Introduction to omic data

Juan R González
Bioinformatics Research Group in Epidemiology

ISGlobal Instituto de Salud Global Barcelona

Institució CERCA Centres de Recerca de Catalunya

Una iniciativa de:

Obra Social "la Caixa"

Clínic Barcelona Hospital Universitari · Parc de Salut Mar · Universitat de Barcelona · Universitat Pompeu Fabra Barcelona · Generalitat de Catalunya · Gobierno de España · Ajuntament de Barcelona · Fundación Ramón Areces

OUTLINE

- Omic data
- Genomics, transcriptomics and epigenomics
- Omic databases
- Bioconductor

Omic data

OMIC

- The term ‘‘omic’’ is derived from the Latin suffix ‘‘ome’’ meaning mass or many. Thus, OMICS involve a mass (large number) of measurements/features per outcome (Jackson et al., 2006)

Integration of OMICS data

- Efficient integration of data from different OMICS can greatly facilitate the discovery of true causes and states of disease, mostly done by softwares (Andrew et al., 2006).

Omic data

- In biological context , suffix –**omics** is used to refer to the study of **large sets of biological molecules** (Smith et al., 2005)
- The realization that DNA is not alone regulate complex biological processes (as a result of HGP, 2001), triggered the rapid development of **several fields in molecular biology** that together are described with the term OMICS.
- The OMICS field ranges from
  - Genomics (focused on the genome)
  - Proteomics (focused on large sets of proteins, the proteome)
  - Metabolomics (focused on large sets of small molecules, the metabolome).

Omic data

The field of genomics has been divided into 3 major categories:

- **Genotyping** (focused on the genome sequence): the physiological function of genes and the elucidation of the role of specific genes in disease susceptibility (Syvanen, 2001) [Structural genomics]
- **Transcriptomics** (focused on genomic expression): The abundance of specific mRNA transcripts in a biological sample is a reflection of the expression levels of the corresponding genes (Manning et al., 2007) [Functional genomics]
- **Epigenomics** (focused on epigenetic regulation of genome expression): Study of epigenetic processes (expression activities not involving DNA) on a large (ultimately genome-wide) scale (Feinberg, 2007) [Functional genomics]

Genotyping

Goal

- Identification of the physiological function of genes
- Role of specific genes in disease susceptibility (syvanen et al., 2001)

Common feature used

- Among different variations (insertions, deletions, SNPs, etc.), **single nucleotide polymorphisms** (SNPs) are the most commonly investigated (Sachidanandam et al., 2001) and can be used as markers for diseases.
- Tag SNPs (informative subset of SNPs) and fine mapping are further used to identify true cause of phenotype (patil et al., 2001).

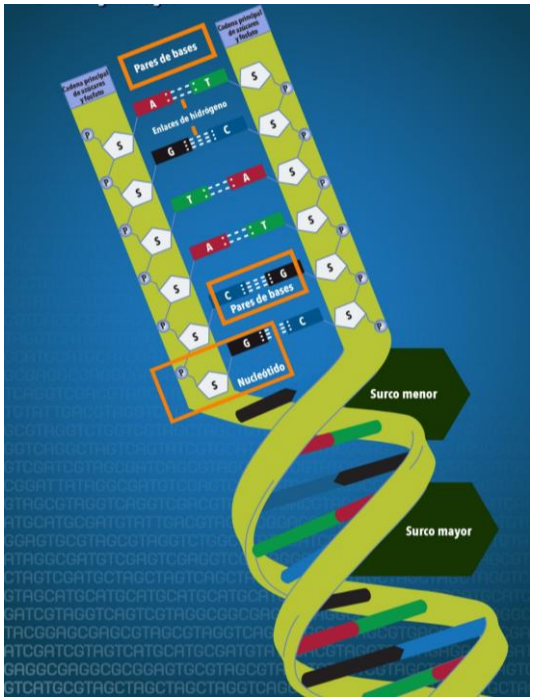
Application

- Identification of genes associated with disease

Recent improvement in genotyping

- **Array-based genotyping** allows the assessment of entire genome (up to 1M SNPs) per assay: GWAS (Jelly et al., 2010)
- Next Generation Sequencing (**WES**, **WGS**)

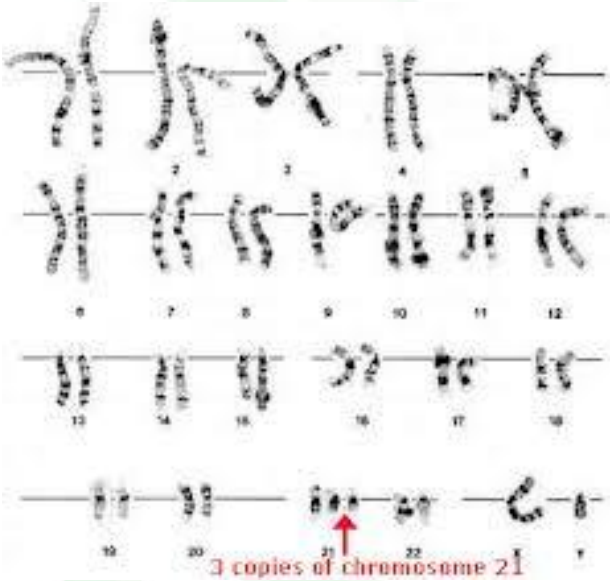
Genotyping



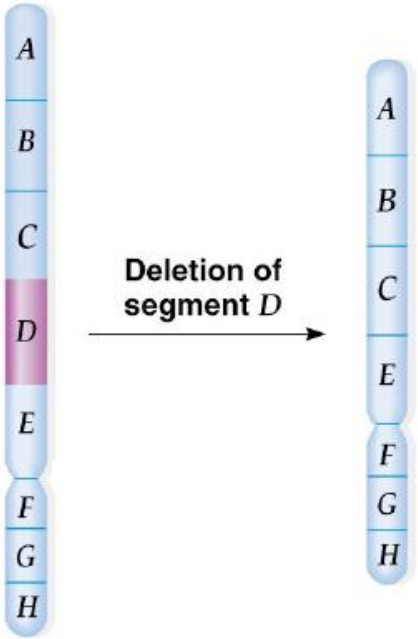
- DNA = deoxyribonucleic acid
- Two antiparallele strands: double helix
- Four nucleotides:
  - Adenine = Timine
  - Cytosine = Guanine
- Human genome size: 3.2 billion bp
- Known variants: 324 M variants (2017)
- Difference between 2 individuals: 20 M bp (0.6%)

Genotyping

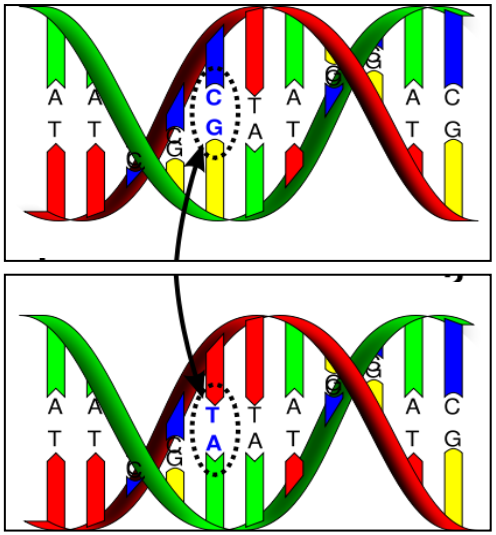
Chromosomal variants



Structural variants



Single nucleotide polymorphism (SNP)



