Omics: The -not so new- technologies

Alex Sánchez[1,2]

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

[1]*Department of Genetics, Microbiology & Statistics, Universitat de Barcelona*

UNIVERSITAT DE BARCELONA

Vall d'Hebron Institut de Recerca

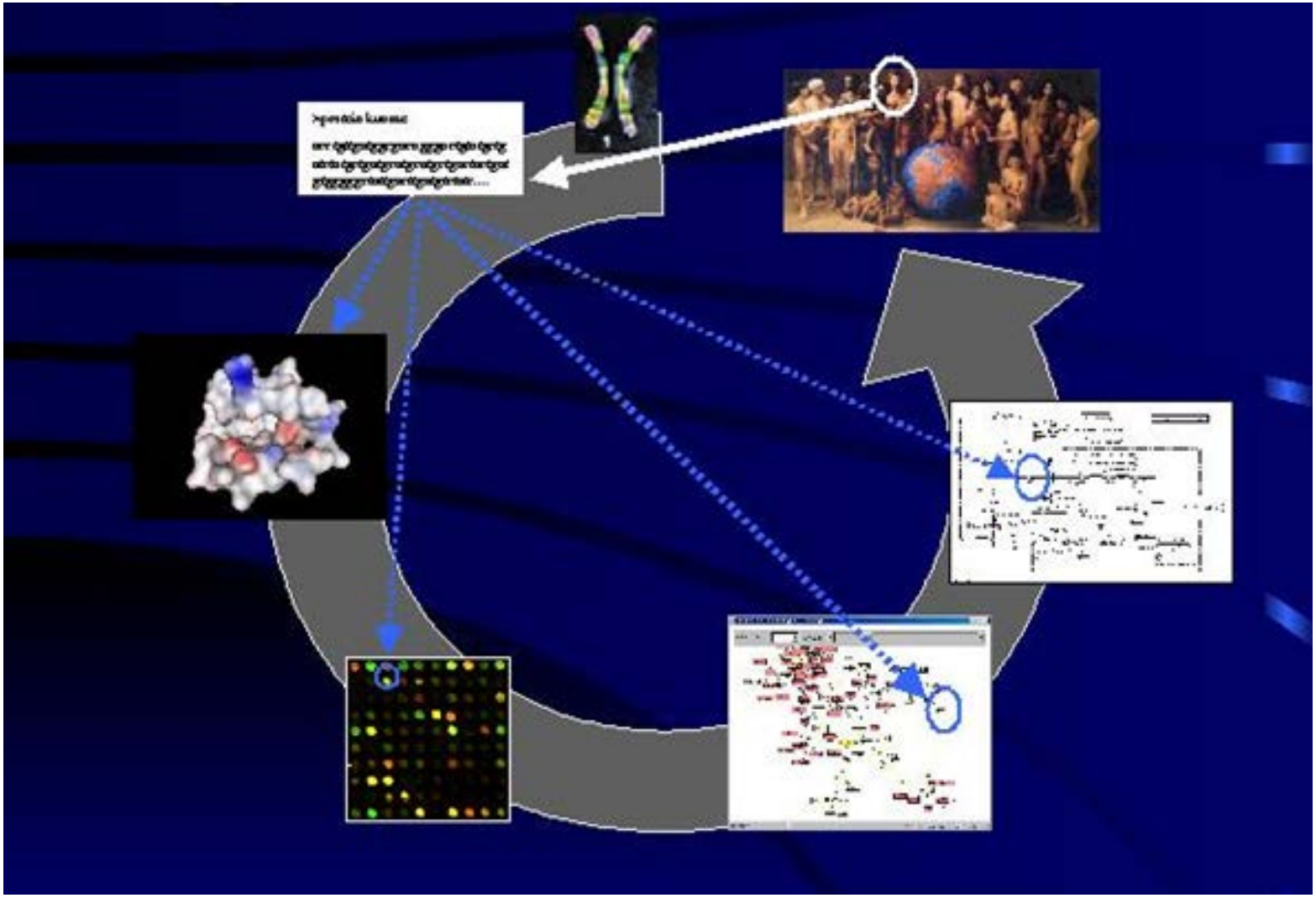
[2]*Statistics and Bioinformatics Unit Vall d'Hebron Institut de Recerca*

GRUP DE RECERCA EN ESTADÍSTICA I BIOINFORMÀTICA

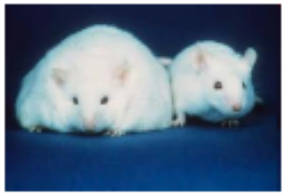
BIOstatnet

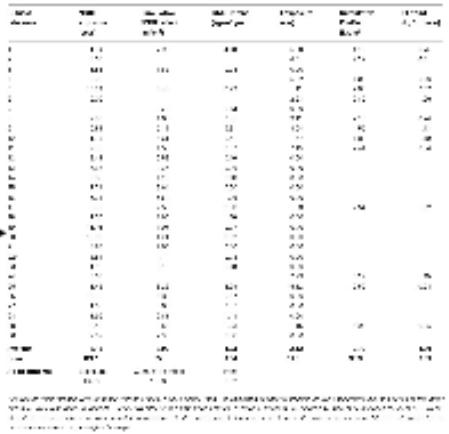
GRBIO

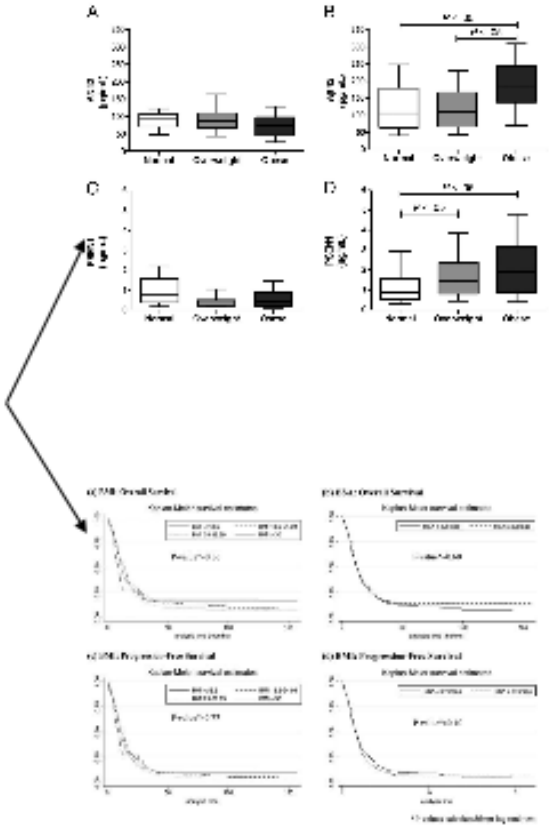
Pre-genomics vision



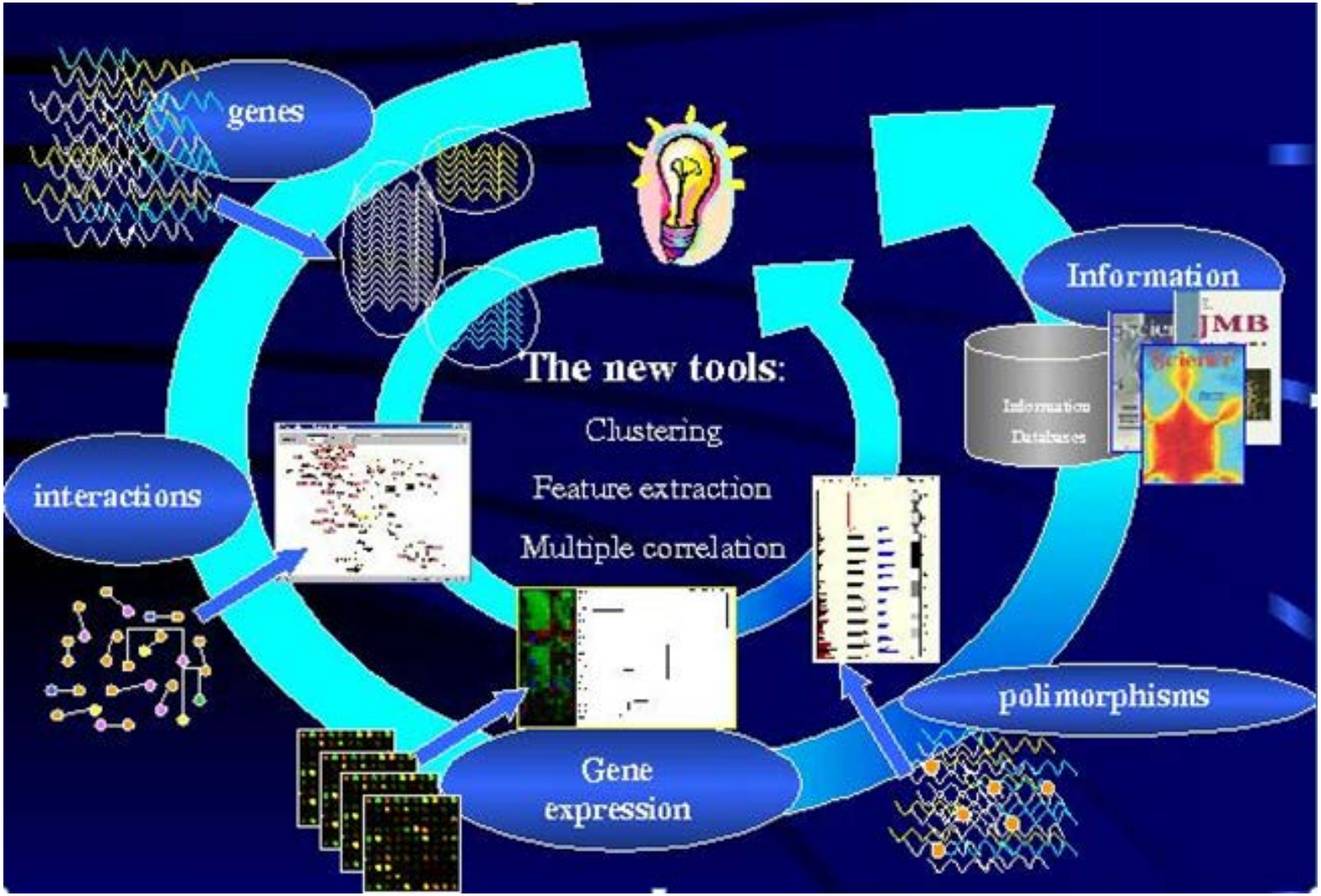
How we studied disease in 1996







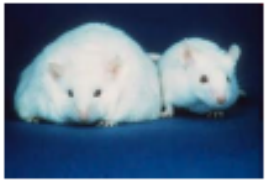
Post-genomics vision

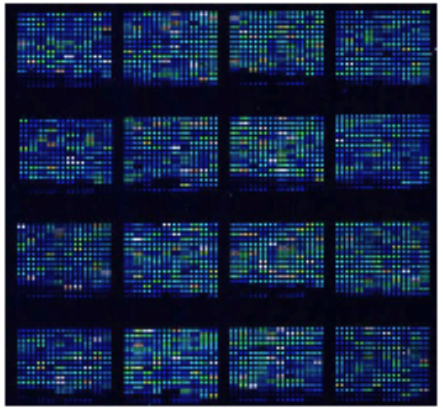


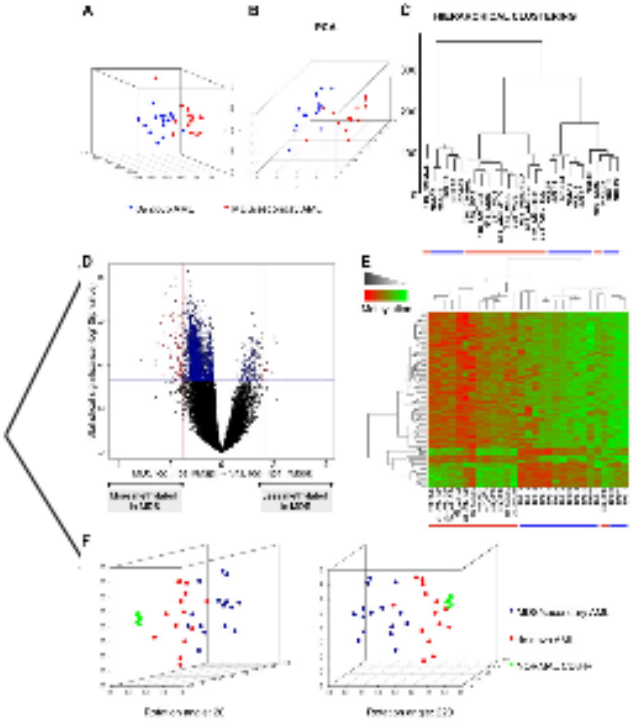
The (1st) paradigm shift

With the same resources we obtain a picture with lower resolution but with a view of the whole context

