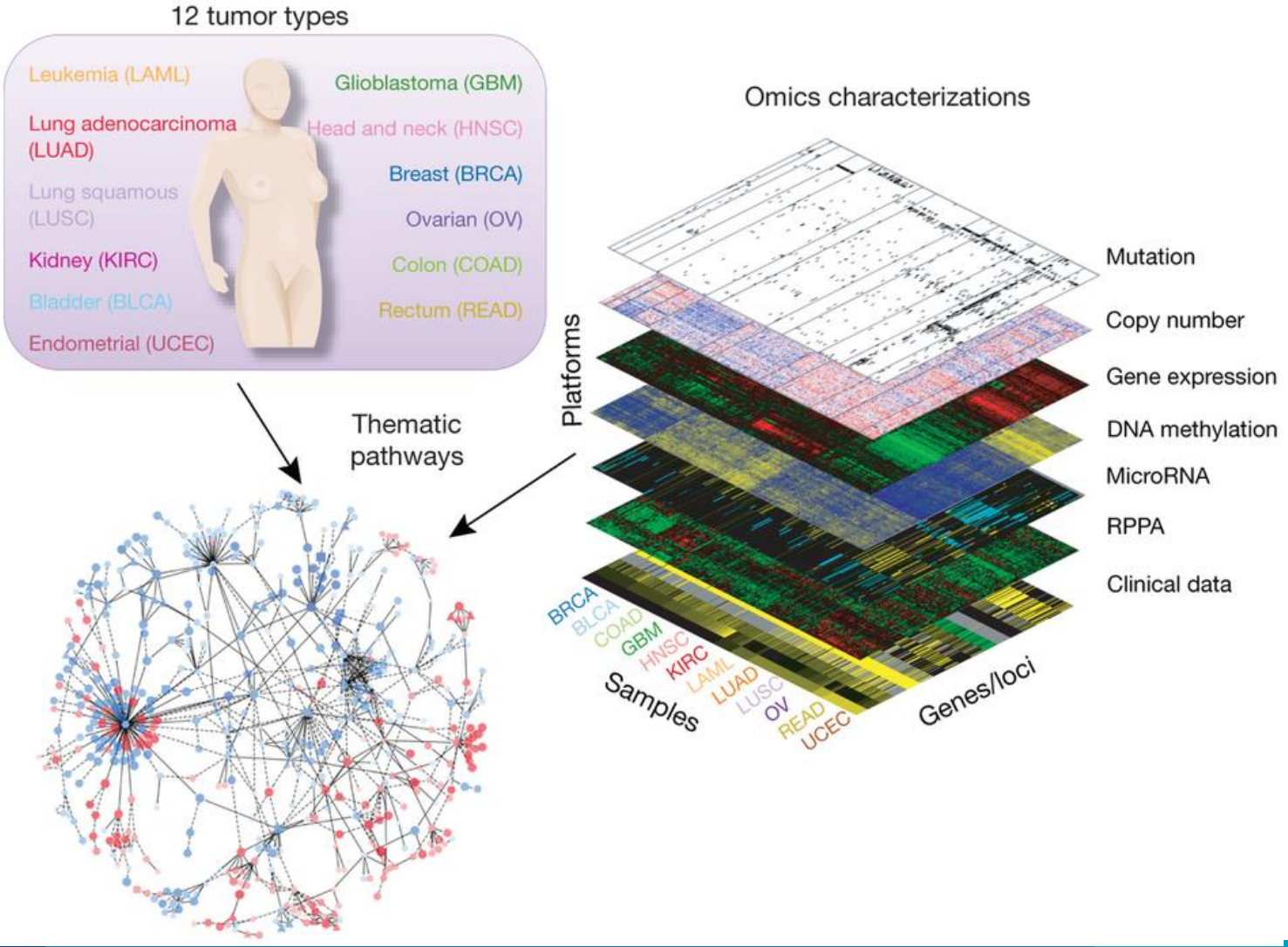
Red: model with Age , Diabetes, Smoking

→ AUC=0.67

Green: model with Age, Diabetes, Smoking HDLcholesterol and systolic blood pressure

→ AUC=0.75

Large scale omics data ; combined to classify patients and predict disease.



Stop goofing around and stand up straight!
This figure is for the grant renewal.