Genotyping

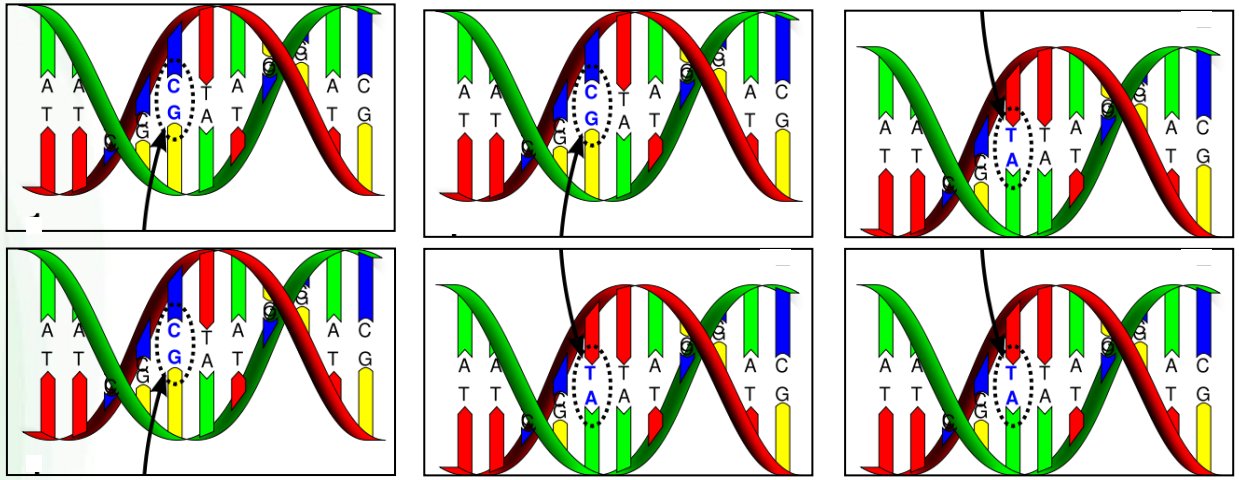
Subject
Homozygote CC

Subject 2
Heterozygote CT

Subject 3
Homozygote TT

Maternal chromosome

Paternal chromosome



Genotype vs allele
Allele: different forms a genetic variant can take

Locus/Loci

Human Reference Genome

- A single consensus representation of the genome

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

Assembly (versions):

- Human genome 19 (hg19)=build 37 (bd37)=GRCh37
- Human genome 38 (hg38)=GRCh38

Strand:

- genome assembly strand (plus (+) or minus (-))
- dbSNP strand (forward (F) or reverse (R))
- Illumina (top (T) or bottom (B))

Human genome browsers

- System for displaying, viewing and accessing genome annotation data
- NCBI (National Center for Biotechnology Information): http://www.ncbi.nlm.nih.gov/
- ENSEMBL (a joint project between EMBL (European Molecular Biology Laboratory)-EBI and the Wellcome Trust Sanger Institute): http://www.ensembl.org/index.html
- UCSC (Universtiy California Santa Cruz) Genome Bioinformatics: http://genome.ucsc.edu/

1000 Genomes Project

ARTICLE

doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

Genetic variants database: SNPs

- dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP
- ENSEMBL: http://www.ensembl.org/info/website/tutorials/gene_snps.html

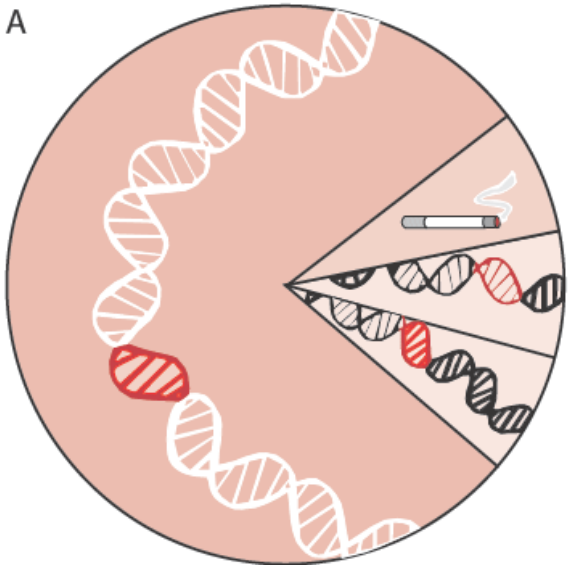
SNPs have a unique identifier “rs#”

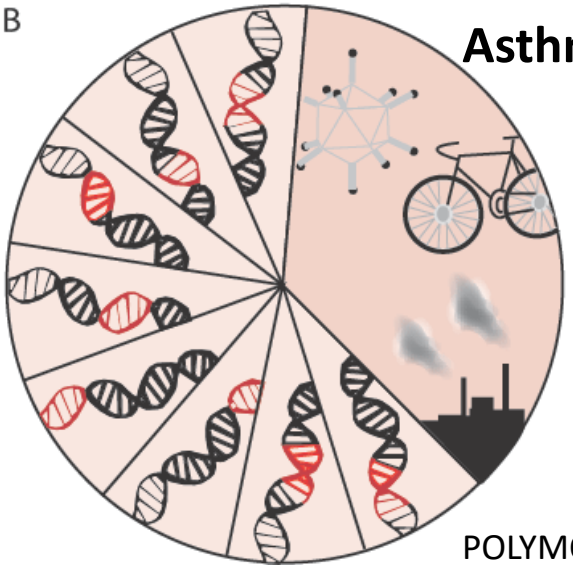
Ex. rs1695

Mendelian vs complex diseases

Phenylketonuria

- mutations in PAH gene
- decreased metabolism of phenylalanine
- treatment: food with low levels of phenylalanine



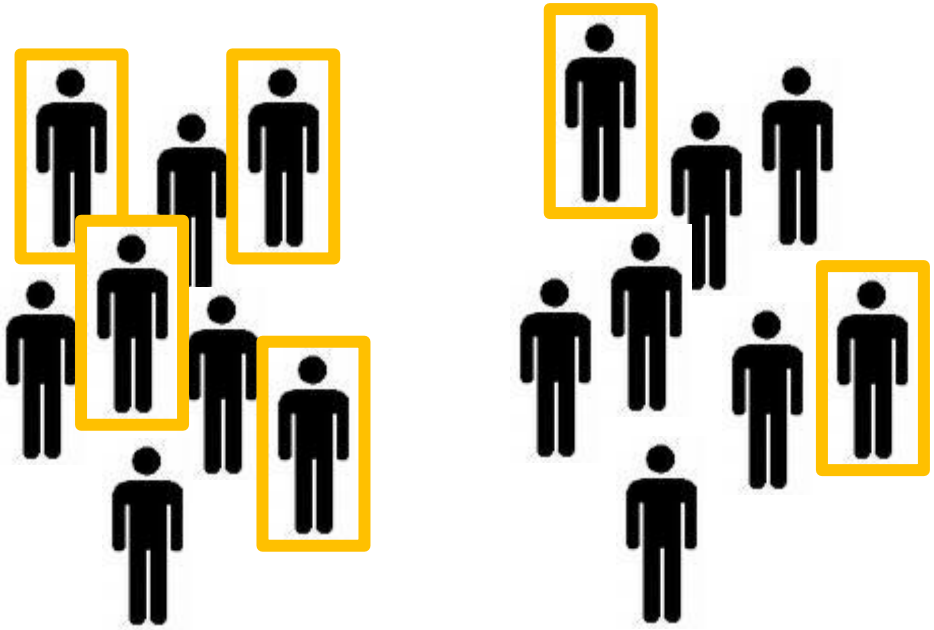


Asthma

MUTATION

POLYMORPHISM

Association studies: Genetic marker vs. phenotypic trait



Variant (SNP) allele

Genotyping

- Hypothesis driven study: SNP, gene/locus, pathway…
- Agnostic screening
- Genome wide association study (GWAS) – arrays
- Whole genome/exome sequencing (WGS, WES)

GENOTYPING Particular points in the genome	**SEQUENCING** All the nucleotides in a genomic region
….AGCTAAATGATAGCATCAT….	….AGCTAAATGATAGCATCAT….