By 2006 first omics had arrived

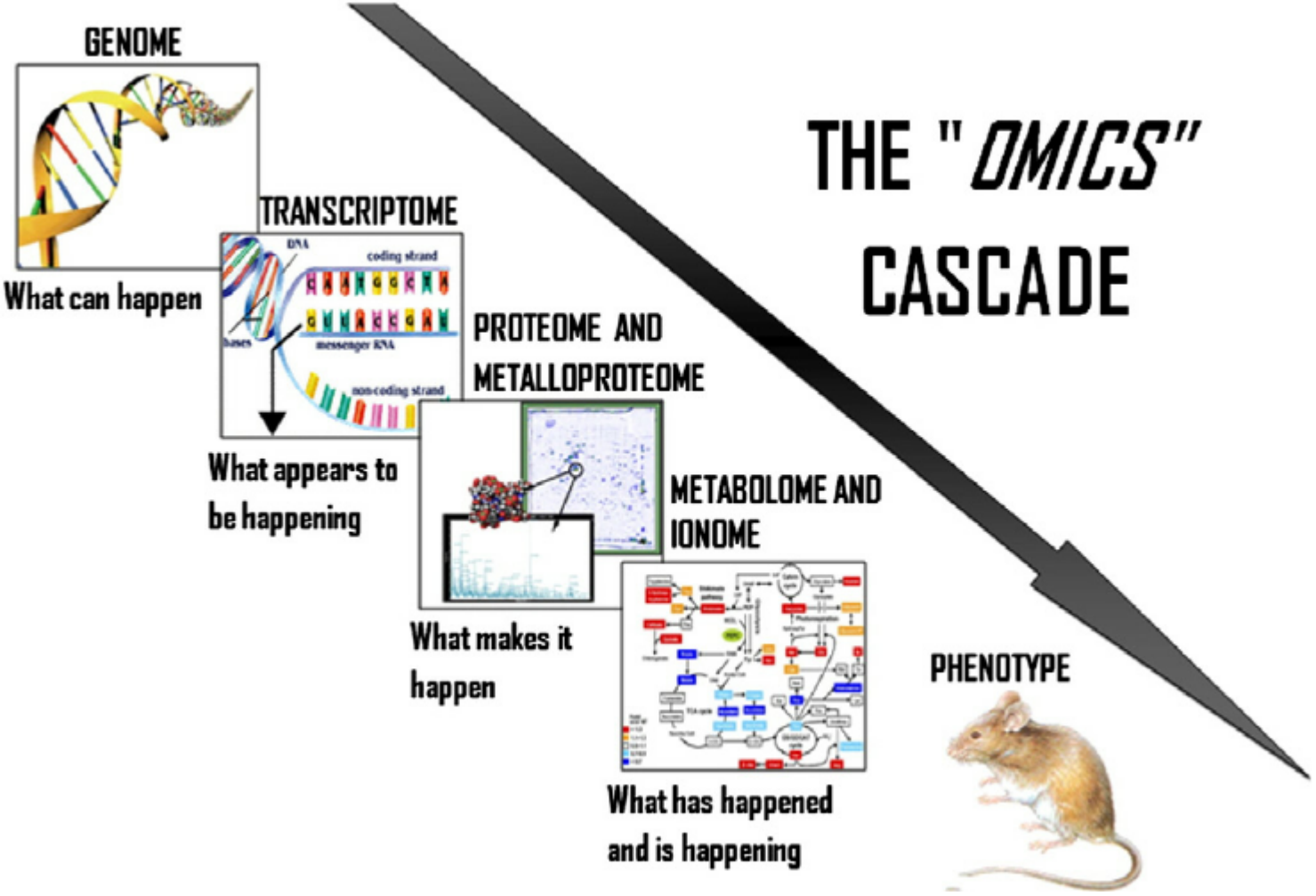




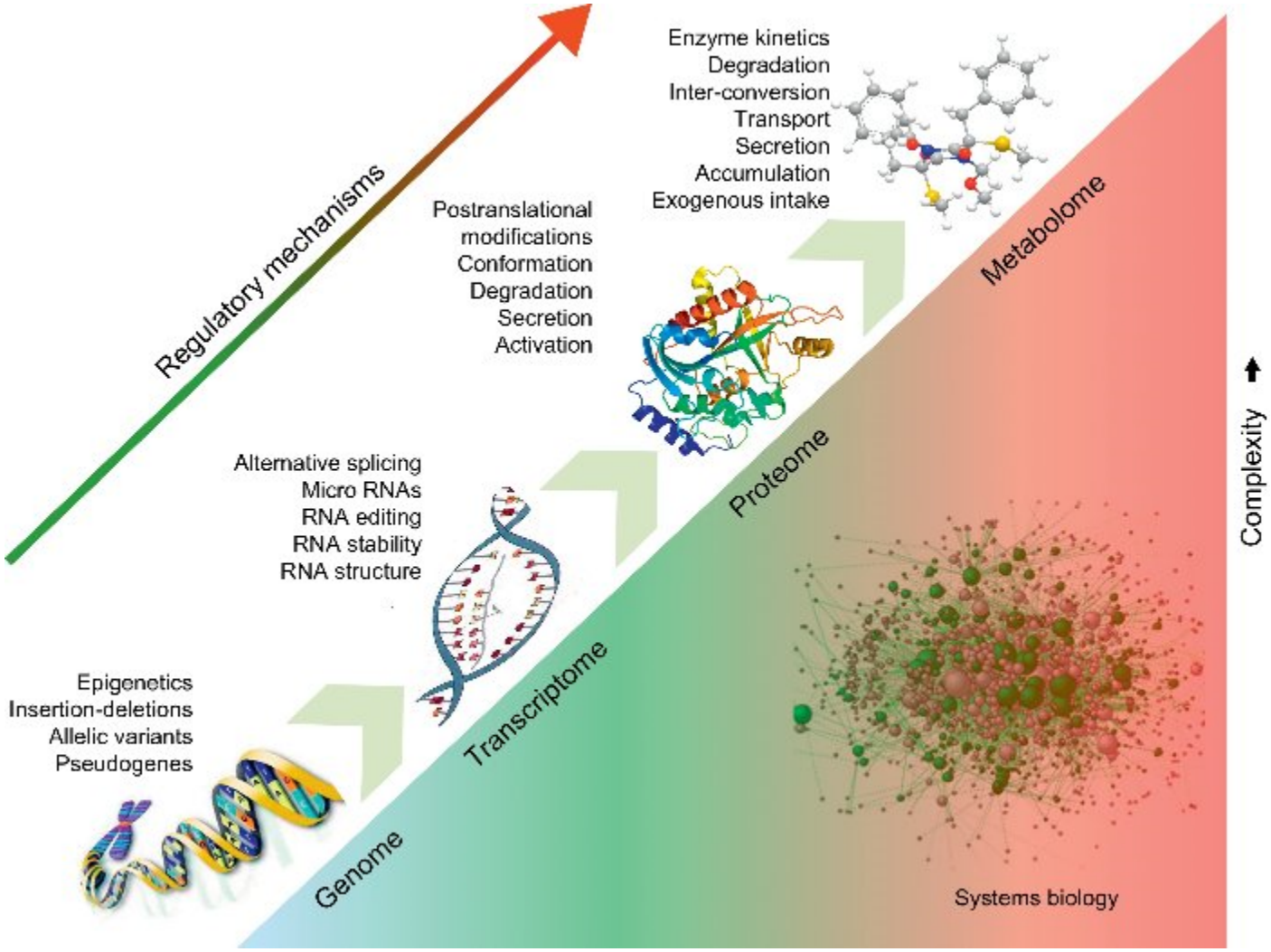


- With the availability of **new information**, such as the *human genome sequence* and **new technologies** the study of *different components* participating and/or regulating *complex biological processes*, triggered the development of several fields described together with the term OMICS.
- "omics = *large sets of biological molecules*"

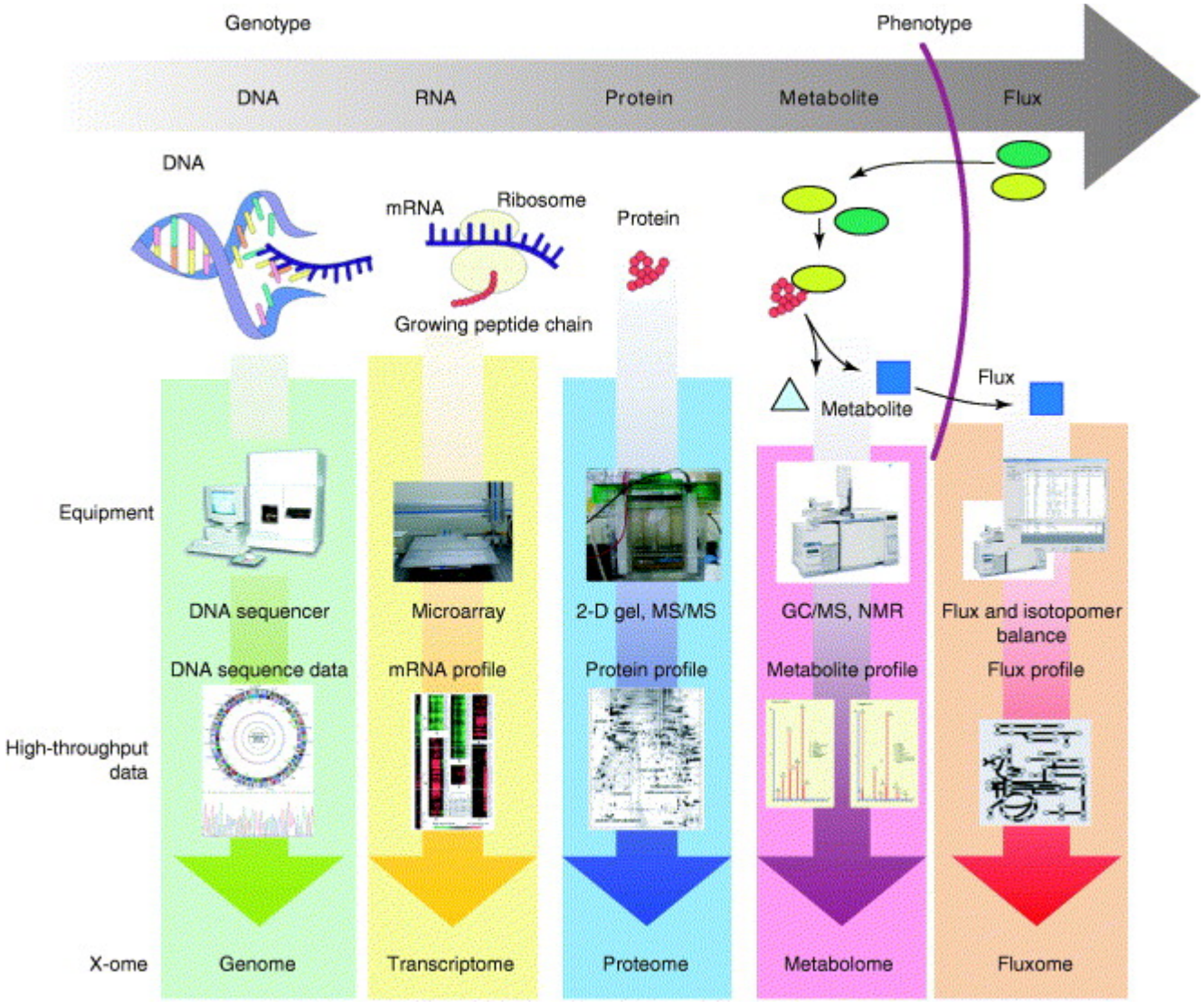
Life at distinct levels: The *OMES*



Even more complex: The REGULOME

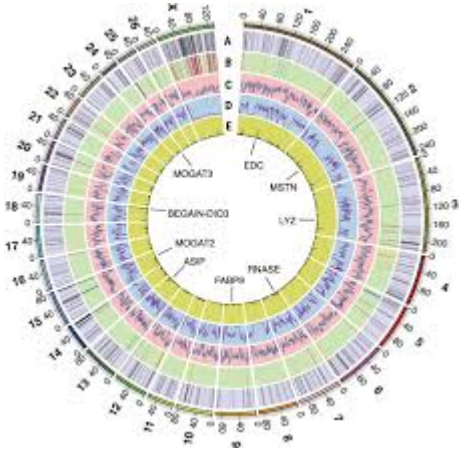


Omics technologies to study *omes*



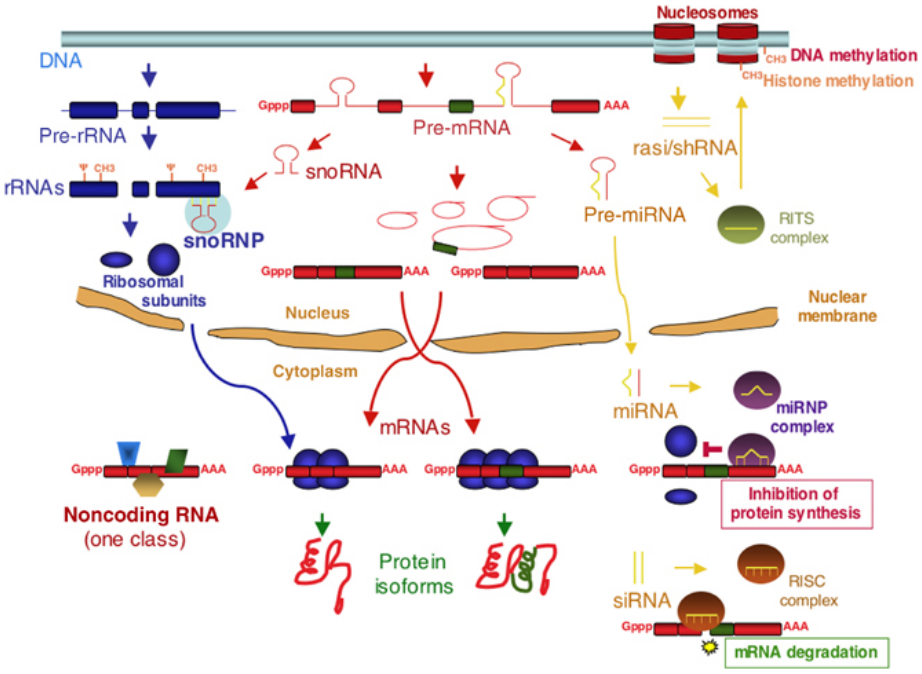
Genomics

- The complete set of DNA found in each cell is known as the *genome*
- Genomics uses *sequencing technologies* to identify and/or characterize all the *genes* in a given cell, tissue or organ.
- Genomic applied to study variation at the DNA level is used to find variants (SNPs or other) that can be associated with disease.



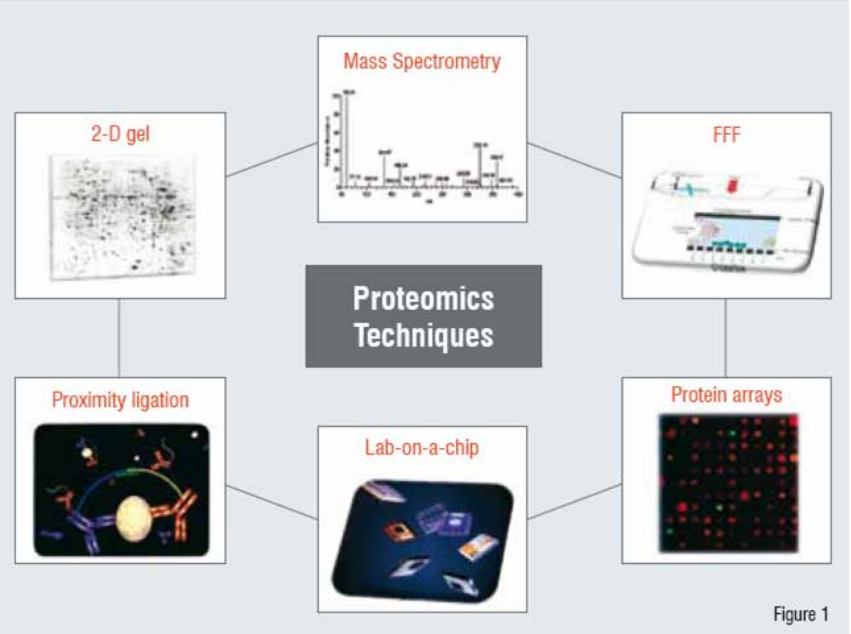
Transcriptomics

- The transcriptome is the set of all RNA molecules, in one or a population of cells.
- ***Transcriptomics***, examines expression levels of mRNAs in a given cell population, often using high-throughput techniques
- Most popular technologies:
  - **Microarrays**
  - **RNA-seq**