GWAS catalog: https://www.ebi.ac.uk/gwas/

REVIEW

10 Years of GWAS Discovery: Biology, Function, and Translation

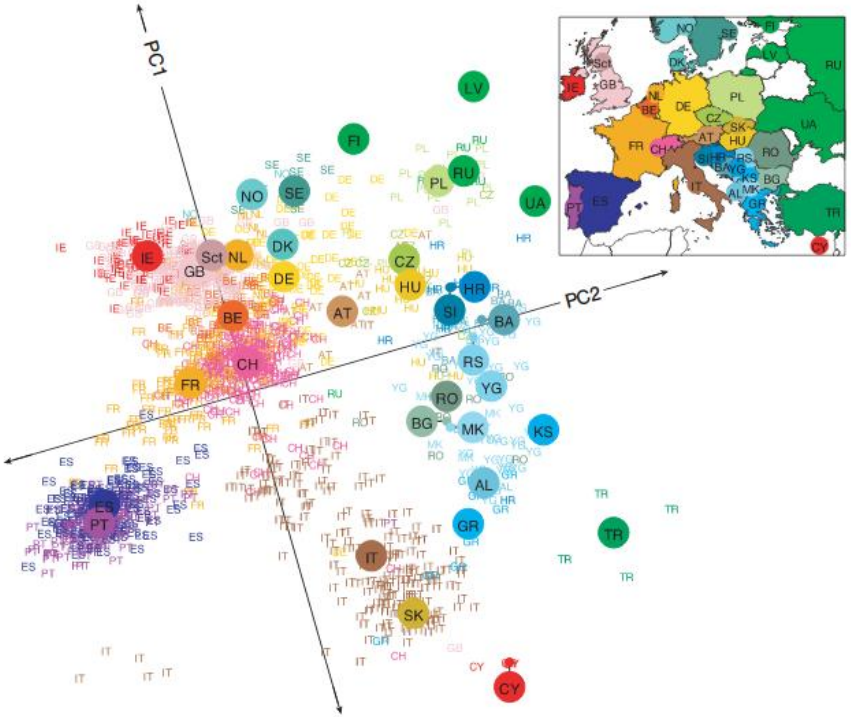
Peter M. Visscher,[1,2,*] Naomi R. Wray,[1,2] Qian Zhang,[1] Pamela Sklar,[3] Mark I. McCarthy,[4,5,6] Matthew A. Brown,[7] and Jian Yang[1,2]

Application of the experimental design of genome-wide association studies (GWASs) is now 10 years old (young), and here we review the remarkable range of discoveries it has facilitated in population and complex-trait genetics, the biology of diseases, and translation toward new therapeutics. We predict the likely discoveries in the next 10 years, when GWASs will be based on millions of samples with array data imputed to a large fully sequenced reference panel and on hundreds of thousands of samples with whole-genome sequencing data.

Genotyping

- **Selection of individuals**
  - Phenotype
  - DNA quality
  - **Population stratification** (ethnicity, technical…)

Novembre et al 2008



Ethnicity:

European and African different genetic background

European (in blue) high prevalence of disease X

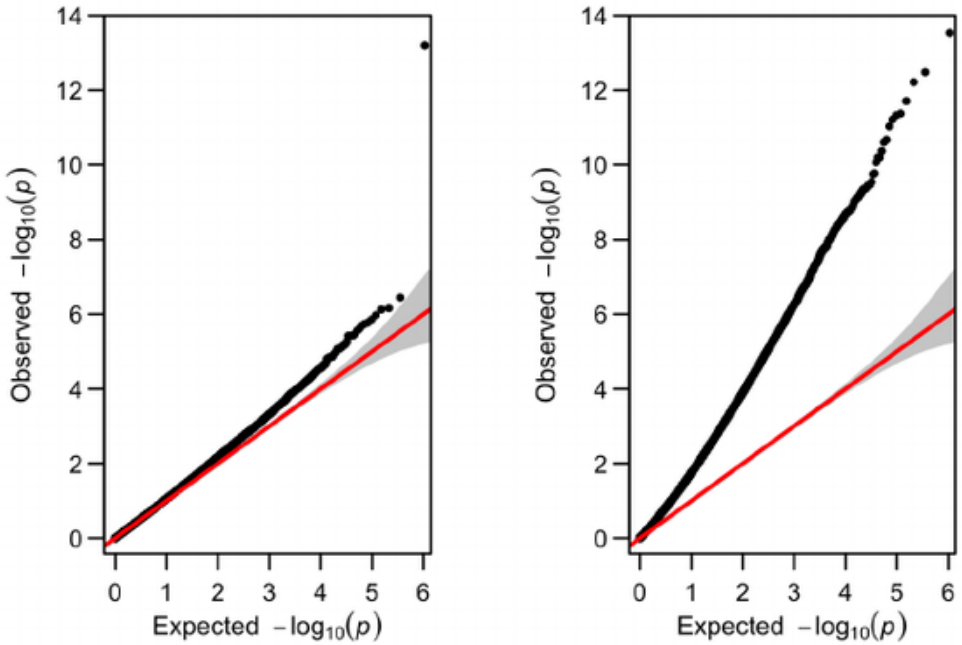
African low prevalence of disease X

GWAS multipletesting correction

- Many variants, then multipletesting correction needed
- Genome wide significant:

  P value < 5E-08 (number of independent LD blocks in genome)
- Suggestive association:

  P value < 1E-05



Manhattan plot

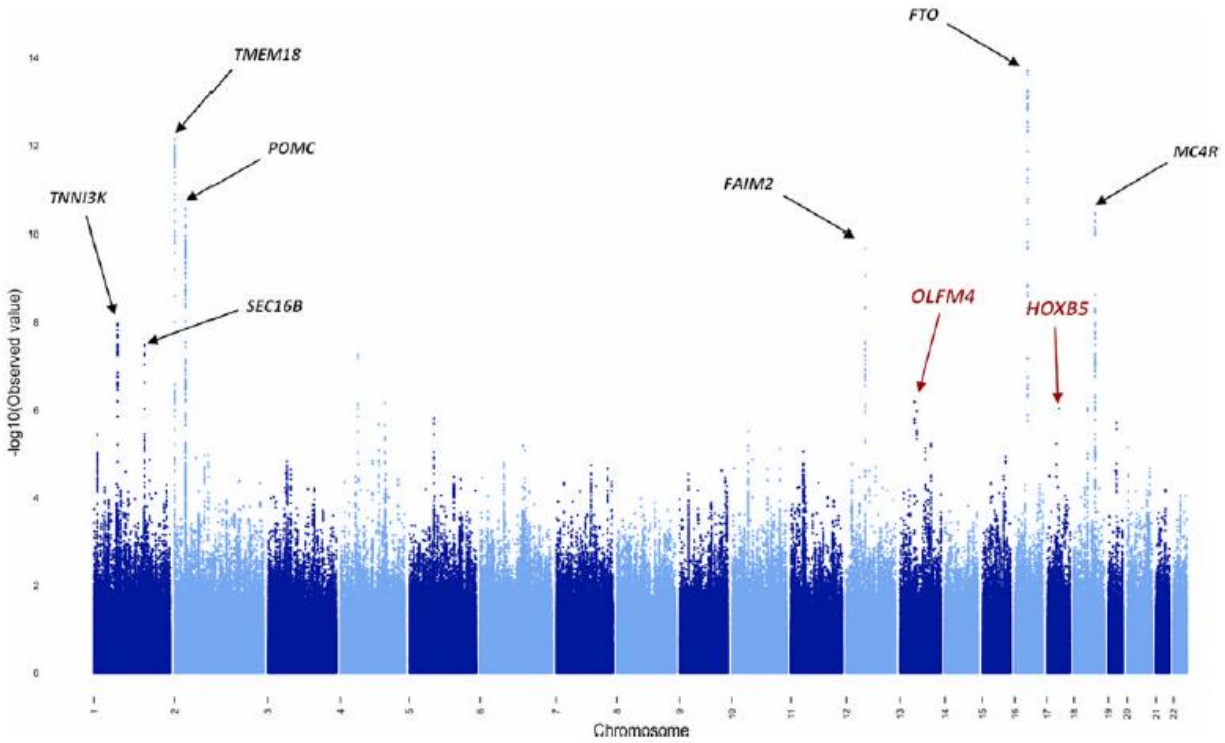


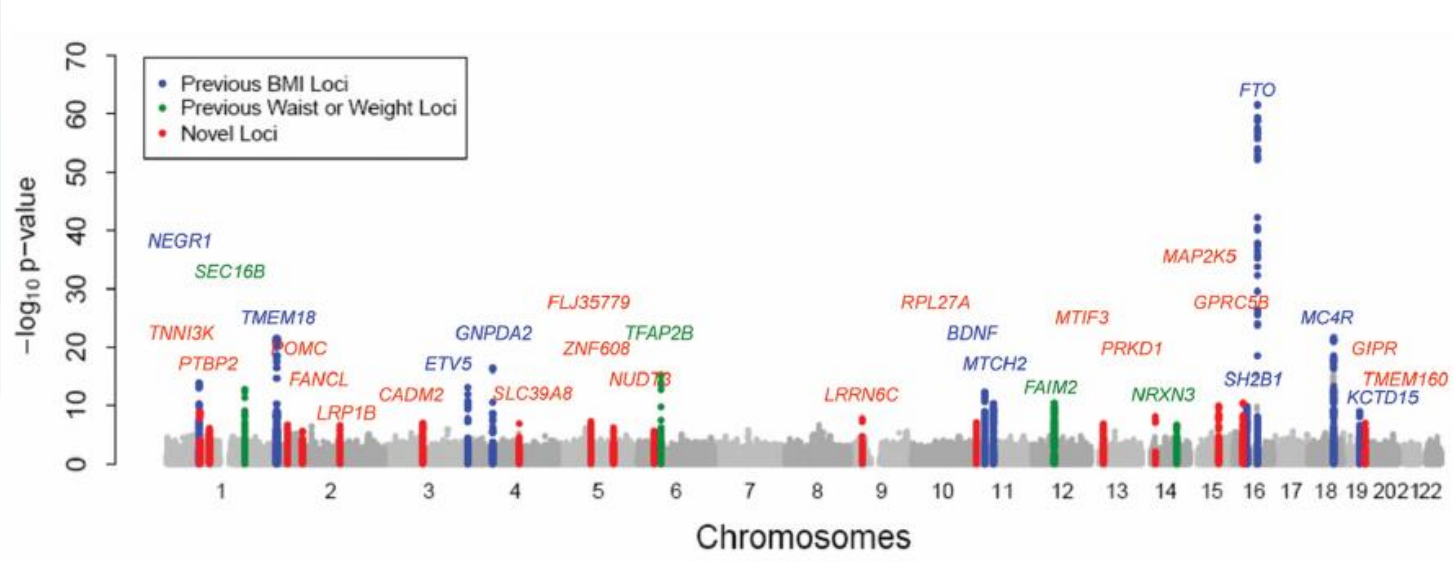
Figure 1.
Manhattan Plot of the meta-analysis of childhood obesity GWAS runs in the discovery stage (5,530 cases and 8,318 controls), with each locus achieving genome wide significance ($P < 5\times10^{-8}$) indicated in black text. In addition, the novel loci uncovered in this study are indicated in red text.

Personalized medicine

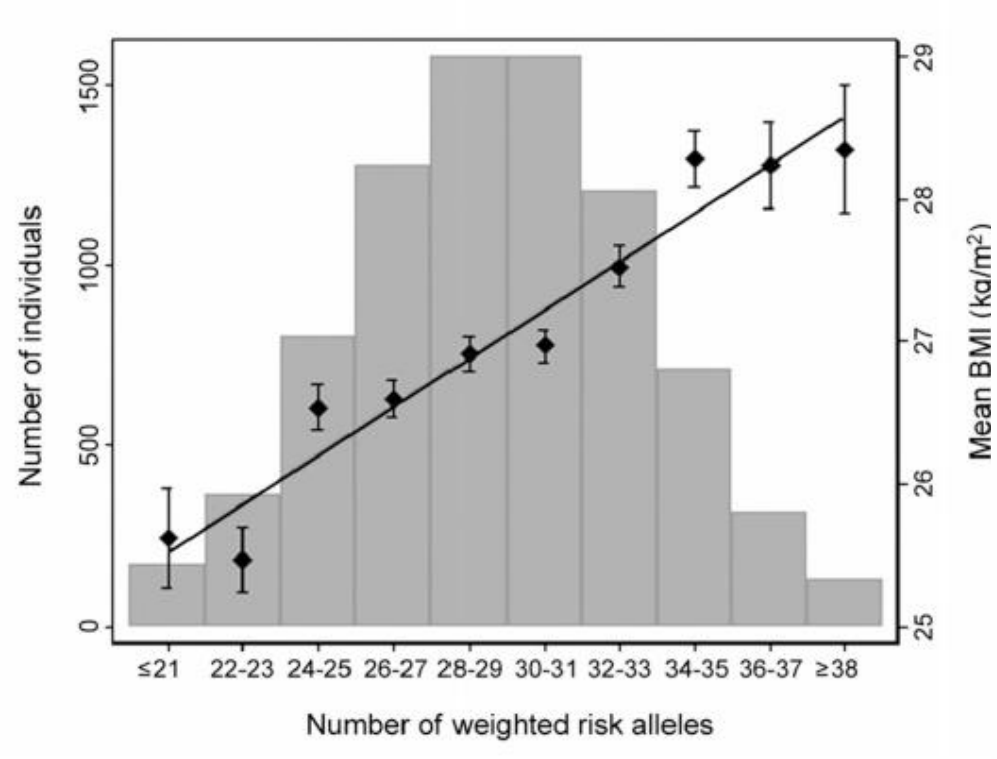
Nat Genet. 2010 November ; 42(11): 937–948. doi:10.1038/ng.686.

Association analyses of 249,796 individuals reveal eighteen new loci associated with body mass index

Elizabeth K. Speliotes[1,2,*], Cristen J. Willer[3,*], Sonja I. Berndt[4,*], Keri L. Monda[5,*], Gudmar Thorleifsson[6,*], Anne U. Jackson[3], Hana Lango Allen[7], Cecilia M. Lindgren[8,9], Jian'an Luan[10], Reedik Mägi[8], Joshua C. Randall[8], Sailaja Vedantam[1,11], Thomas W. Winkler[12], Lu



Personalized medicine



ROC curve

Transcriptomics

Gene expression profiling

- The identification and characterization of the mixture of mRNA that is present in a specific sample

Principle

- The abundance of specific mRNA transcripts in a biological sample is a reflection of the expression levels of the corresponding genes (Manning et al., 2007). Information obtained from microarrays

Application

- To associate differences in mRNA mixtures originating from different groups of individuals to phenotypic differences between the groups (Nachtomy et al., 2007)