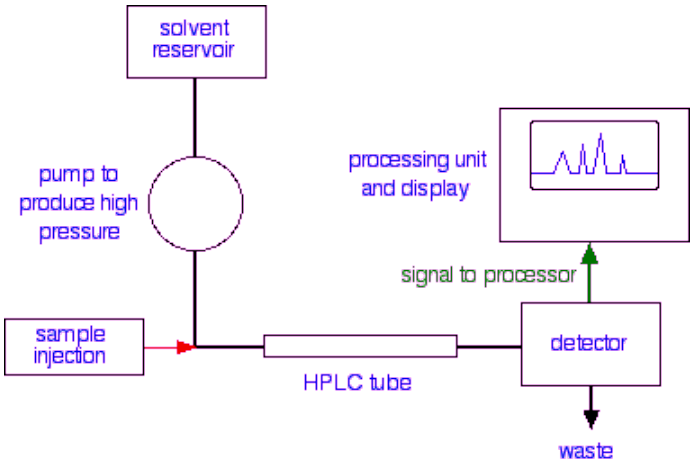
Proteomics

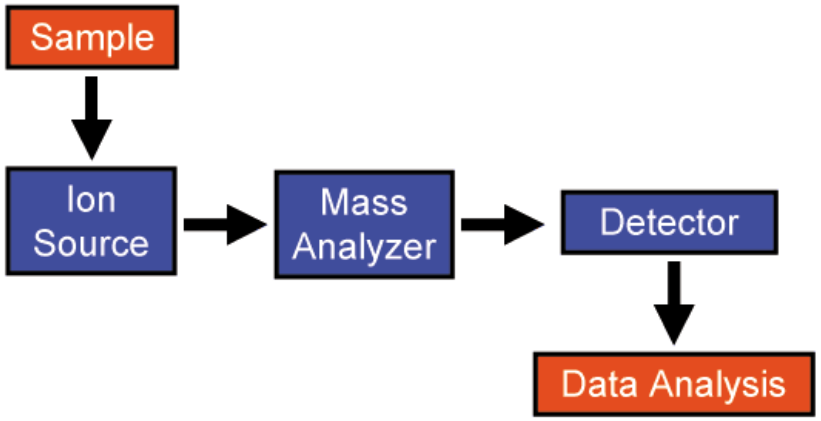
- Proteomics is about the large-scale study of proteins (the *proteome*) particularly their (3D) structures and functions.
- Relies on a wide spectra of techniques
  - 2D gel based
  - Mass Spectrometry (MS)
  - Seldi-TOF (MS)
  - Protein arrays,
- Opinion about its relevance varies (no consensus yet)



Metabolomics

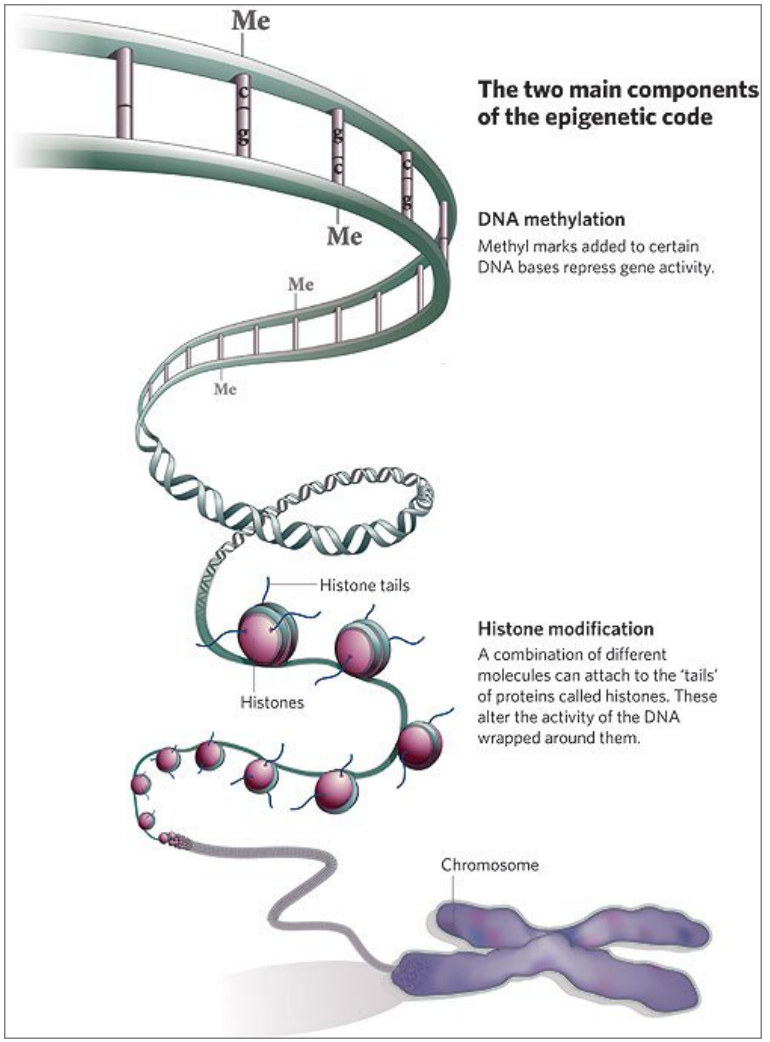
- Comprehensive and simultaneous systematic determination of
  - metabolite levels in the metabolome and
  - their changes over time as a consequence of stimuli.
- Relies on
  - Separation techniques
    - GC, CE, HPLC, UPLC
  - Detection techniques
    - NMR, MS
- Increase in use and relevance attributed in recent years



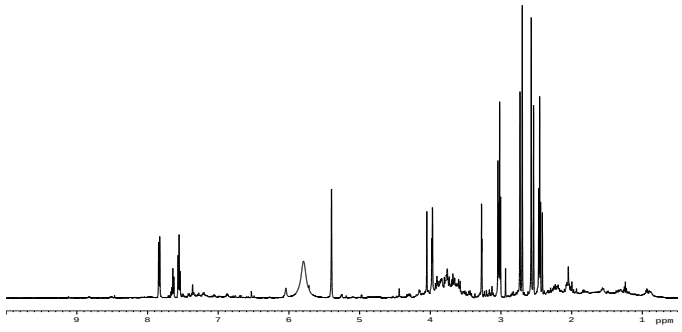


Epigenomics

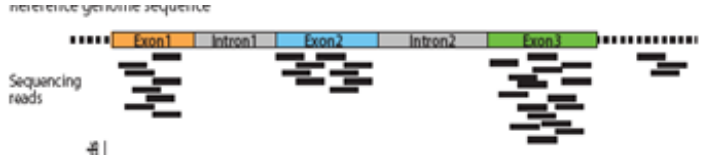
- Epigenetics is the study of changes in the phenotype or gene expression caused by other mechanisms than changes in the underlying DNA sequence.
  - DNA methylation
  - Histone modifications
- Epigenetics refers to the study of single genes or sets of genes.
- Epigenomics refers to global analyses of epigenetic changes across the entire genome.
- Considered to be more relevant in recent years



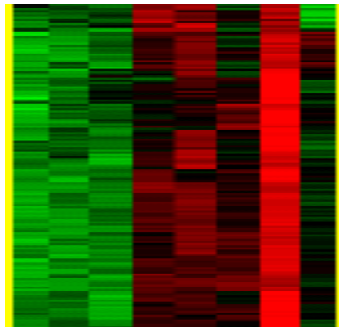
In 2016 researchers have access to studying disease at many levels