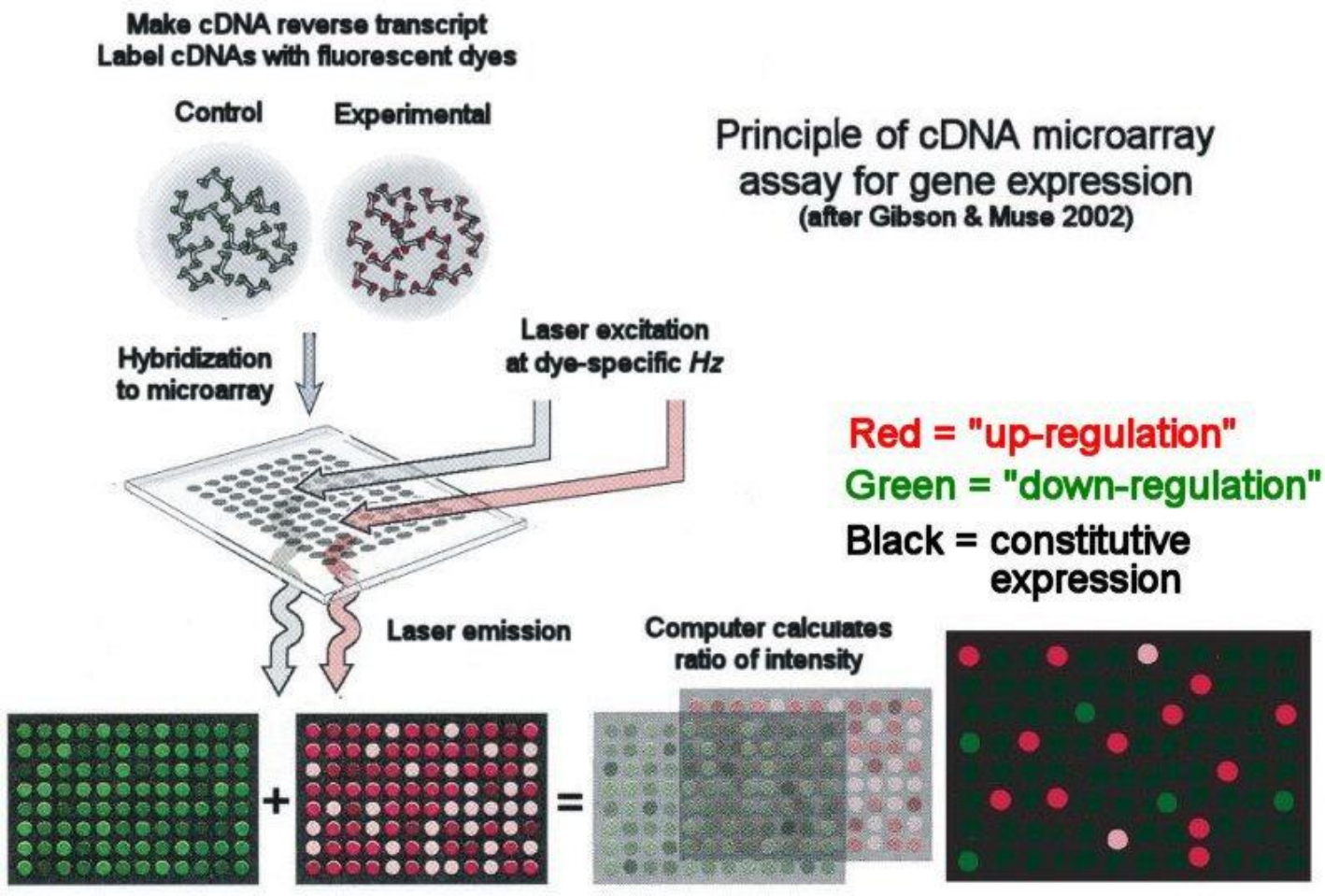
Challenge

- The transcriptome in contrast to the genome is highly variable over time, between cell types and environmental changes (Celis et al., 2000).