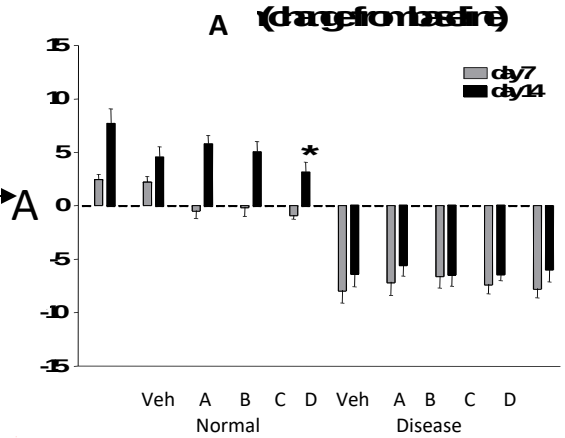
H NMR metabolites



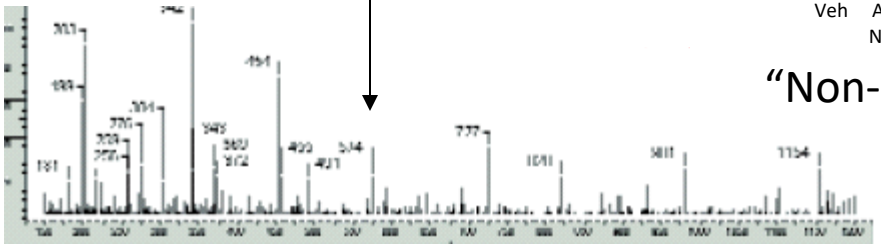
“NGS-Sequences



Affy Transcriptome

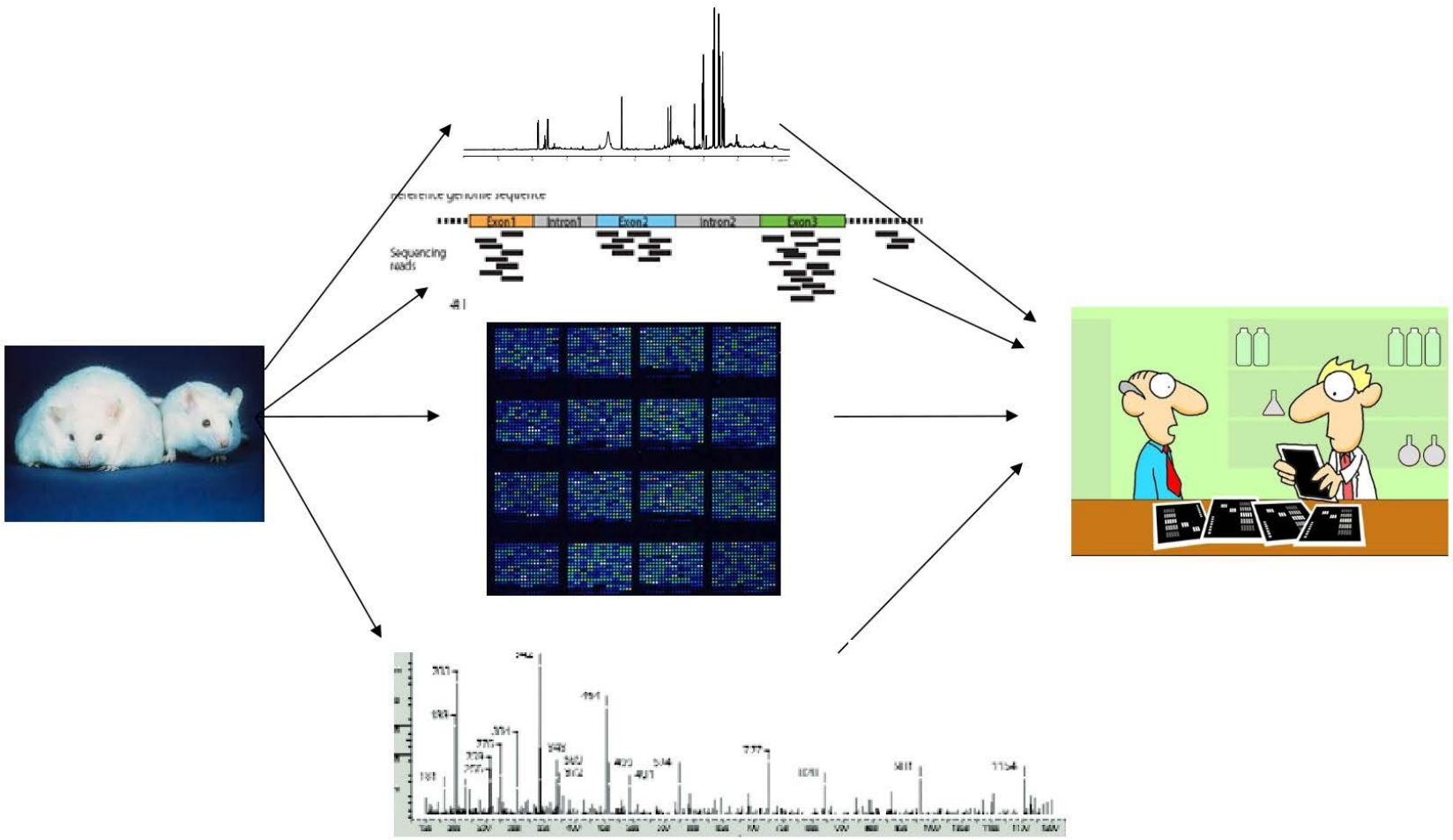


“Non-omic” markers

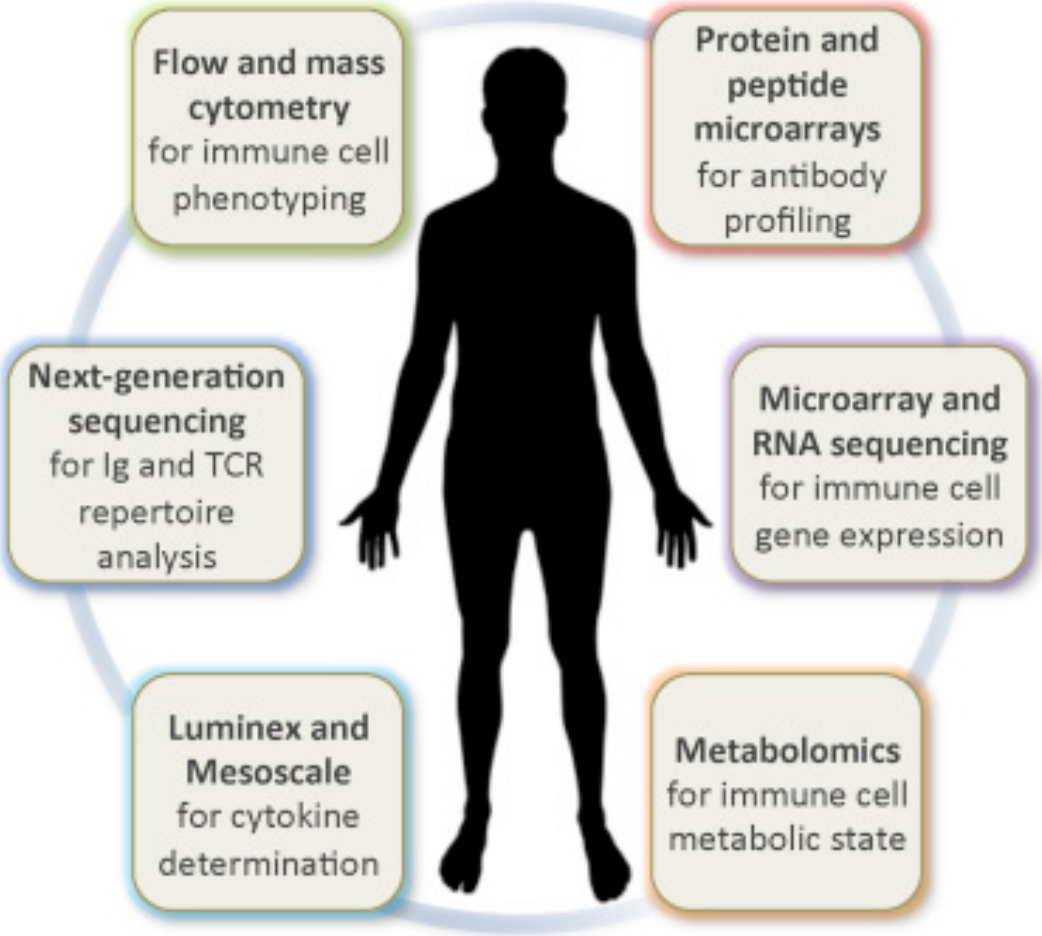


LC-MS proteomicss

The new challenge: IntegrOmics

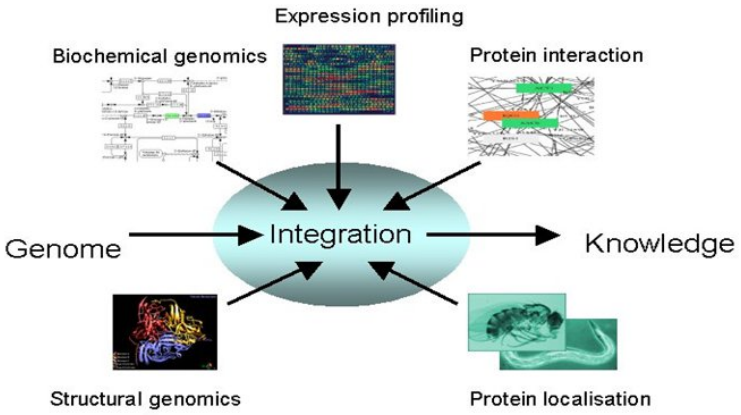


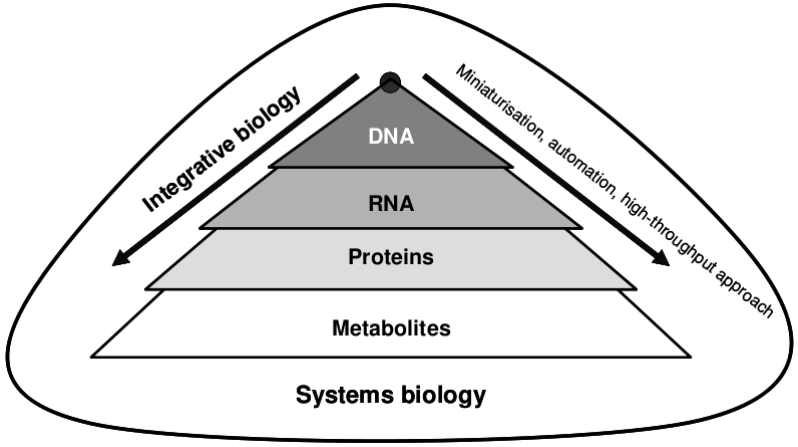
Multi-level high-throughput analysis of the human immune system



Furman D[1], Davis MM[2]. Vaccine. 2015 Sep 29;33(40):5271-81

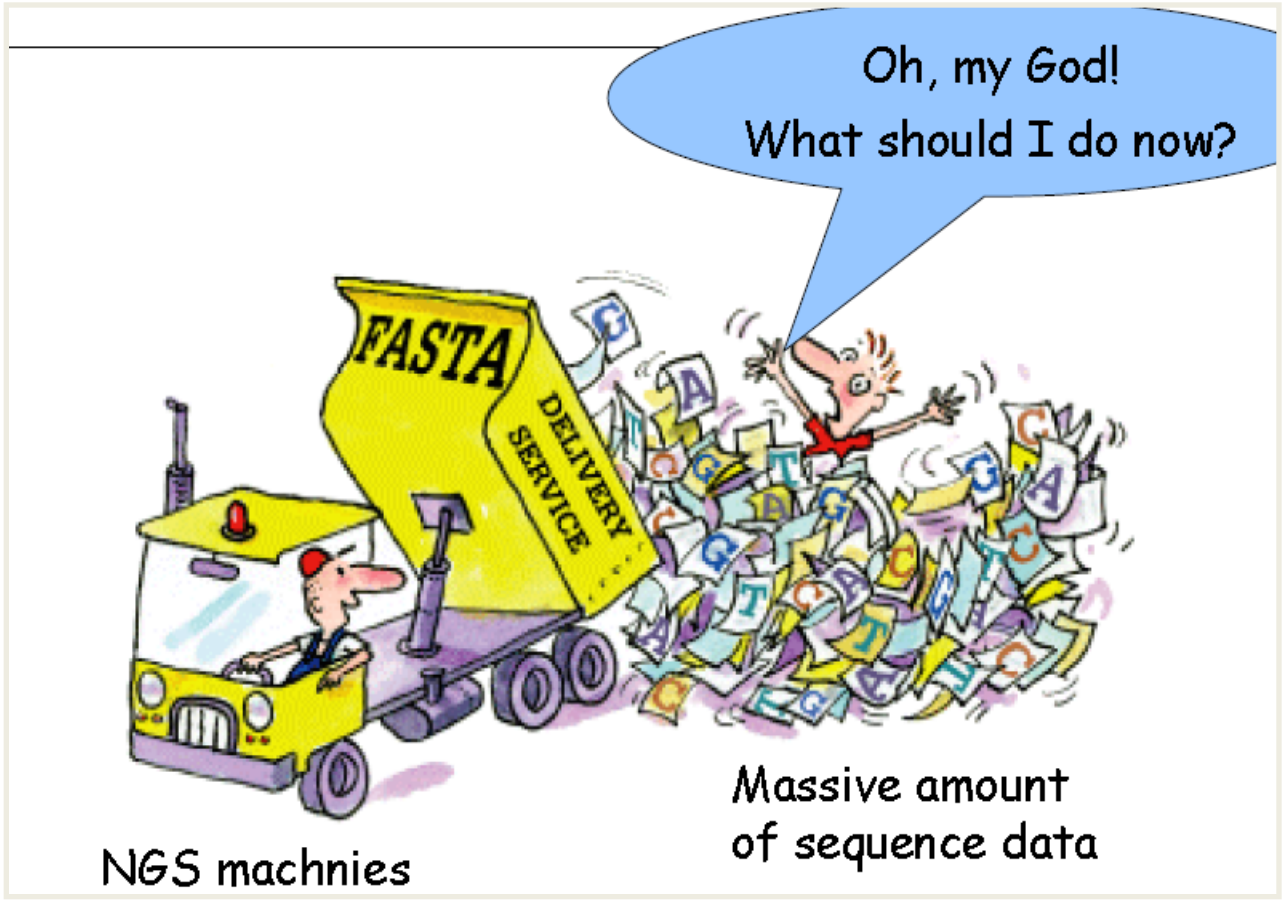
Integromics : a Systems Biology approach





- It is expected that the integrated collection and analysis of diverse types of data,
- jointly modelled and analyzed in a systems biology approach can shed light on the global functioning of the immune system.

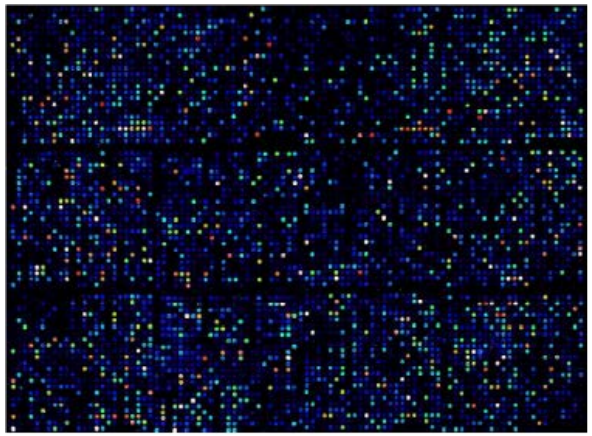
How can we deal with the data?

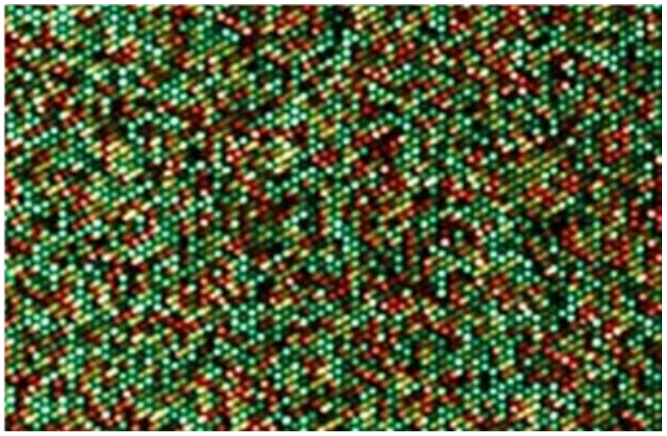


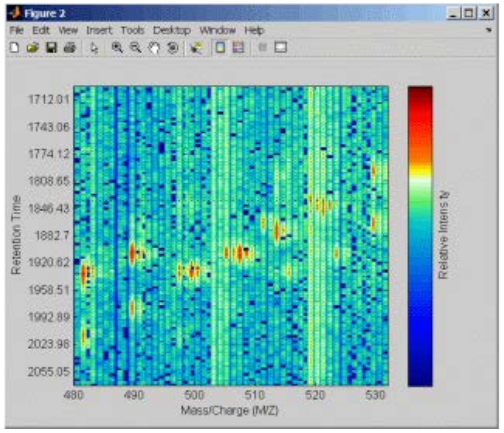
‘omics’ are high throughput

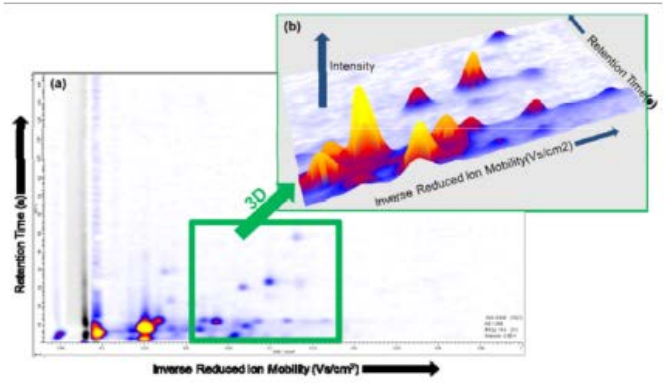
- Most ‘omic’ approches generate huge quantities of data.
- The management, storage, analysis and interpretation of these high throughput datasets cannot be conceived without all type of computing and quantitative ressources
- ***Biostatistics and Bioinformatics are a must for omics sciences***

Omics Data are high throughput



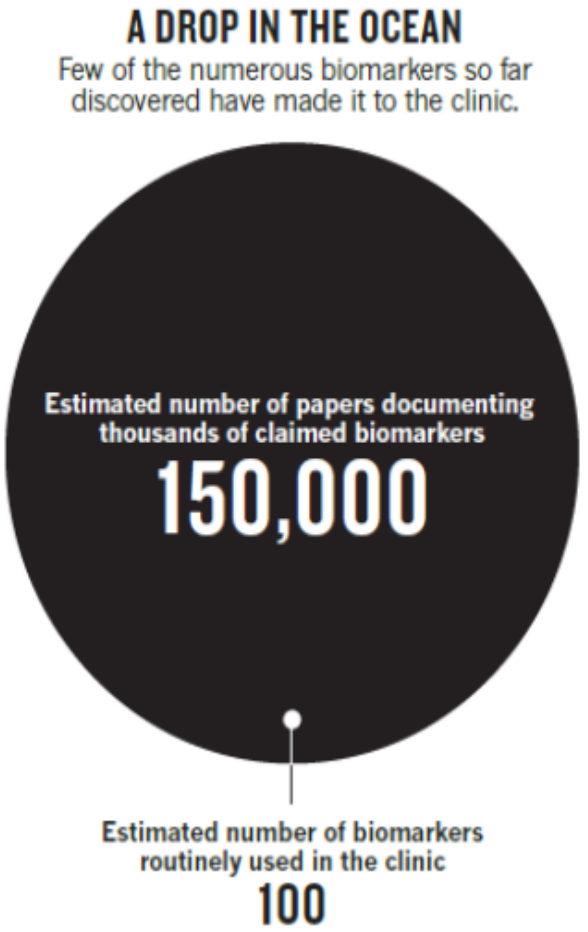






Biostatistics and Bioinformatics are a must for omics sciences

Not to talk of noise ...



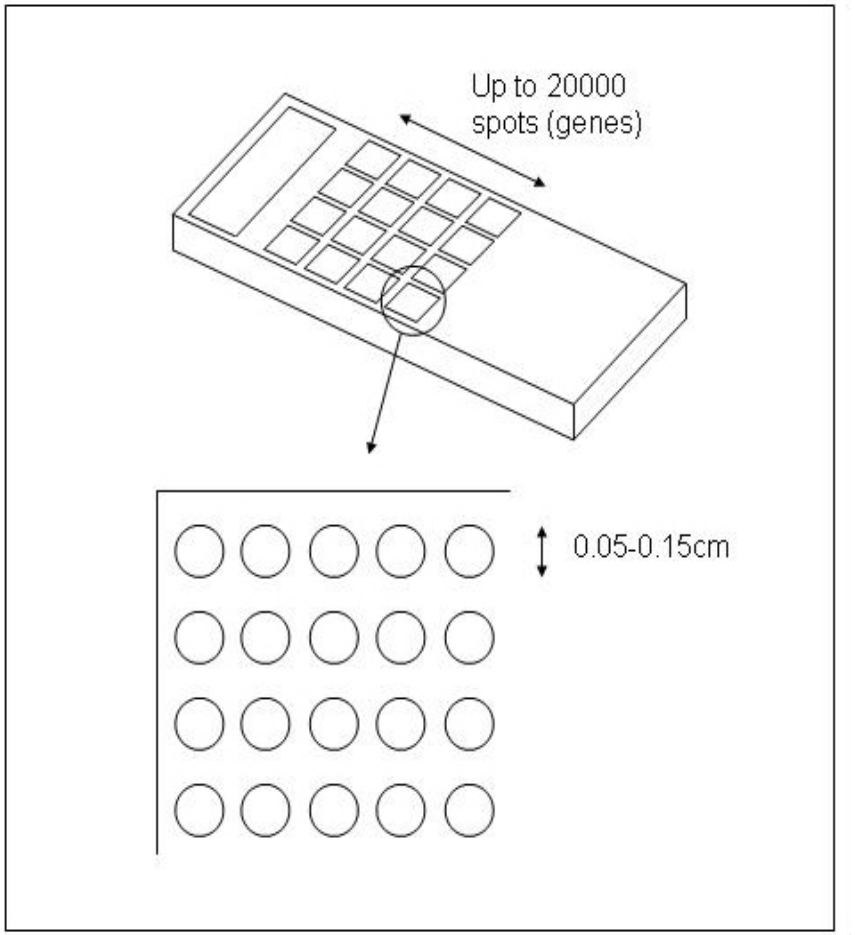
An Omics Analysis model

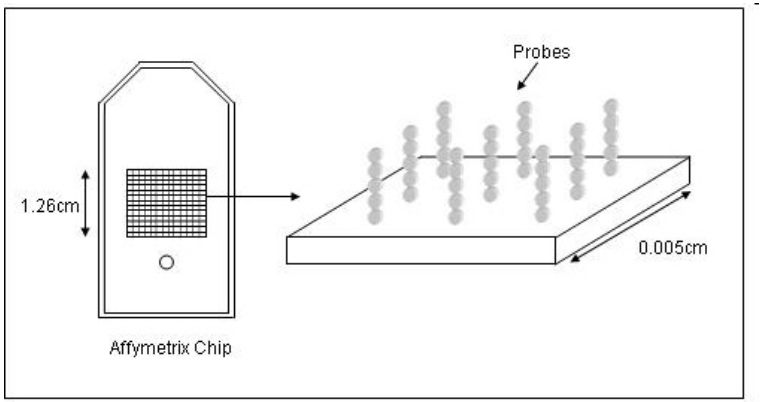
TranscriptOMIC analysis with microarrays or RNAseq

Using and analyzing genomes

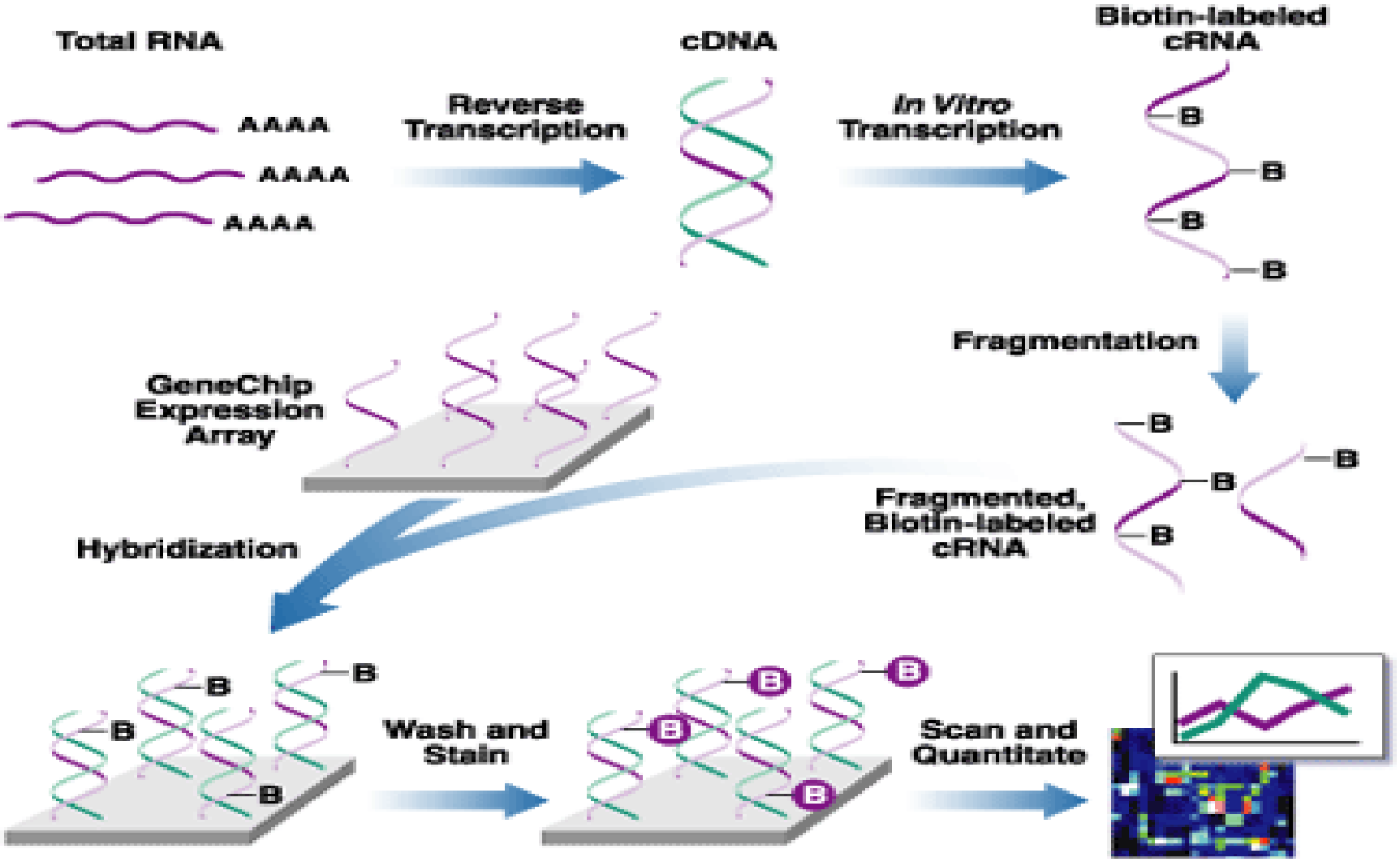
- Once genome sequences started being available new technologies appeared allowing the simultaneous study of all elements of an “ome”
- Microarrays
  - Study the expression of all genes simultaneously.
- (Next Generation) Sequencing
  - Sequence everything that’s available in:
    - Genomes
    - Transcriptome: RNAseq
    - Regulome
      - Epigenome: Bisulfite sequencing
      - MicroRNA: RNAseq

What is a microarray



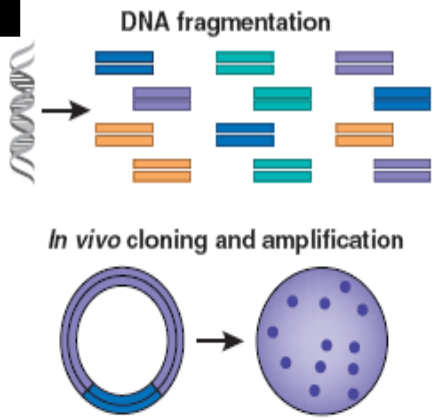


In the lab: how do microarrays work

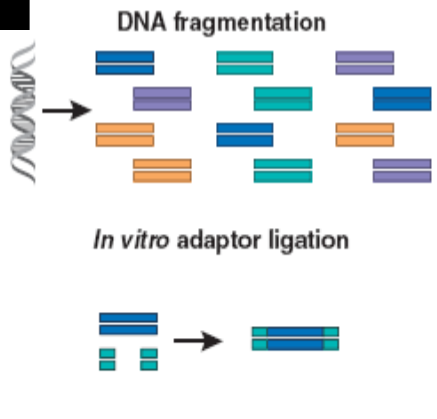


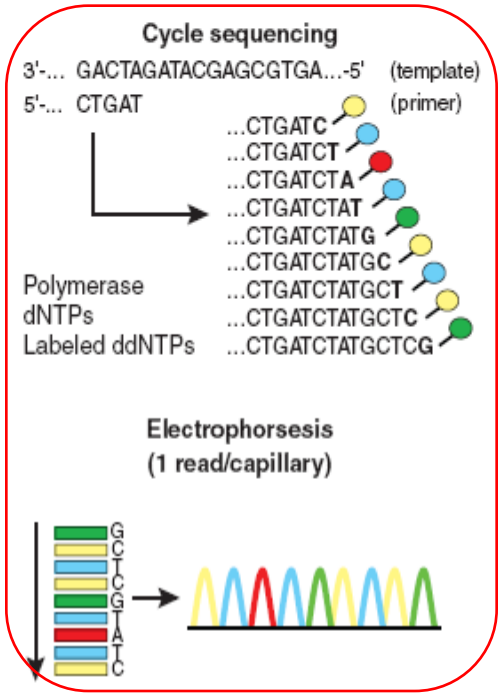
What is Next-generation DNA sequencing

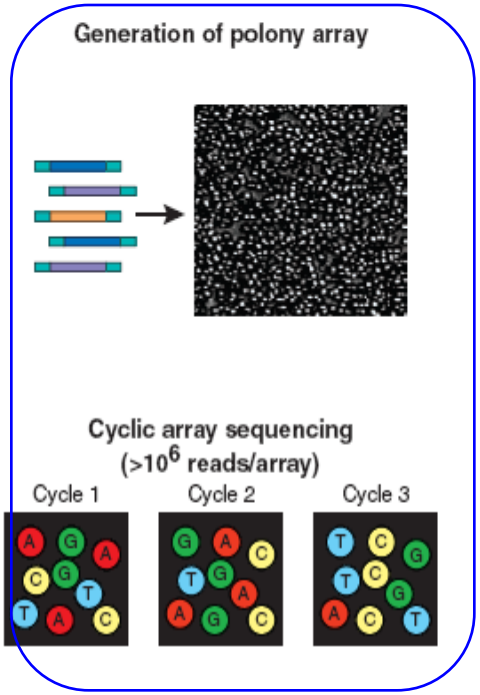
Sanger sequencing



Next-generation sequencing







Advantages:

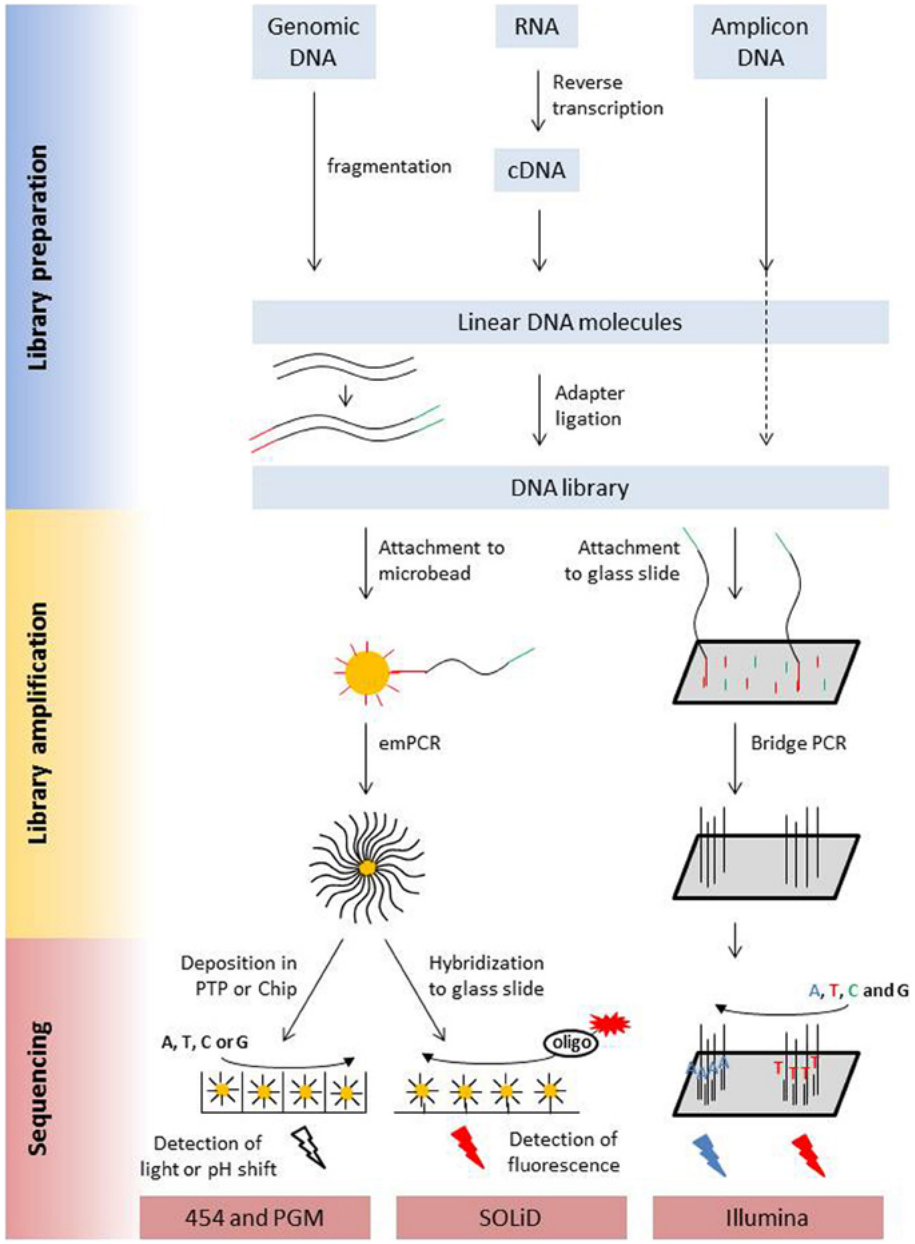
- Construction of a sequencing library → clonal amplification to generate sequencing features