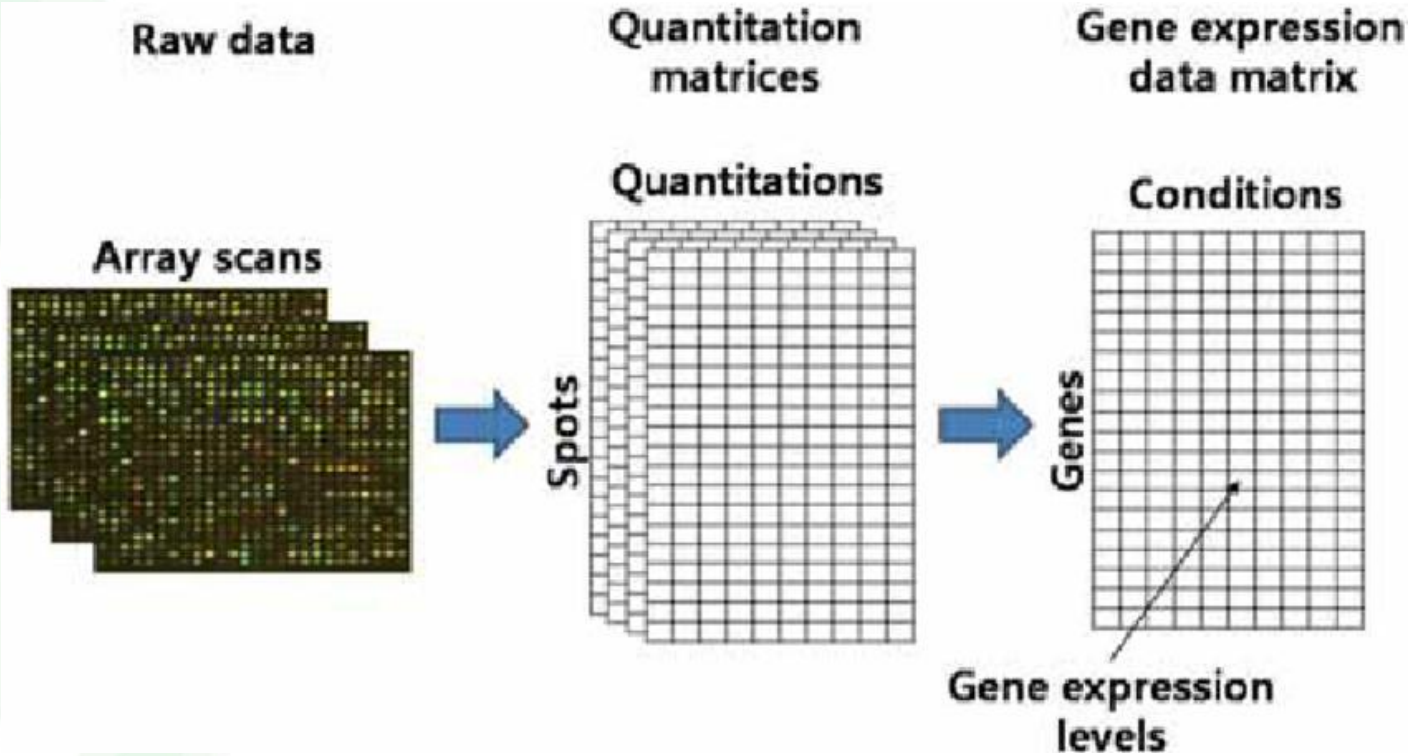
Recent improvement transcriptomics

- RNAseq

Transcriptomics

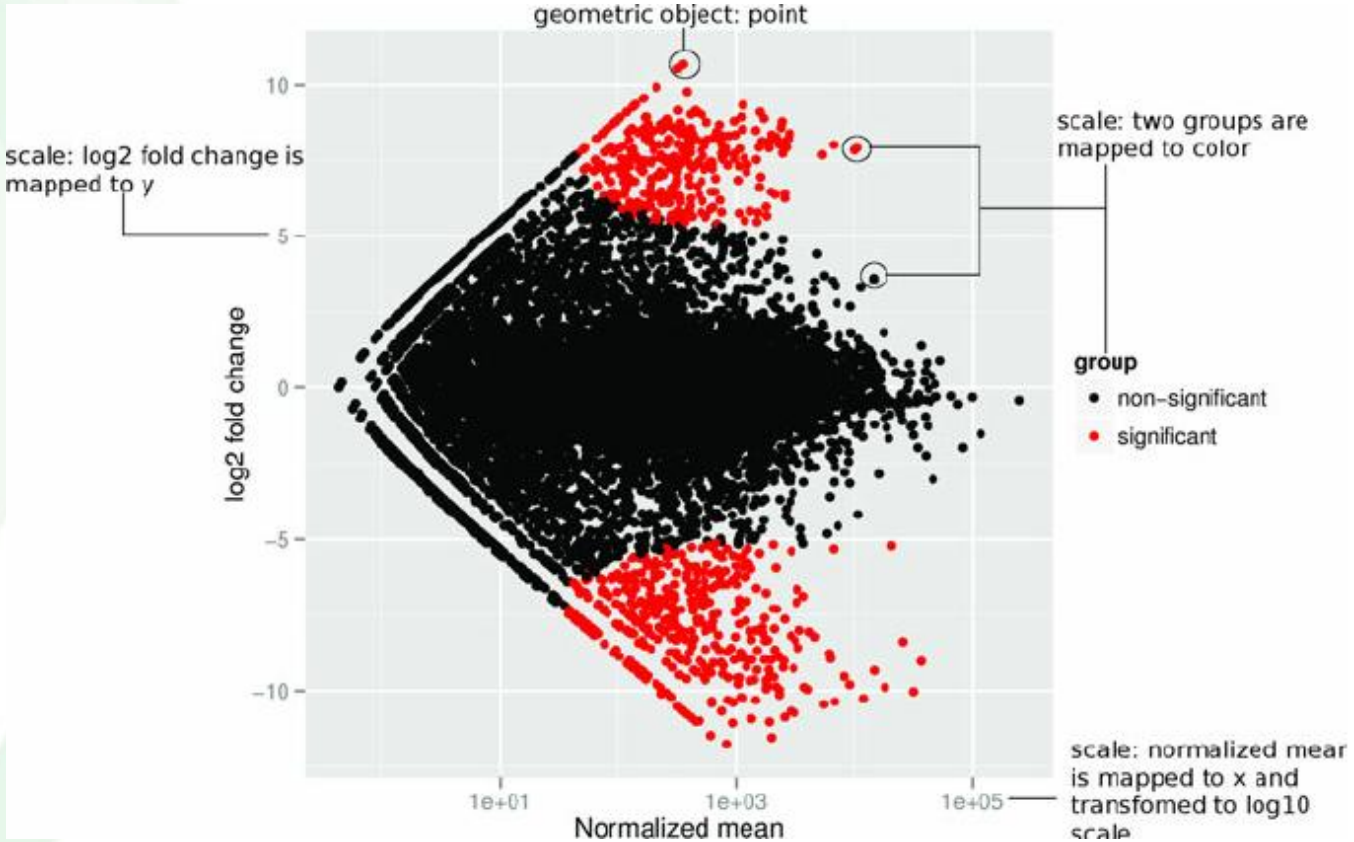


Transcriptomics

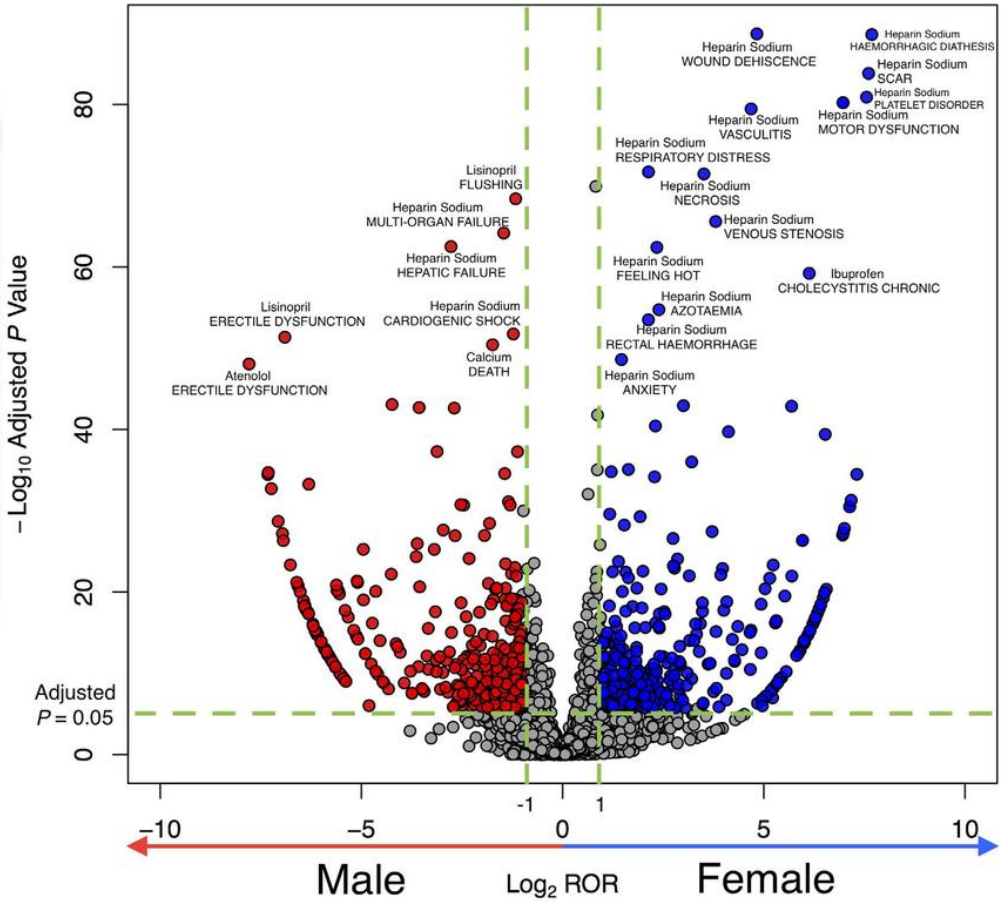


Statistical analysis: linear models (one per gene) + empirical Bayes (**limma**)
Multiple comparisons: FDR

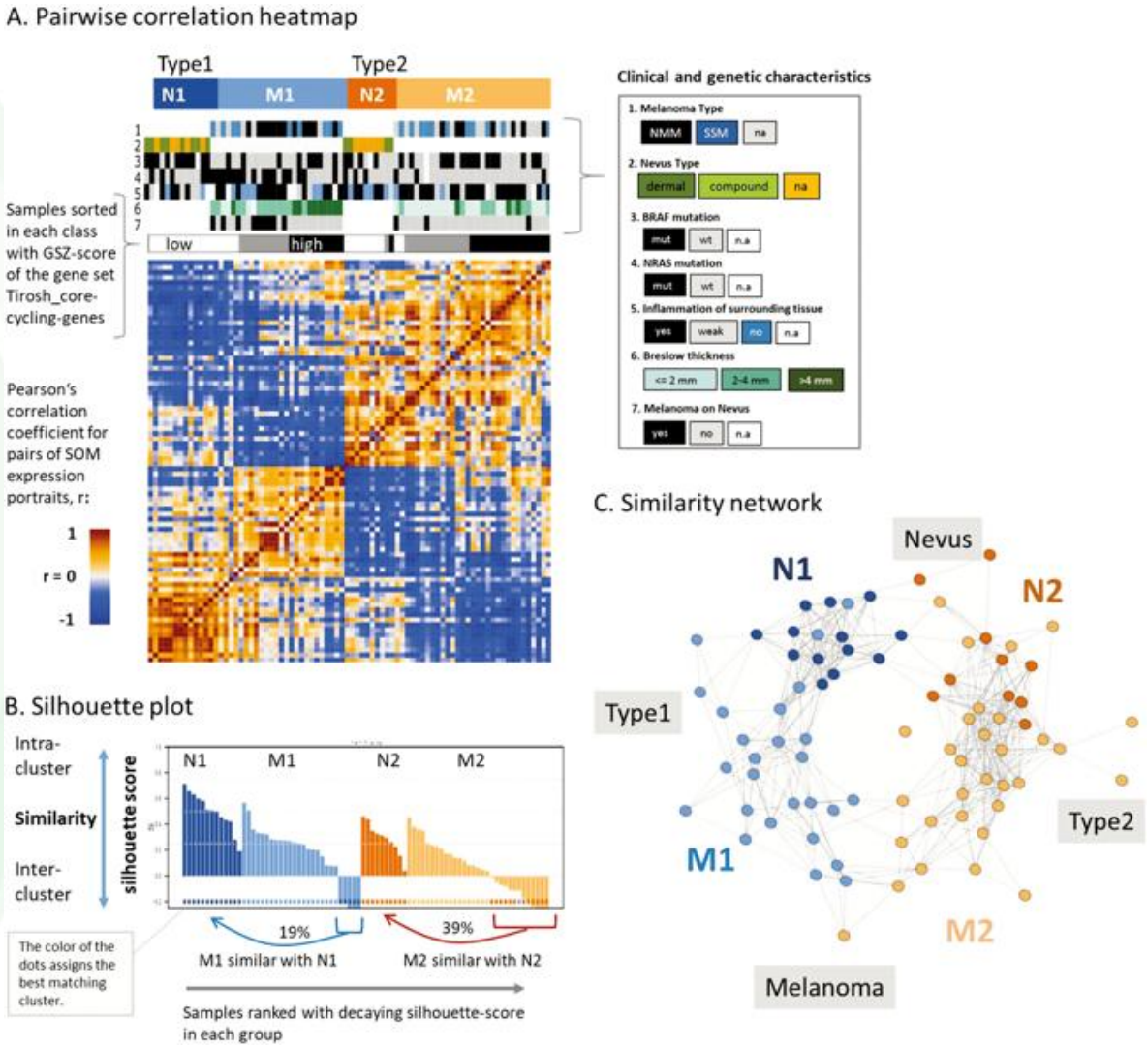
MA plot


geometric object: point
scale: log2 fold change is mapped to y
scale: two groups are mapped to color
log2 fold change
10
5
0
−5
−10
group
non−significant
significant
1e+01
1e+03
1e+05
Normalized mean
scale: normalized mear is mapped to x and transfomed to log10 scale