  ✓ No in vivo cloning, transformation, colony picking…

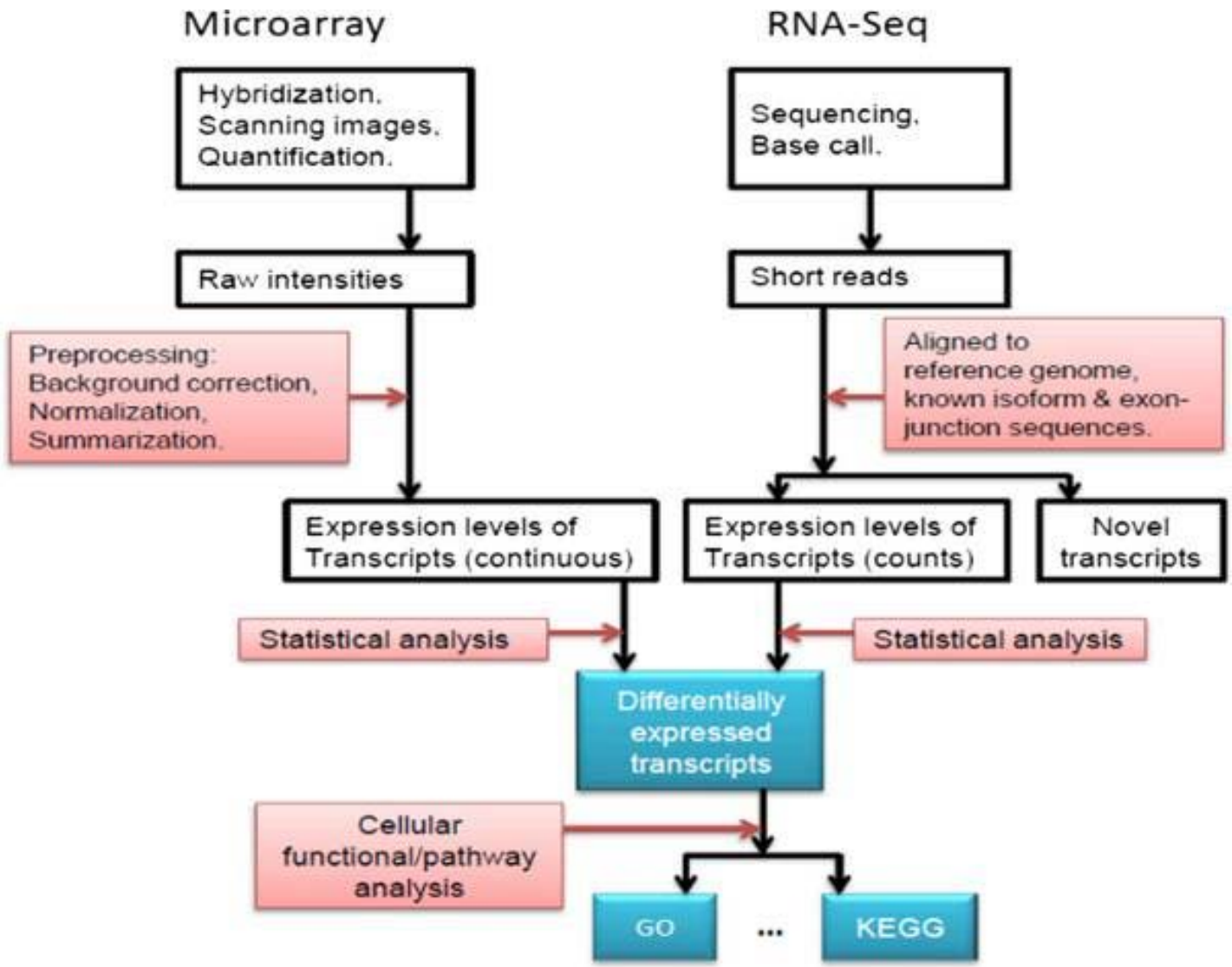
- Array-based sequencing

  ✓ Higher degree of parallelism than capillary-based sequencing

In the lab: Sequencing Genomes, RNA or Amplicons

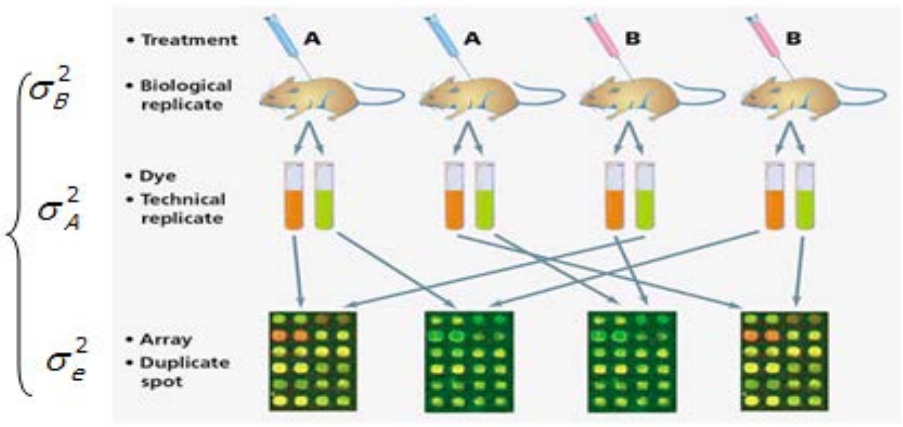


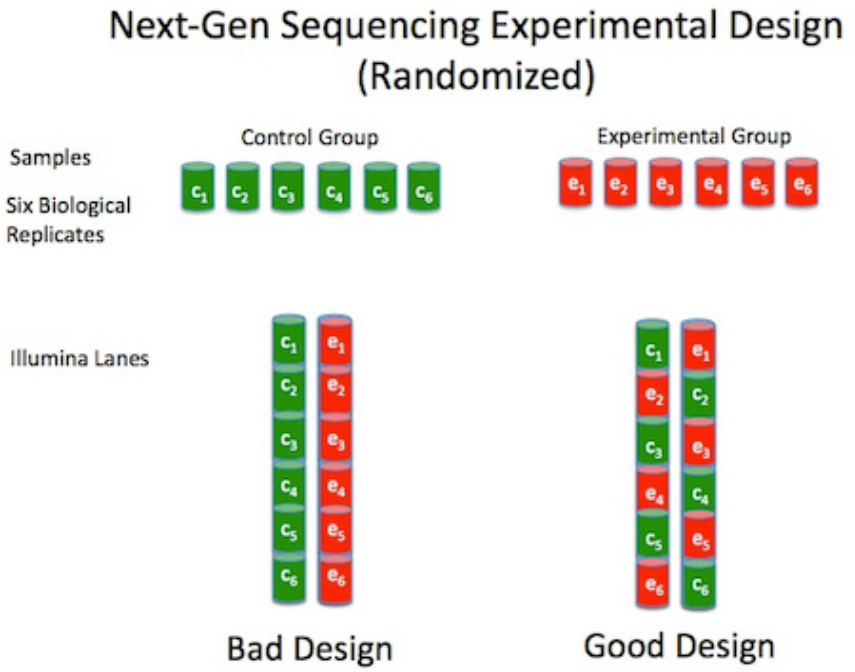
The (omics) data analysis process



Fang et al. *Cell Biosci. 2012; 2: 26.*

(0) Experimental design





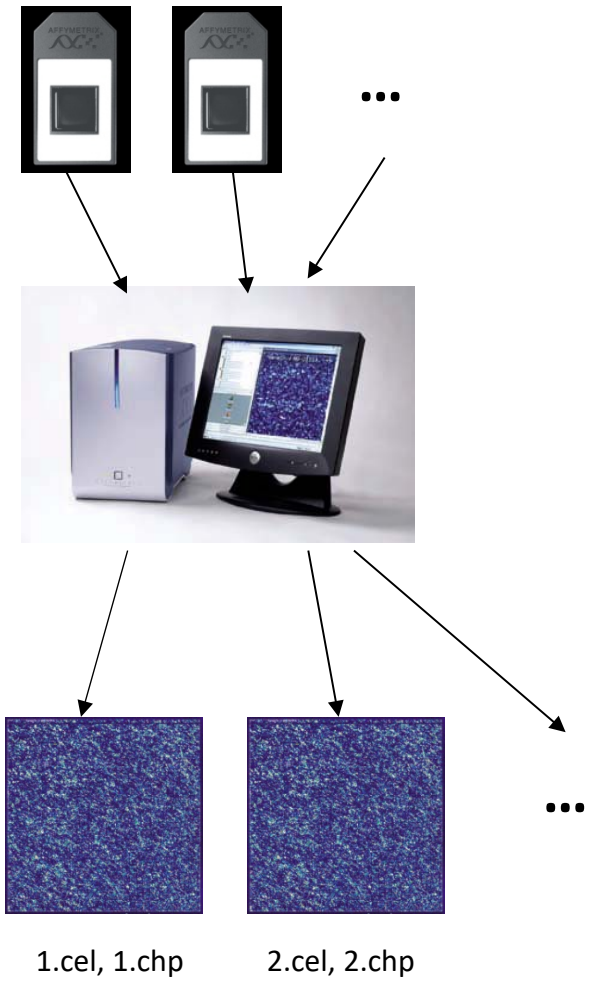
	Awful design :-(				Balanced design :-)		
Sample	Treatment	Sex	Batch	Sample	Treatment	Sex	Batch
1	A	Male	1	1	A	Male	1
2	A	Male	1	2	A	Female	2
3	A	Male	1	3	A	Male	2
4	A	Male	1	4	A	Female	1
5	B	Female	2	5	B	Male	2
6	B	Female	2	6	B	Female	1
7	B	Female	2	7	B	Male	1
8	B	Female	2	8	B	Female	2

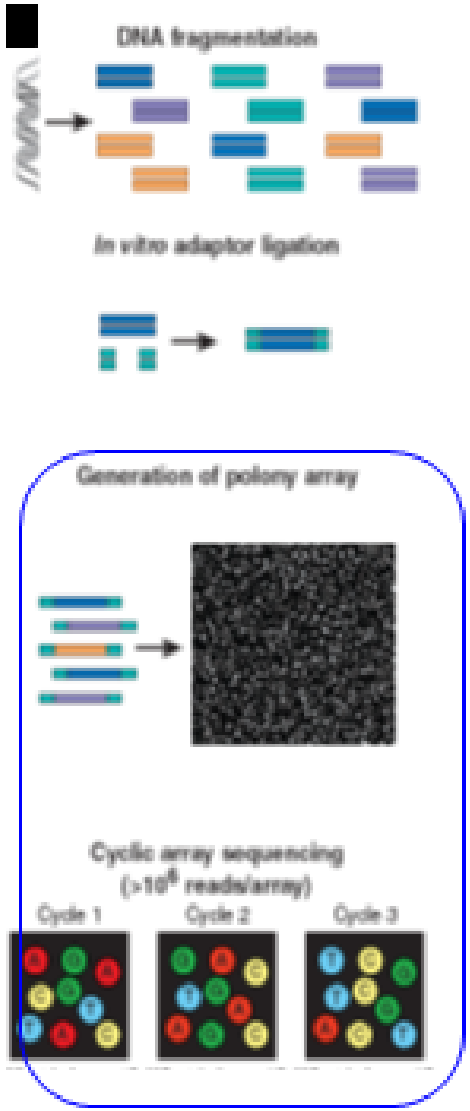
 # replicates?, pooling?, platform? , array type?

 # reads?, # replicates?, platform?, multiplexing?

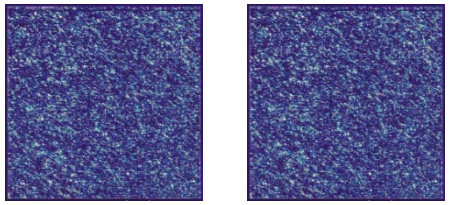
Randomization, Replication, Blocking, ...

(1) Image obtention





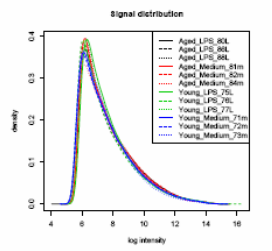
(2) Low level quality control

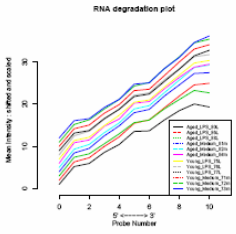


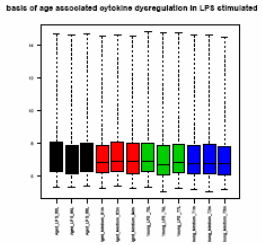
...

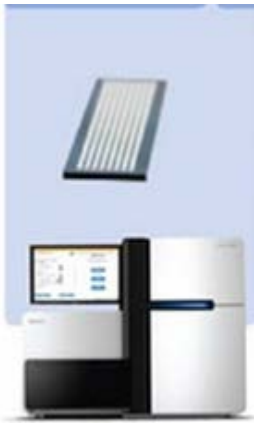
1.cel, 1.chp 2.cel, 2.chp

Affymetrix Expression Console™ Software

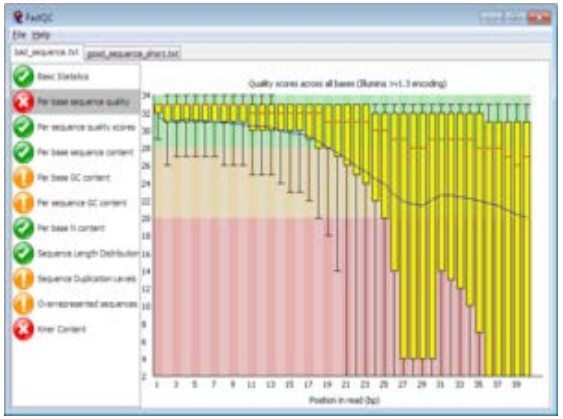




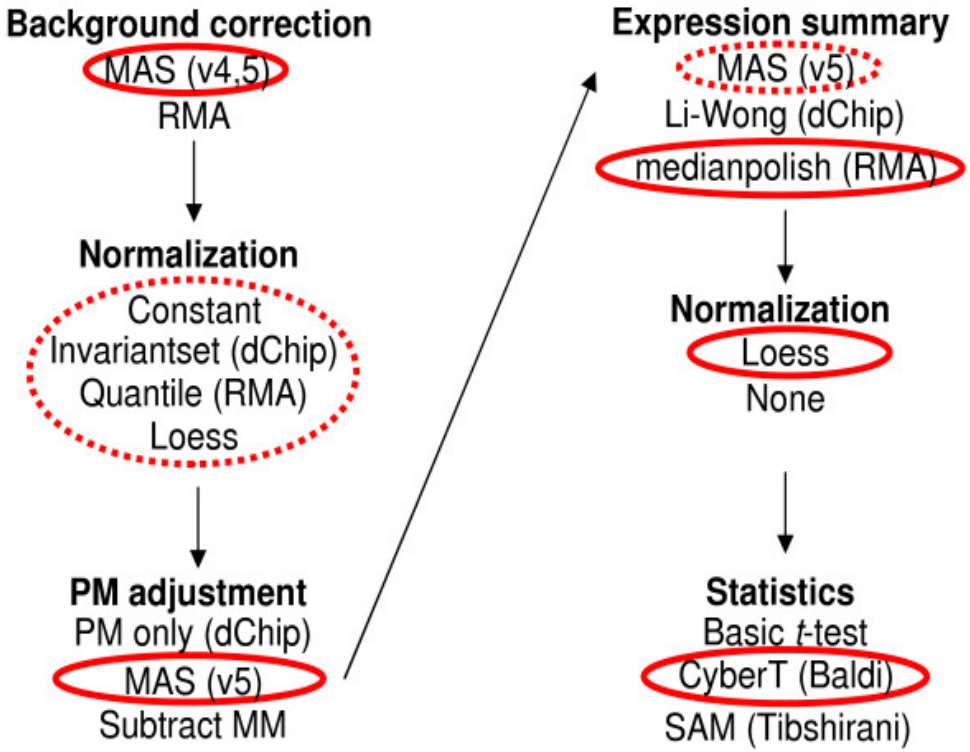




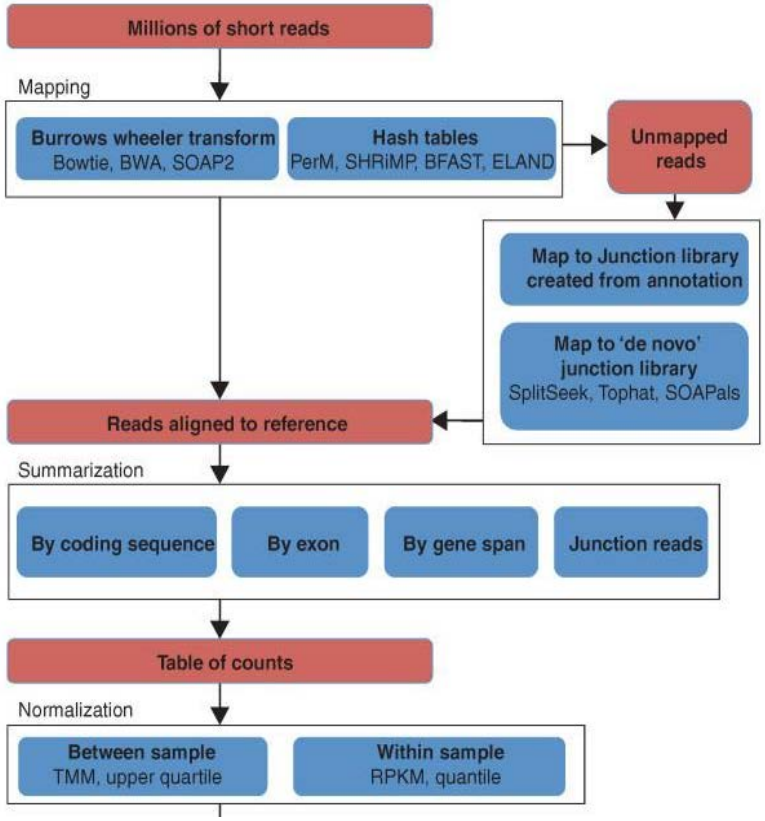
FastQC



(3) Preprocessing: Normalization, Summarization & Filtering



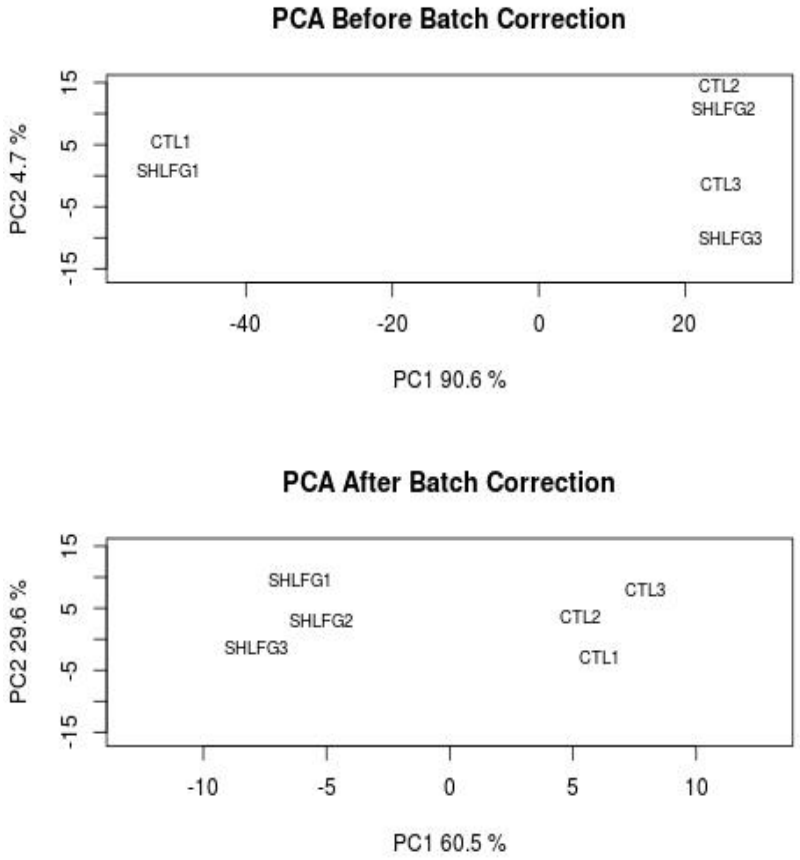
Preferred analysis methods for Affymetrix GeneChips
*Genome **Biology*** 2005, **6:**R16



From RNA-seq reads to differential expression results
Genome Biology 2010, **11**:220

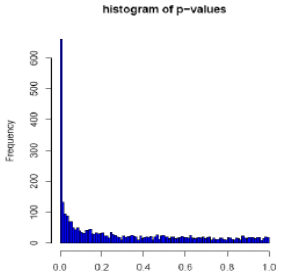
(4) Exploratory Data Analysis

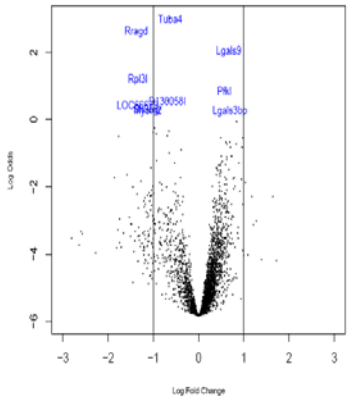
- Goal: Visualize data in order to highlight possibly (uncontrolled) experimental issues such as batch effects.
- Batch effect can be
  - Controlled (if predicted) with experimental design
  - Removed (in balanced designs) with statistical approaches

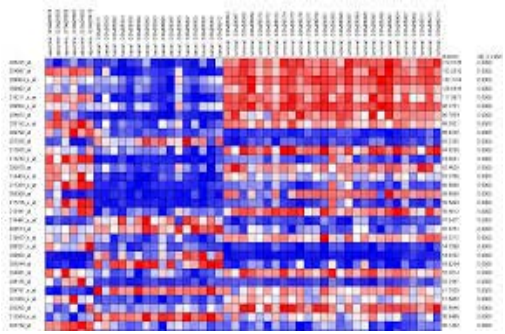


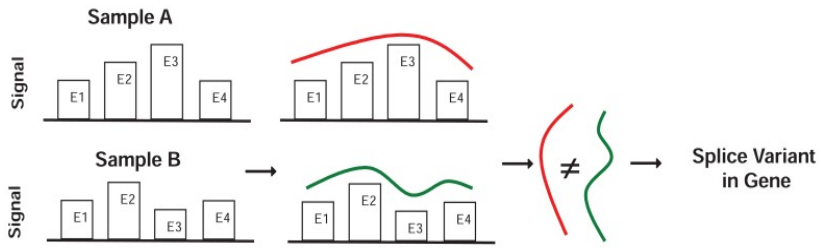
(5) Statistical analysis (1): Differentially expressed *what*

$$t = \frac{R_g}{\sqrt{\frac{d_0 \cdot SE_0^2 + d \cdot SE_g^2}{d_0 + d}}}$$









$$\log(Y) \mid T, E, P = \mu + T + P + E + T * E + S(P * T) + \varepsilon$$

Generalised linear model of the negative binomial family

$$N_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_{ij})$$ **Noise part**

$$\log \mu_{ij} = s_j + \sum_k \beta_{ik} x_{kj}$$ **Systematic part**

μ_{ij} expected count of gene i in sample j
s_j library size effect
x_{kj} design matrix
β_{ik} (differential) expression effects for gene i

DEXSeq

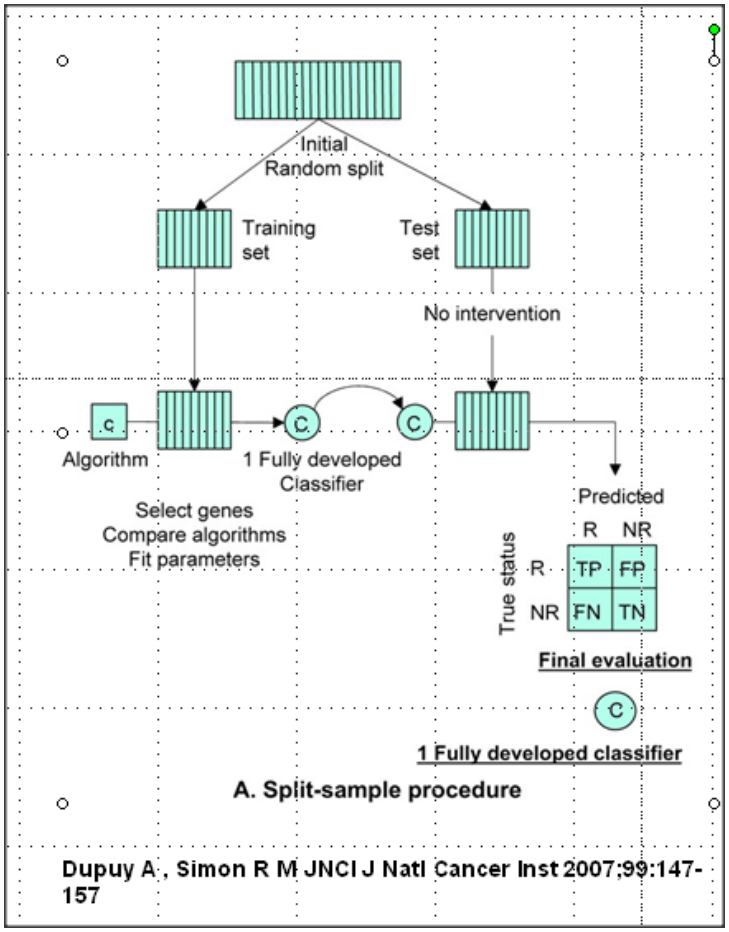
$$K_{ijl} \sim \mathrm{NB}(s_j \mu_{ijl}, \alpha_{il})$$

counts in gene i, sample j, exon l — size factor — dispersion

$$\log \mu_{ijl} = \beta_i^0 + \beta_{il}^{\mathrm{E}} x_l^{\mathrm{E}} + \beta_{ij}^{\mathrm{T}} x_j^{\mathrm{T}} + \beta_{ijl}^{\mathrm{ET}} x_l^{\mathrm{E}} x_j^{\mathrm{T}}$$

- expression strength in control
- fraction of reads falling onto exon l in control
- change in expression due to treatment
- change to fraction of reads for exon l due to treatment

(5a) Statistical analysis (ii):
Building and validating a predictor



Input:

- Expression matrix

Process

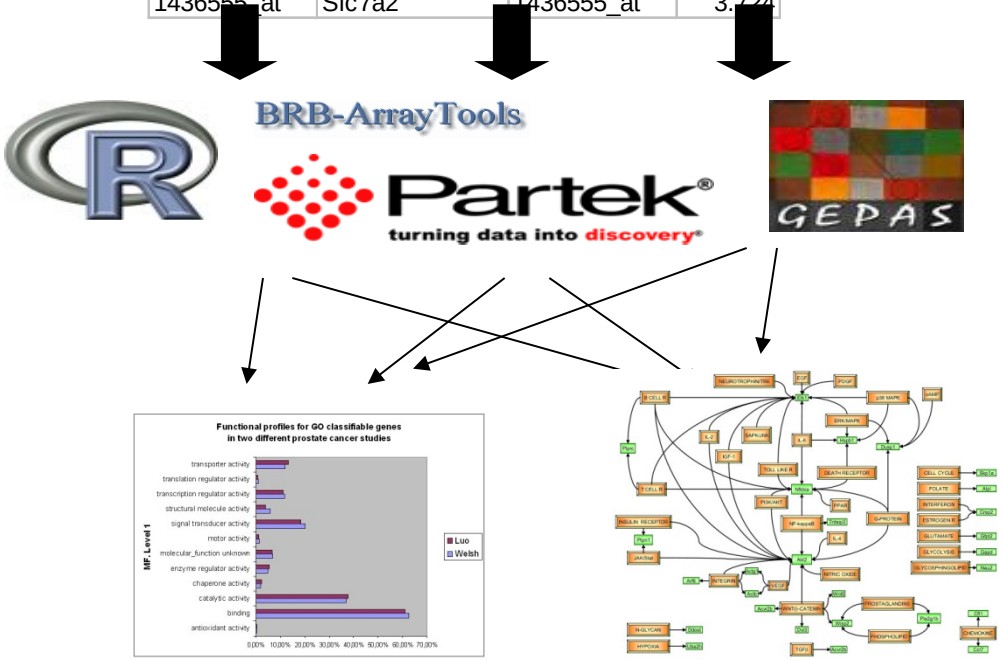
- Variable selection
- Model fitting
- Validation

Output

- Predictive models
- Measures of sensitivity and reproducibility

(6) Biological significance

ProbeSet	gene	ID	logFC
1450826_a_at	Saa3	1450826_a_at	4.911
1457644_s_at	Cxcl1	1457644_s_at	4.286
1415904_at	Lpl	1415904_at	-4.132
1449450_at	Ptges	1449450_at	5.164
1419209_at	Cxcl1	1419209_at	5.037
1416576_at	Socs3	1416576_at	3.372
1450330_at	Il10	1450330_at	4.519
1455899_x_at	Socs3	1455899_x_at	3.648
1419681_a_at	Prok2	1419681_a_at	3.709
1436555_at	Slc7a2	1436555_at	3.724



Input

- Gene lists

Process

- GEA, GSEA,
- Network analysis

Output:

- Relevant GO or KEGG terms
- Relevant pathways
- Networks

Potential applications in MS

- To identify transcriptional differences
  - between MS patients and healthy controls, or
  - between different clinical forms or activity phases of the disease
- To identify differentially expressed genes at the time of the first neurological event suggestive of demyelinating disease
- To identify molecular pathways involved in the inflammatory and neurodegenerative processes taking place in the CNS of MS patients.
- To investigate the transcriptional changes associated with the effects of therapies used in the treatment of patients with MS.