Volcano plot



Transcriptomics



Epigenomics

Epigenetic processes

- ❑ Mechanisms other than changes in DNA sequence that cause effect in gene transcription and gene silencing.
- ❑ Number of mechanisms of epigenomics but is mainly based on two mechanisms: DNA methylation and histone modification

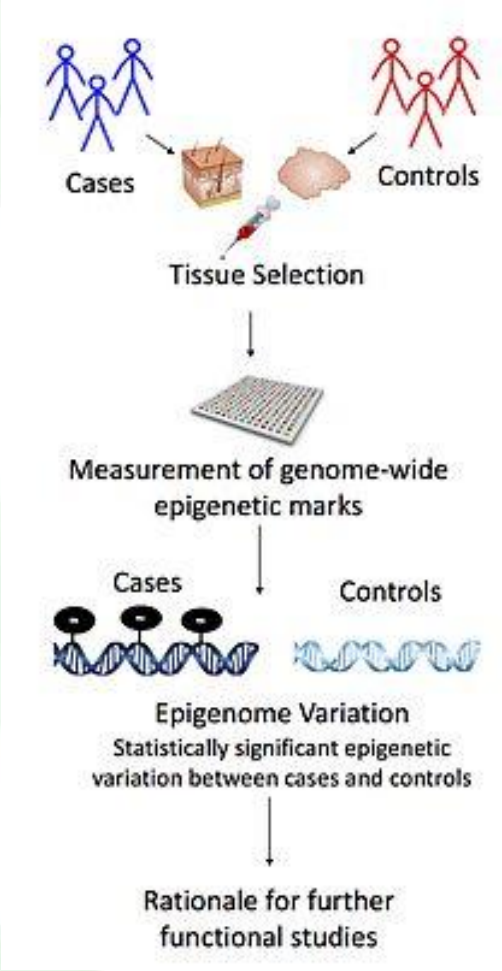
Goal

- ❑To study epigenetic processes on a large (ultimately genome-wide) scale to assess the effect on **disease** (450K and EPIC Illumina arrays)

Association with disease

- ❑ Hypermethylation of CpG islands located in promoter regions of genes is related to gene silencing
- ❑ Altered gene silencing plays a causal role in human disease
- ❑ Histone proteins are involved in the structural packaging of DNA in the chromatin complex. Post translational histone modifications such as acetylation and methylation are believed to regulate chromatin structure and therefore gene expression

Epigenomics



- ❑ 450K / EPIC Illumnina arrays (**CpGs**)
- ❑ Statistical analyses using **linear models** (**robust regression** or beta regression)
- ❑ Multiple comparisons: **FDR**

Proteomics

Epigenetic processes

- Role proteins in biological systems. It consists of all proteins present in specific cell types or tissue and highly variable over time, between cell types and will change in response to changes in its environment (Fliser et al., 2007).
- The overall function of cells can be described by the proteins (intra- and intercellular and the abundance of these proteins (Sellers et al., 2003)
- Although all proteins are directly correlated to mRNA (transcriptome) , post translational modifications (PTM) and environmental interactions impede to predict from gene expression analysis alone (Hanash et al., 2008)

Tools for proteomics

- Mass spectrometry (MS)
- Protein microarrays using capturing agents such as antibodies.

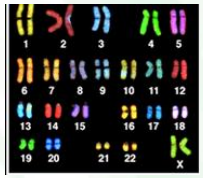
Major focuses

- the identification of proteins and proteins interacting in protein-complexes
- Then the quantification of the protein abundance. The abundance of a specific protein is related to its role in cell function (Fliser et al., 2007)

Metabolomics

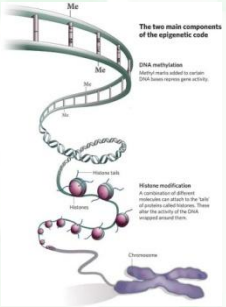
- The metabolome consists of small molecules (e.g. lipids or vitamins) that are also known as metabolites (Claudino et al., 2007).
- Metabolites are involved in the energy transmission in cells (metabolism) by interacting with other biological molecules following metabolic pathways.
- Metabolic phenotypes are the by-products of interactions between genetic, environmental, lifestyle and other factors (Holmes et al., 2008).
- The metabolome is highly variable and time dependent, and it consists of a wide range of chemical structures.
- An important challenge of metabolomics is to acquire qualitative and quantitative information with preturbance of environment (Jelly et al., 2010)
- Recent research on linking the metabolome with complex diseases or changes in enviromental exposures

Omics

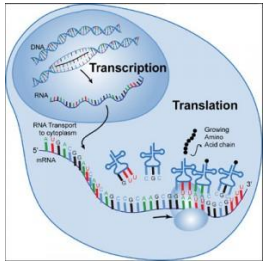


GENOME
hereditary information (DNA)
stable
>99% equal between individuals
1.5% coding genes

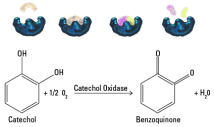
EXPOSOME
dynamic
diet, metalls, air pollution, stress…



EPIGENOME
changes in gene expression caused by mechanisms other than DNA sequence
tissue and time especific



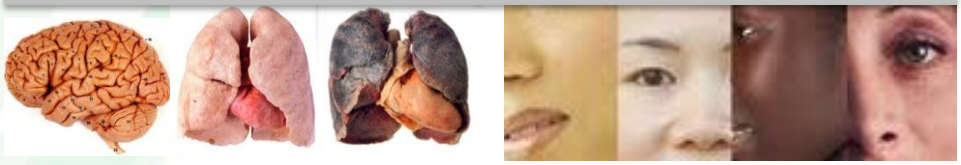
TRANSCRIPTOME
gene expression (RNA)
tissue and time especific



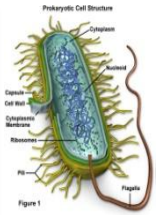
PROTEOME
tissue and time especific

METABOLOME
tissue and time especific

DISEASOME (PHENOTYPE)



METAGENOME
(metatranscriptome, virome…)
bacteria and virus
1-3% body's mass
trilions of microorganisms



Others

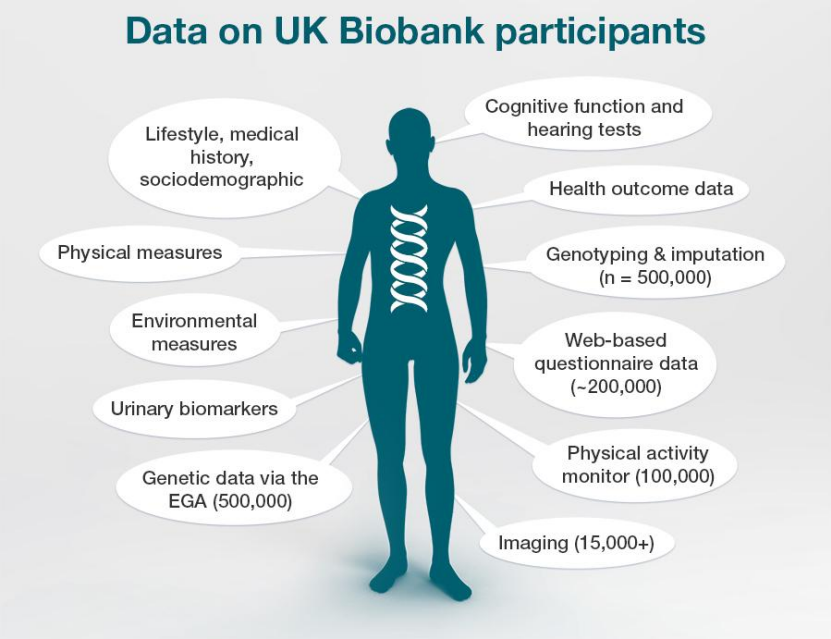
- Metagenome
- Immunome
- Pathogenome
- Regenome
- Psychogenome
- Phenome
- Exposome
- …

Omic data repositories

dbGaP (genomic data https://www.ncbi.nlm.nih.gov/gap

EGA (genomic data http://ega.crg.eu/)

UK Biobank (genomic data https://www.bdi.ox.ac.uk/)



GEO (transcriptomics – BioC: GEOquery)