Example applications in MS[1]

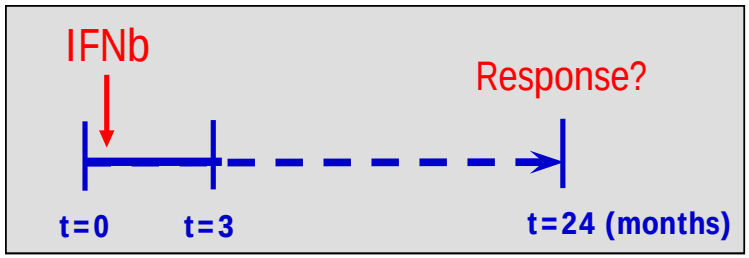
doi:10.1093/brain/awp228 Brain 2009: 132; 3353–3365 | 3353

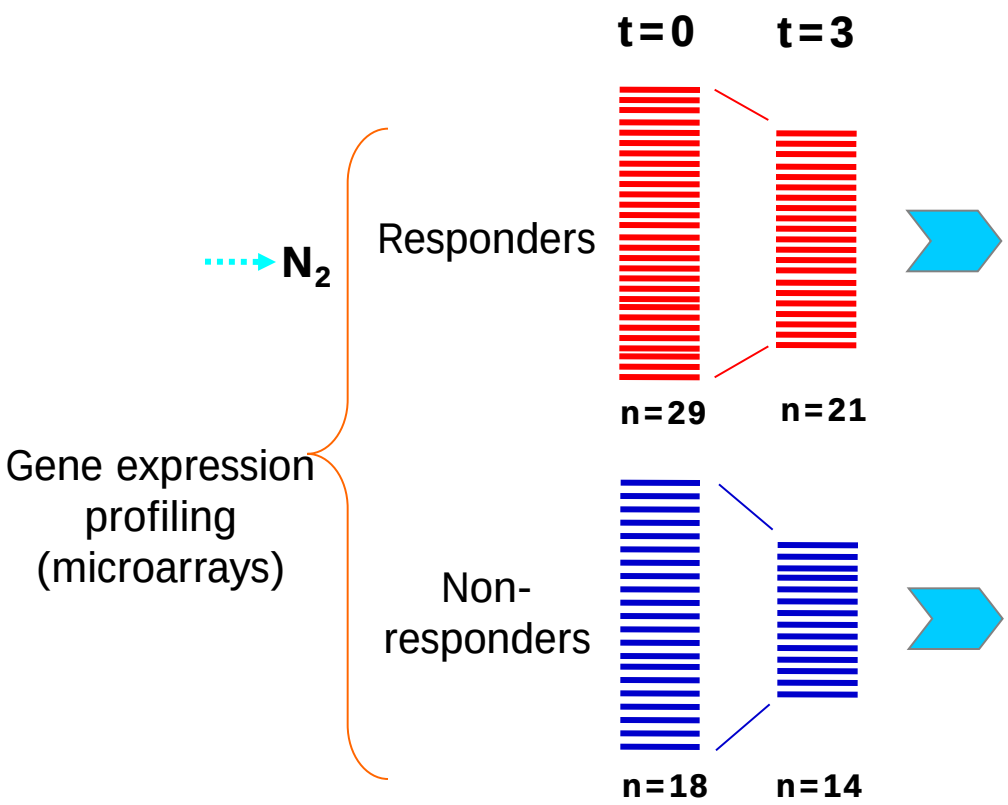
BRAIN
A JOURNAL OF NEUROLOGY

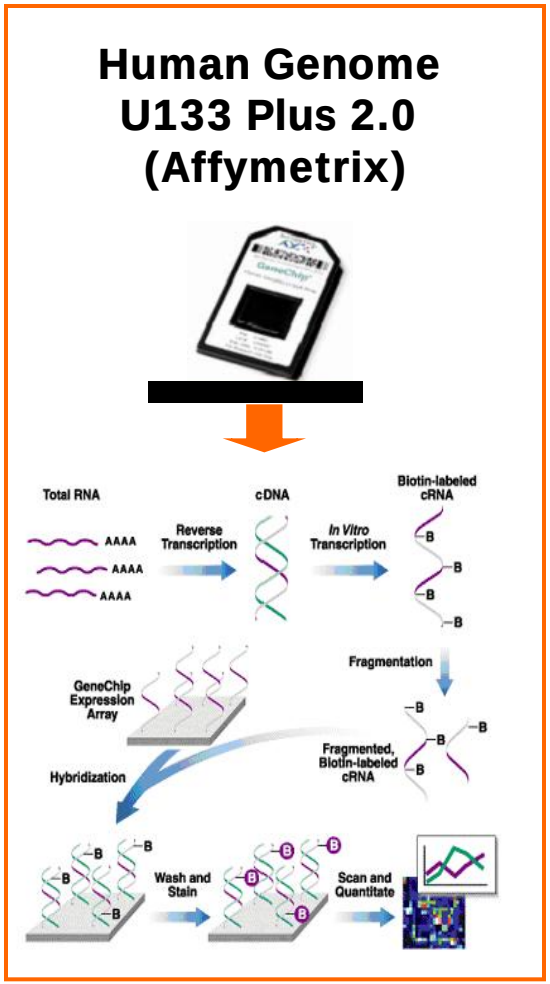
A type I interferon signature in monocytes is associated with poor response to interferon-β in multiple sclerosis

M. Comabella,[1] J. D. Lünemann,[2,†] J. Río,[1] A. Sánchez,[3] C. López,[1] E. Julià,[1] M. Fernández,[1] L. Nonell,[1] M. Camiña-Tato,[1] F. Deisenhammer,[4] E. Caballero,[5] M. T. Tortola,[5] M. Prinz,[6] X. Montalban[1,*] and R. Martin[7,*,‡]

Prediction of response to IFNβ treatment





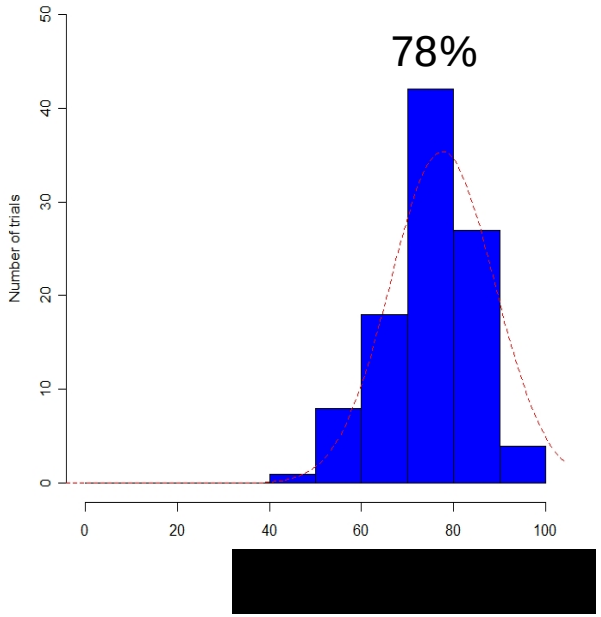


Prediction of response to IFNβ treatment

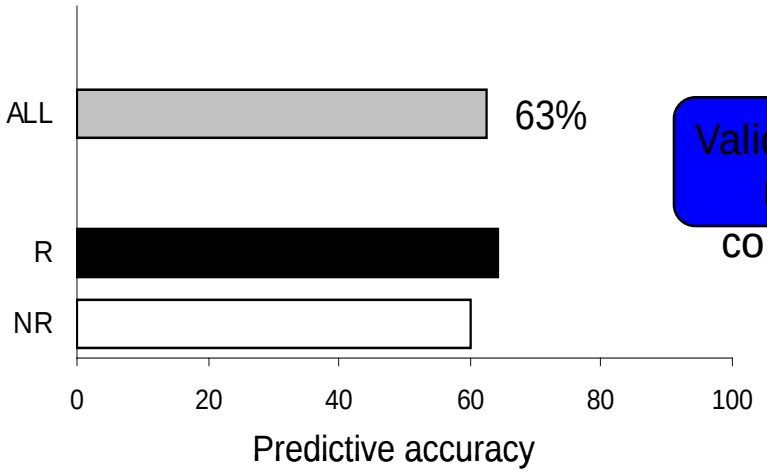
Type I IFN signature

Affymetrix probe set	*Gene Symbol*
229450_at	IFIT3
203153_at	IFIT1
205660_at	OASL
230233_at	RASGEF1B
214059_at	IFI44
217502_at	IFIT2
208962_s_at	FADS1
201670_s_at	MARCKS

Predominantly induced by type I IFNs



Original cohort



Validation cohort

Example application (2)

Technological Innovation and Resources

Protein-Based Classifier to Predict Conversion from Clinically Isolated Syndrome to Multiple Sclerosis*

Eva Borràs‡§, Ester Cantó¶, Meena Choi‖, Luisa Maria Villar**, José Carlos Álvarez-Cermeño**, Cristina Chiva‡§, Xavier Montalban¶, Olga Vitek‖, Manuel Comabella¶‡‡, and Eduard Sabidó‡§‡‡

Example application (2)

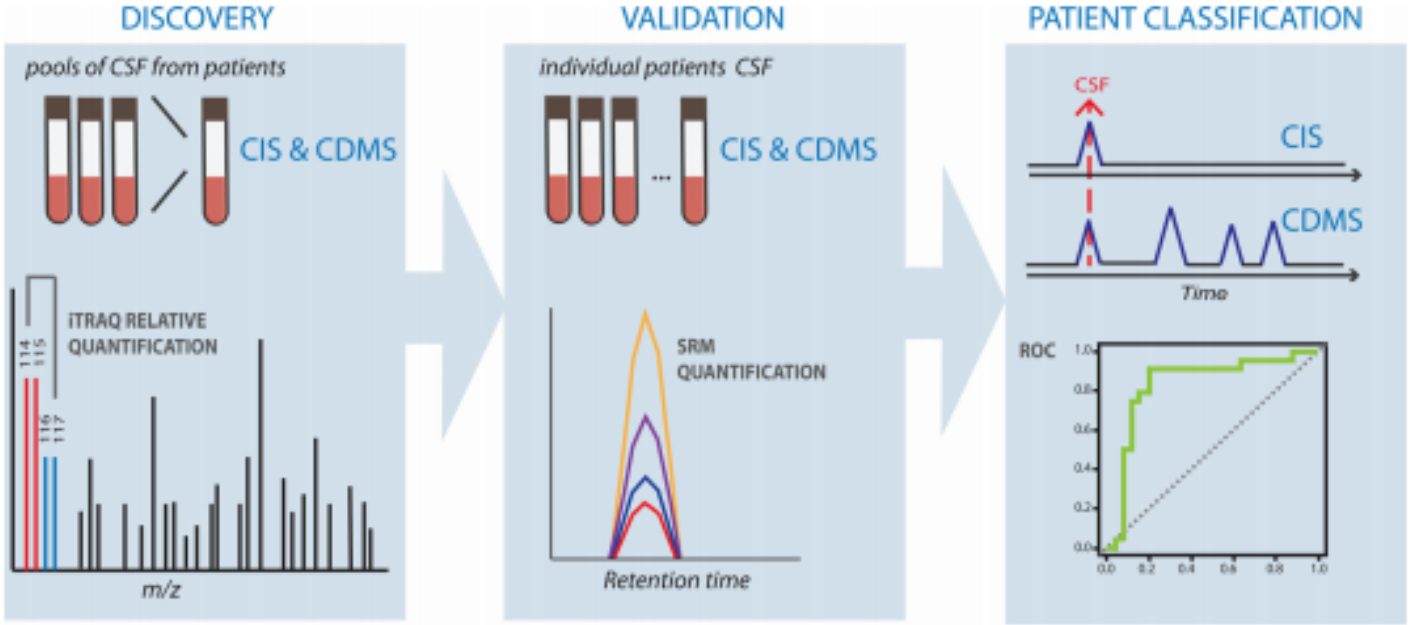


FIG. 1. **General workflow used in the present study.** Initially, protein candidates identified in our previous discovery studies—together with several proteins described by other groups—were selected and quantified by targeted mass spectrometry (SRM) in a relatively large cohort individual patients. Protein quantities were then evaluated by their capability of classifying patients with clinical isolated syndrome, and thus, the best prognostic protein combination was identified.

Transcriptional differences between MS and HC

Subjects studied	Arrays used	Summary	Reference
RRMS (n=15) / **HC** (n=15)	4,329 cDNA clones GeneFilters GF211 DNA array	34 genes discriminated between RRMS patients and HC	*Ramanathan et al. 2001*
MS (n=4) / **IDDM** (n=5) / **RA** (n=20) / **SLE** (n=24) / **HC** (n=9)	4,329 cDNA clones GeneFilters GF211 DNA array	Gene expression profiling was similar among autoimmune disorders	*Maas et al. 2002*
MS (n=24) / **HC** (n=19)	cDNA clones from ResGen and lymphochip	Identification of pairs of genes whose expression discriminated well between MS and HC	*Bomprezzi et al. 2003*
HC (n=18) / **RRMS** (n=26: 12 in relapse and 14 in remission)	12,000 genes, Affymetrix Human Genome U95Av2	Identification of 1,109 genes whose expression was significantly different between MS and HC irrespective of treatment and disease activity	*Achiron et al. 2004*
MS (n=17) / **HC** (n=7)	6,800 genes, Affymetrix HuGeneFL	Upregulation in MS patients of multiple genes belonging to the E2F pathway	*Iglesias et al. 2004*
RRMS (n=13) / **SLE** (n=5) **HC** (n=18)	12,000 genes, Affymetrix Human Genome U95Av2	A common autoimmunity signature containing 541 genes differentiated between MS and SLE patients and HC	*Mandel et al. 2004*
MS (n=72: 65 **RRMS** and 7 **SPMS**) / **HC** (n=22)	1,258 genes, Hitachi cDNA microarrays	173 genes in T cells and 50 genes in non-T cells differentially expressed between MS and HC with high representation of apoptosis-related genes	*Satoh et al. 2005*
RRMS (n=10) / **PPMS** (n=8) / **HC** (n=12)	Affymetrix Human Genome U95Av2	Decreased expression of CX3CR1 in the NK cell population of MS patients compared to HC	*Infante-Duarte et al. 2005*

Transcriptional differences between MS and HC

Subjects studied	Arrays used	Summary	Reference
RRMS (n=29; 9 patients were in relapse) / **HC** (n=25)	In-house 43K cDNA microarrays	A sub-population of RRMS patients exhibit an activated immune defense program that resembles a virus response program	*van Baarsen et al. 2006*
RRMS (n=42: 22 in relapse and 20 in remission)	22,000 genes, Affymetrix U133A2	Among DEG between relapse and remission there was a high representation of apoptotic-related genes mainly involved in the caspase-dependent pathway	*Achiron et al. 2007*
RRMS (n=63); 24 patients were on treatment	Genechip array U95Av2 and HU-133A	A predictive signature of clinical outcome was overepresented with genes related to zinc-ion binding and cytokine activity regulation pathways	*Achiron et al. 2007*

Discussion and conclusions

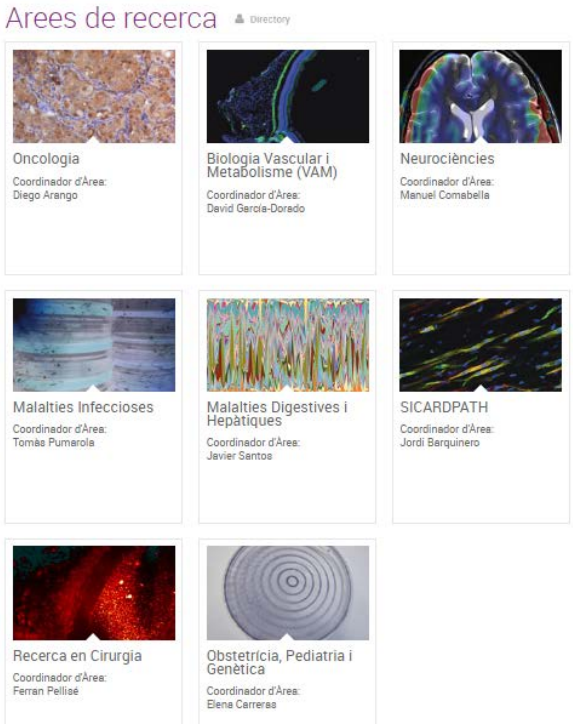
- Omics technologies enable global approaches to the study of biology.
  - A wider extempt with the cost of a lesser depth
- Many omics
  - Different technologies
  - Similar yield (numerical data tables)→ Similar analyses, similar issues to deal with
- Hardest part
  - Integrative analysis (integration of analyses)
  - Biological interpretation
- Still a long way to go …

Acknowledgements

UNIVERSITAT DE BARCELONA

GRBIO

Vall d'Hebron
Institut de Recerca



Integrative **O**mics **D**ata **A**nalysis

Thanks for your attention!