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

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

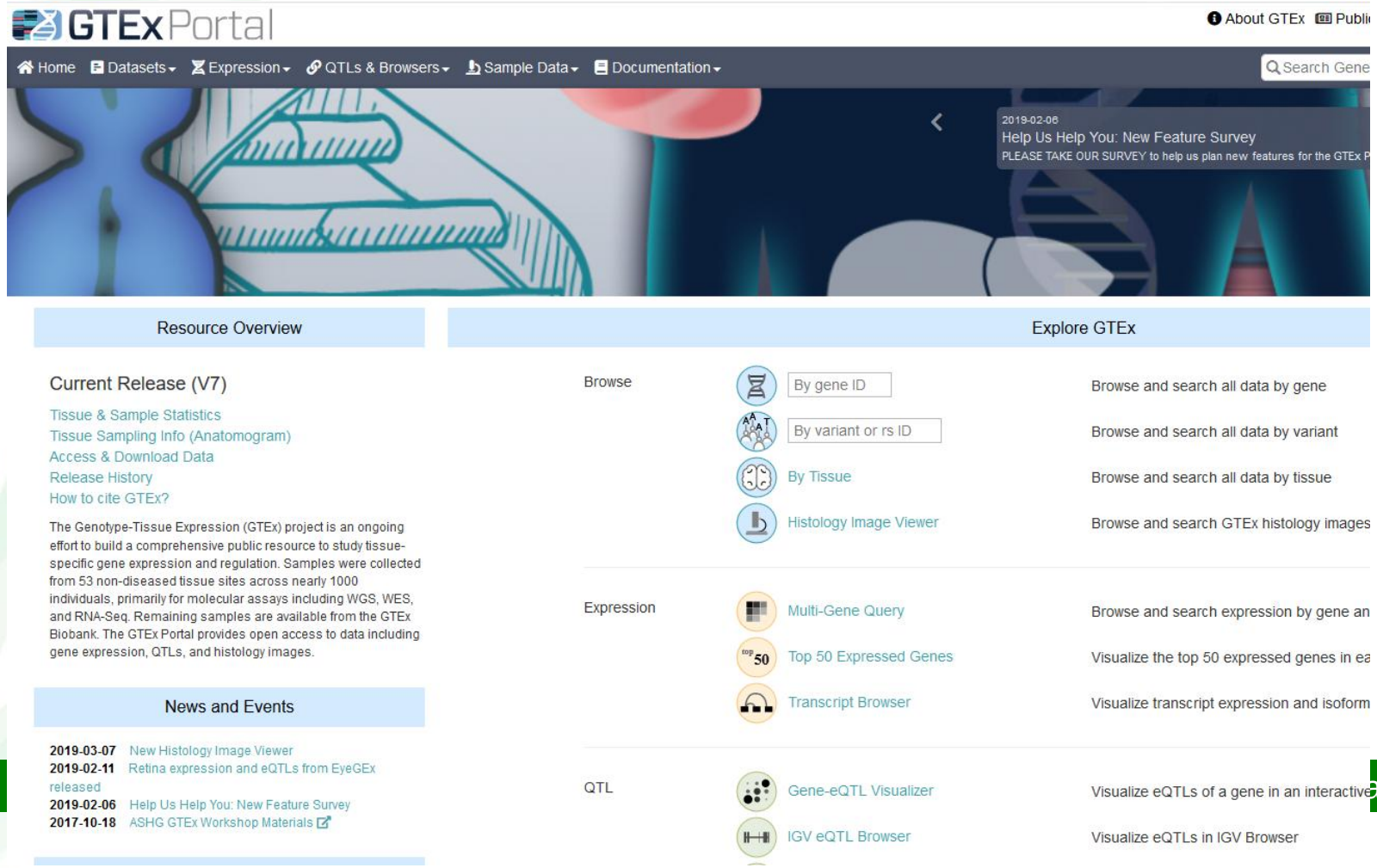
Keyword or GEO Accession

Getting Started	Tools	Browse Content	
Overview	Search for Studies at GEO DataSets	Repository Browser	
FAQ	Search for Gene Expression at GEO Profiles	DataSets:	4348
About GEO DataSets	Search GEO Documentation	Series:	110236
About GEO Profiles	Analyze a Study with GEO2R	Platforms:	19461
About GEO2R Analysis	Studies with Genome Data Viewer Tracks	Samples:	2926596
How to Construct a Query	Programmatic Access		
How to Download Data	FTP Site		

Omic data repositories

GTeX(transcriptomics – genomics)

https://gtexportal.org/home/



TCGA (multiomic – BioC: several)

https://www.cancer.gov/

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Keyword or GEO Accession

Getting Started	Tools	Browse Content	
Overview	Search for Studies at GEO DataSets	Repository Browser	
FAQ	Search for Gene Expression at GEO Profiles	DataSets:	4348
About GEO DataSets	Search GEO Documentation	Series:	110236
About GEO Profiles	Analyze a Study with GEO2R	Platforms:	19461
About GEO2R Analysis	Studies with Genome Data Viewer Tracks	Samples:	2926596
How to Construct a Query	Programmatic Access		
How to Download Data	FTP Site		

Omic data repositories

recount (transcriptomic – BioC: several)

https://jhubiostatistics.shinyapps.io/recount/

recount2: analysis-ready RNA-seq gene and exon counts datasets

Datasets | Popular datasets | GTEx | TCGA | Documentation | Download data with R | Accessing recount2 via SciServer

Transcript counts are now available thanks to the work of Fu et al, bioRxiv, 2018. Exon counts are now from dis
further information.

recount2

A multi-experiment resource of analysis-ready RNA-seq

recount2 is an online resource consisting of RNA-seq gene and exon counts as well as covera
data were processed with Rail-RNA as described in the recount2 paper and at Nellore et al, Genome Biology, 2016 which created the coverage
extracted phenotype data, which we provide in their raw formats as well as in RangedSummarizedExperiment R objects (described in the Summ
file, which can be used with the derfinder Bioconductor package to perform annotation-agnostic differential expression analysis at the expresse
RangedSummarizeExperiment objects, phenotype tables, sample bigWigs, mean bigWigs, and file information tables are ready to use and free
data for a specific study. By taking care of several preprocessing steps and combining many datasets into one easily-accessible website, we m

Main publication

- **Collado-Torres L**, **Nellore A**, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. Reproducible RNA-seq analy

Omic data repositories

Datasets | Popular datasets | GTEx | TCGA | Documentation | Download data with R | Accessing recount2 via SciServer | Contribute your data

This tab shows the information for the TCGA project. Due to its size, we also provide ranged summarized experiment objects (RSE) by tissue at the gene and exon levels.

Show 1 entries

accession	number of samples	species	abstract	gene	exon
All	All	All	All	All	All
TCGA	11284	human	The Cancer Genome Atlas (TCGA) is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer. The TCGA dataset, comprising more than two petabytes of genomic data, has been made publically available, and this genomic information helps the cancer research community to improve the prevention, diagnosis, and treatment of cancer.	RSE v2 counts v2 RSE v1 counts v1 RSE by tissue (version 2): adrenal gland bile duct bladder bone marrow brain breast cervix colorectal esophagus eye head and neck kidney liver lung lymph nodes ovary pancreas pleura prostate skin soft tissue stomach testis thymus thyroid uterus RSE by tissue (version 1): adrenal gland bile duct bladder bone marrow brain breast cervix colorectal esophagus eye head and neck kidney liver lung lymph nodes ovary pancreas pleura prostate skin soft tissue stomach testis thymus thyroid uterus	RSE v2 counts v2 RSE v1 counts v1 RSE by tissue (version 2): adrenal gland bile duct bladder bone marrow brain breast cervix colorectal esophagus eye head and neck kidney liver lung lymph nodes ovary pancreas pleura prostate skin soft tissue stomach testis thymus thyroid uterus RSE by tissue (version 1): adrenal gland bile duct bladder bone marrow brain breast cervix colorectal esophagus eye head and neck kidney liver lung lymph nodes ovary pancreas pleura prostate skin soft tissue stomach testis thymus thyroid uterus

Bioconductor

https://www.bioconductor.org/

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | A

About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

Install »

- Discover 1649 software packages available in *Bioconductor* release 3.8.

Get started with *Bioconductor*

- Install *Bioconductor*
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos