Bioinformatic analysis of ‘omic’ data in genetic epidemiology studies

Jornades de consultoria estadística i software II
UAB, Octubre 2013

Juan R Gonzalez
BRGE – Bioinformatics Research Group in Epidemiology
Center for Research in Environmental Epidemiology (CREAL)

Generalitat de Catalunya

Parc de Salut Barcelona MAR

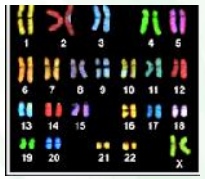
UNIVERSITAT POMPEU FABRA

OMICS

DISEASOME (PHENOTYPE)

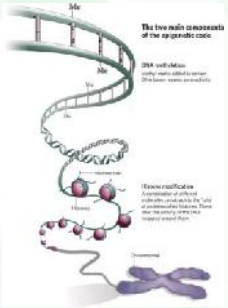
All the disorders and diseases of an organism, viewed as a whole

OMICS

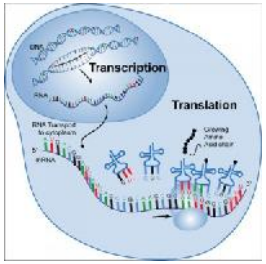


GENOME
hereditary information (DNA)
stable
>99% equal between individuals
1.5% coding genes

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

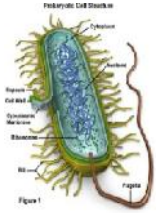
PROTEOME
tissue and time especific

METABOLOME
tissue and time especific

DISEASOME (PHENOTYPE)




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



OMICS

Why OMICS?

To identify biomarkers

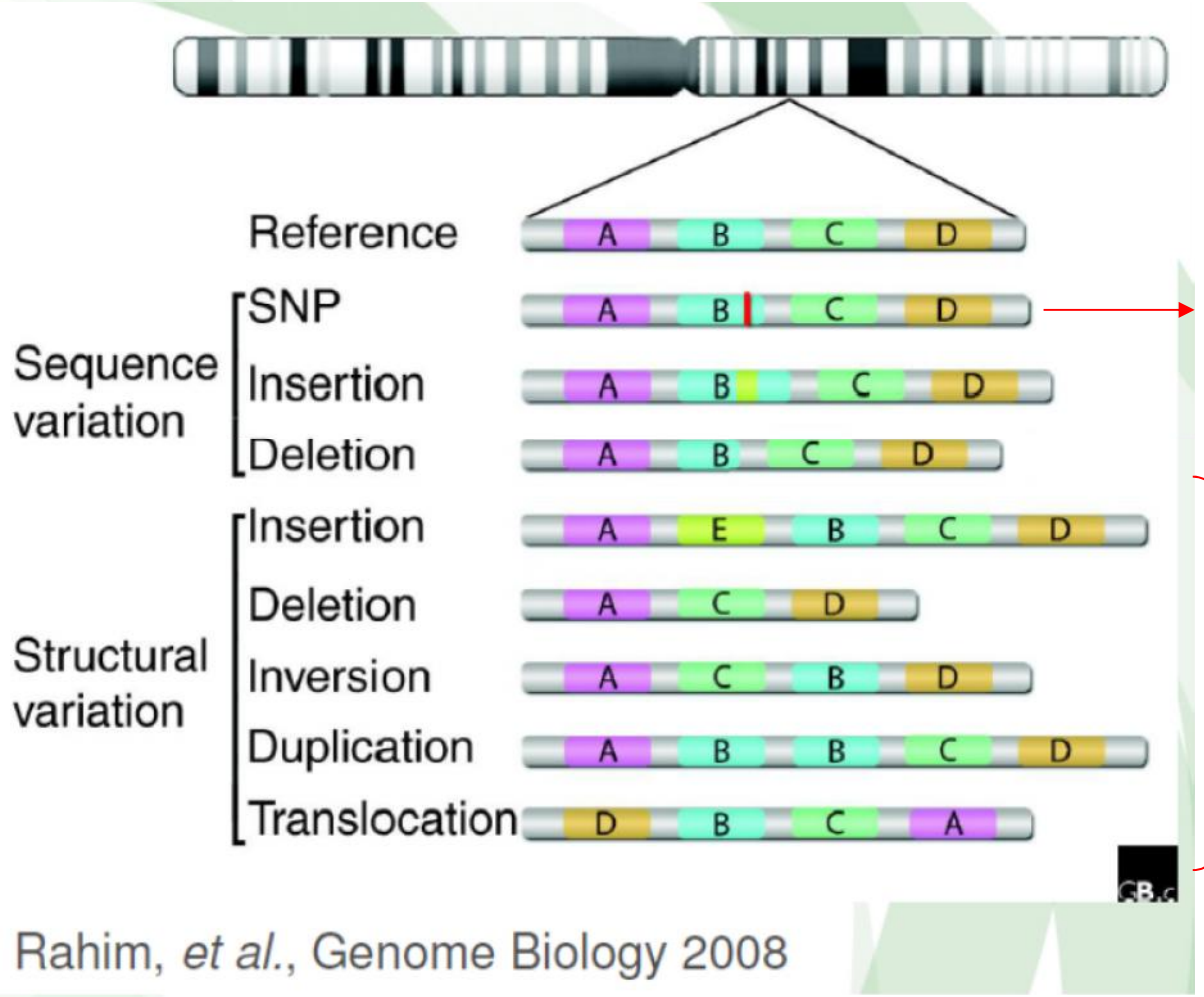
exposure
disease

To understand molecular mechanims

OMICS

- Metabolomic data [serum/plasma]
  - Mass-spectometry data (up to 2500 metabolites)
- Proteomic data [plasma]
  - Quantification of up to 30 proteins
- Transcriptomic data [DNA whole blood]
  - Gene expression and splice variant expression
  - Also information about regulatory non-coding RNAs (snc-RNAs and miRnAs)
- Epigenomic data [DNA whole blood]
  - DNA methylation data
- Genomic data [DNA whole blood] (Existing GWAS)
  - SNP array data

GENOMICS



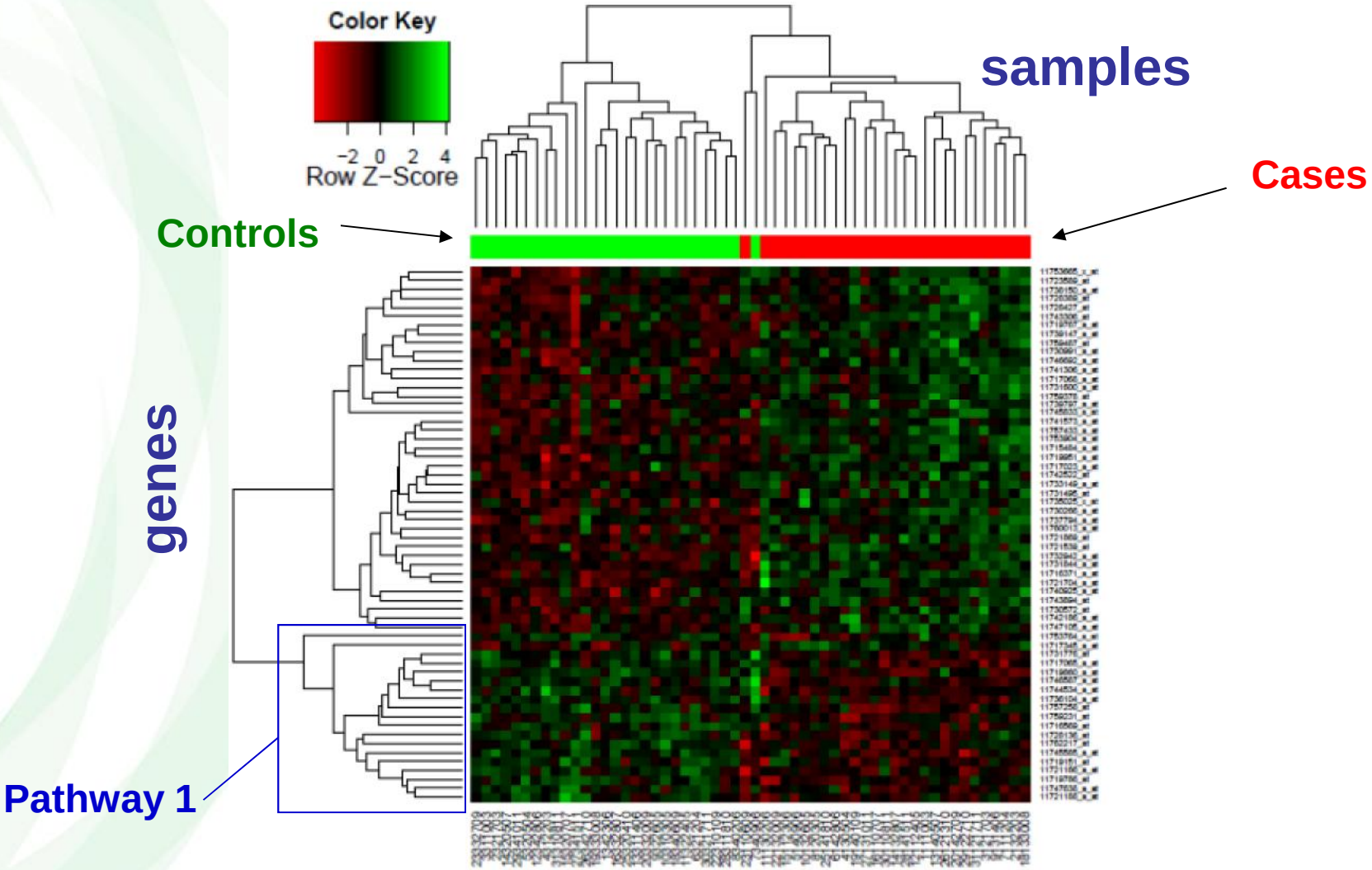
Rahim, *et al.*, Genome Biology 2008

SNP array data (GWAS)

Imputed will be analyzed
SNP -> 4,000,000 markers

Can be predicted using existing GWAS data

MOLECULAR PROFILES

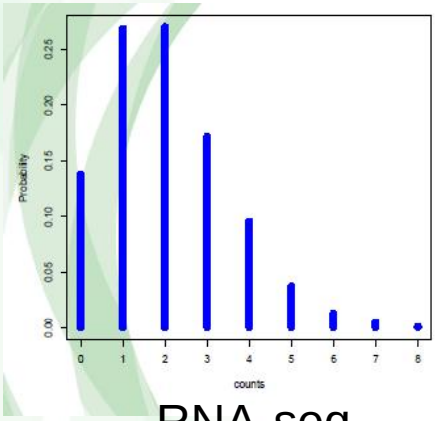


PIPELINE FOR OMIC DATA

- Data pre-processing
  - Quality control (filtering)
  - Normalization (control for batch effect)
- Statistical analysis
- Clustering and enrichment/pathway analysis
- Visualization and Annotation

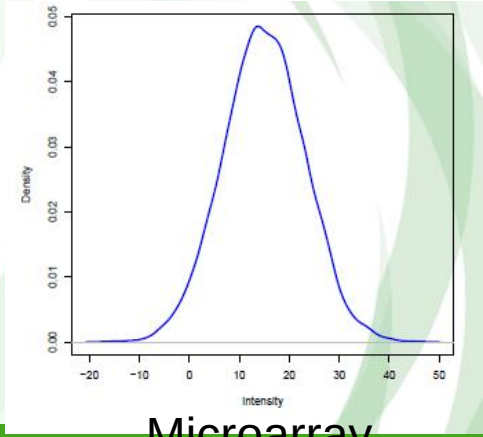
Step 2: Statistical Analysis

Transcriptome



RNA-seq

RNA-seq: 0, 1, 2, ….., 2456, …, 34567, …
Generalized linear models
(Negative Binomial)

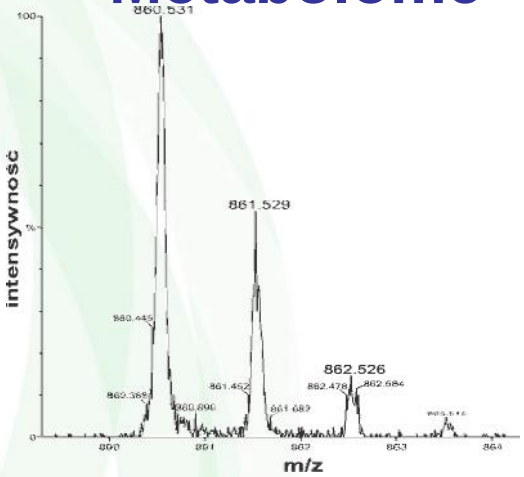


Microarray

Microarrays: 6.1, 6.9, 12.4, 11.4, 8.5, …
Generalized linear models
(Gaussian)

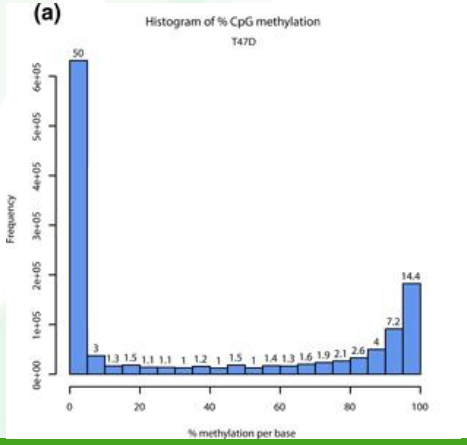
Step 2: Statistical Analysis

Metabolome



Peaks: 11234, 1353, 1234, 12, 455, 122, …
Clustering methods
(Non-parametric)

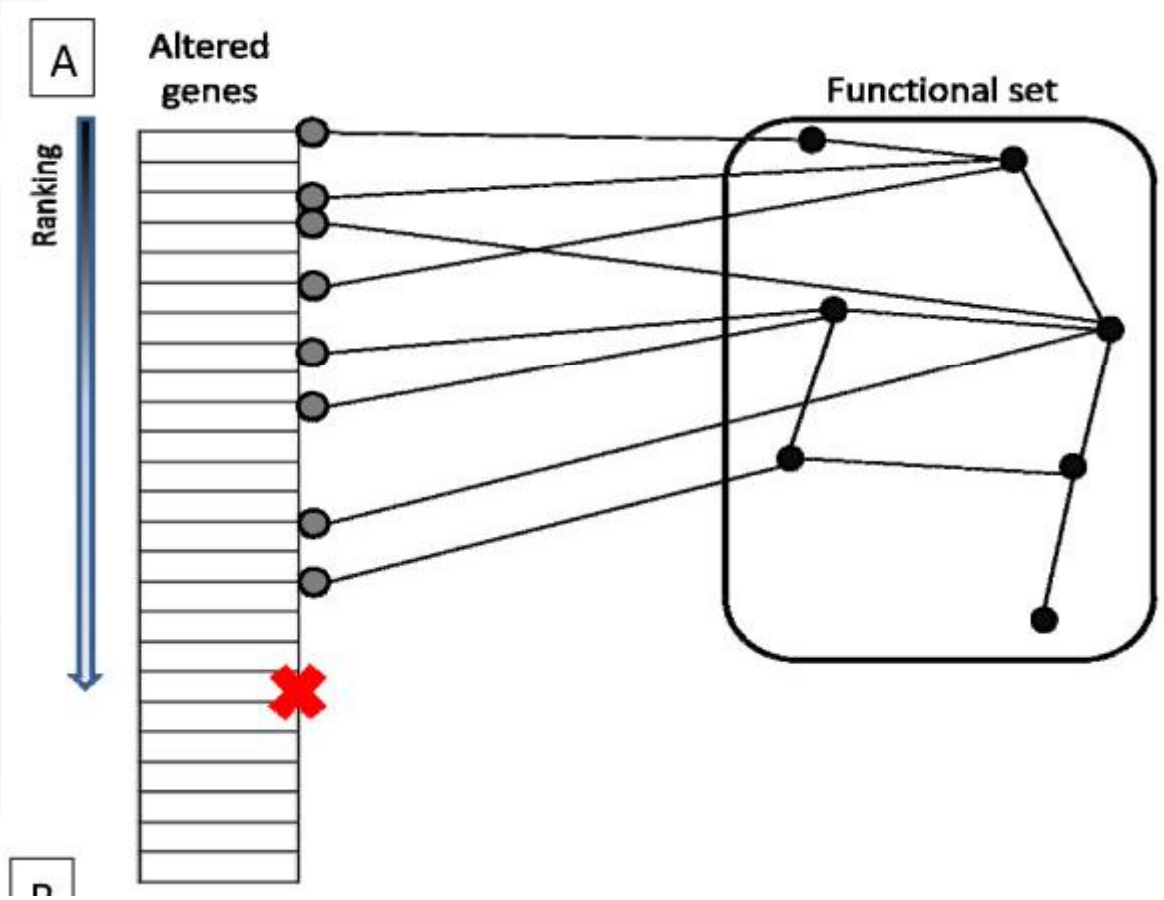
Epigenome



CpG islands: 1, 0, 1, 1, 0, 1, 0, …
Beta regression
(Beta distribution)

Step 3: Enrichment/pathway analysis

Gene set (Under- or over-expressed)

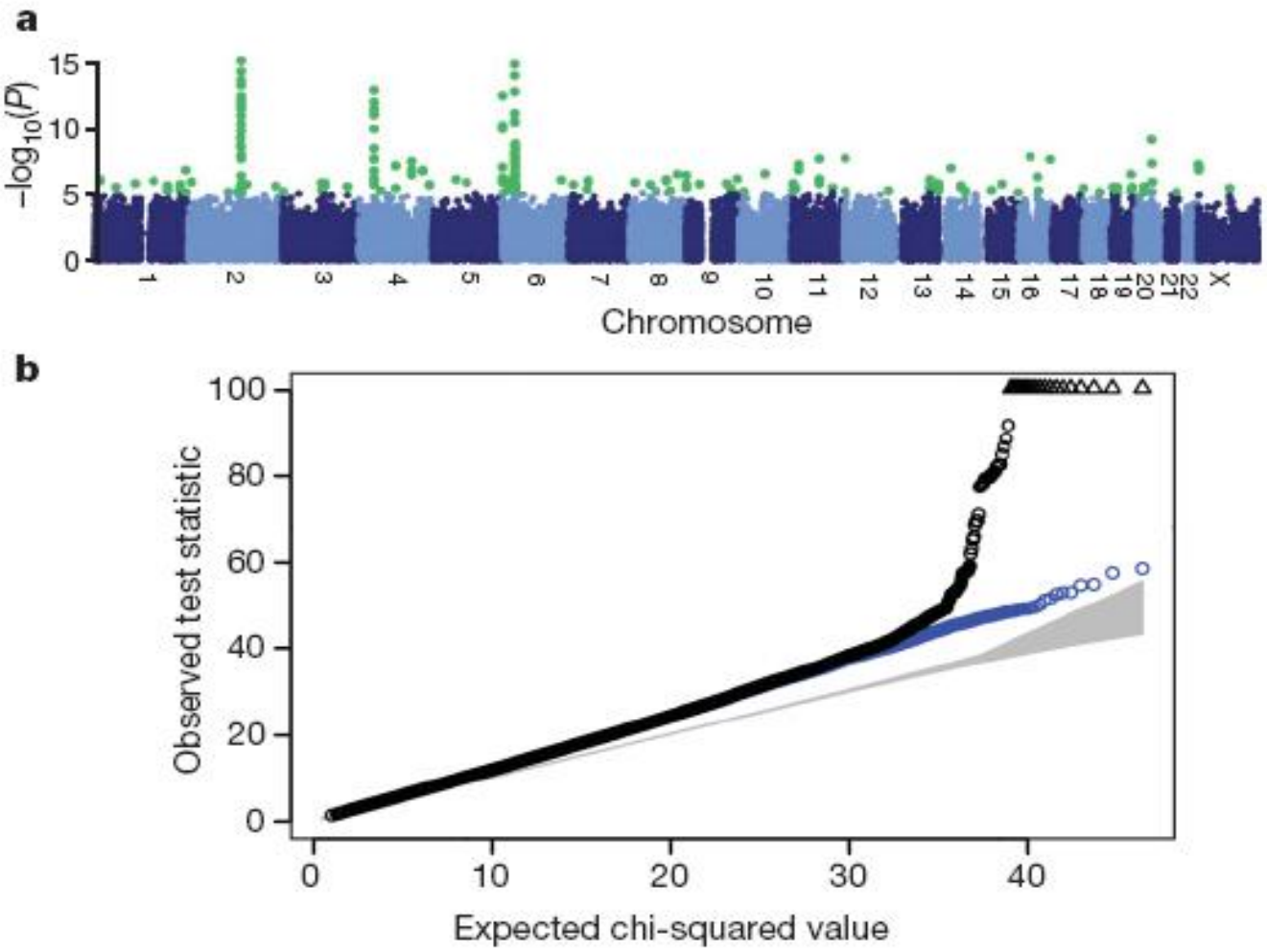


Step 3: Enrichment/pathway analysis

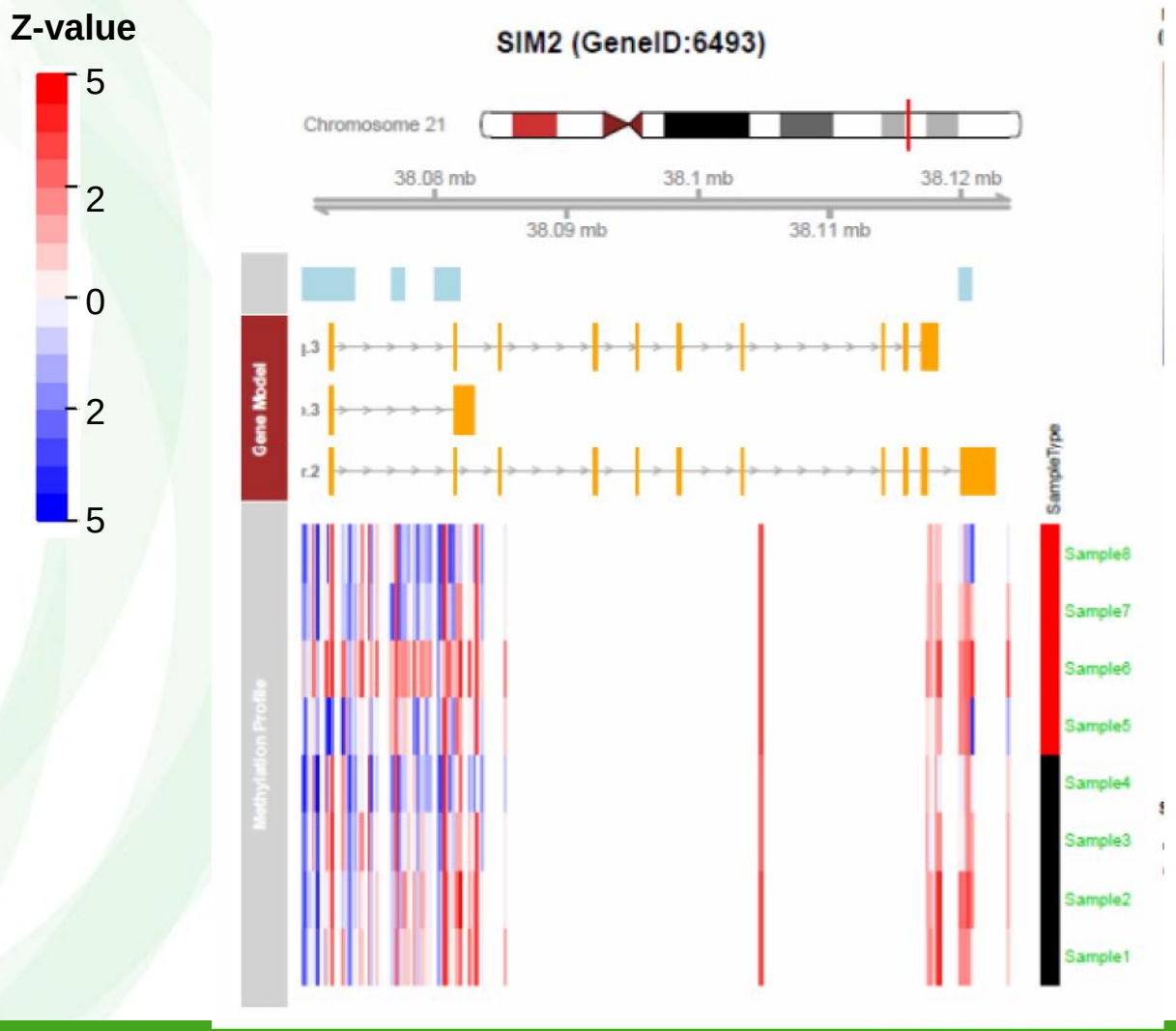
GO, KEGG, "Exposure-related", …

Count	Size	Pvalue	OddsRatio	Term
5	8	0.00	13.72	skeletal muscle tissue development
6	15	0.00	5.50	regulation of lymphocyte differentiation
5	13	0.01	5.13	striated muscle cell differentiation
5	13	0.01	5.13	ribonucleoprotein complex subunit organization
13	35	0.00	5.04	regulation of Rho protein signal transduction
7	20	0.00	4.45	muscle tissue development
6	18	0.01	4.11	positive regulation of Rho GTPase activity
11	35	0.00	3.86	Ras protein signal transduction
16	53	0.00	3.72	regulation of small GTPase mediated signal transduction
6	20	0.02	3.52	muscle organ development
5	17	0.03	3.41	lymphocyte activation involved in immune response
8	28	0.01	3.31	positive regulation of GTPase activity

Step 4: Annotation and visualization



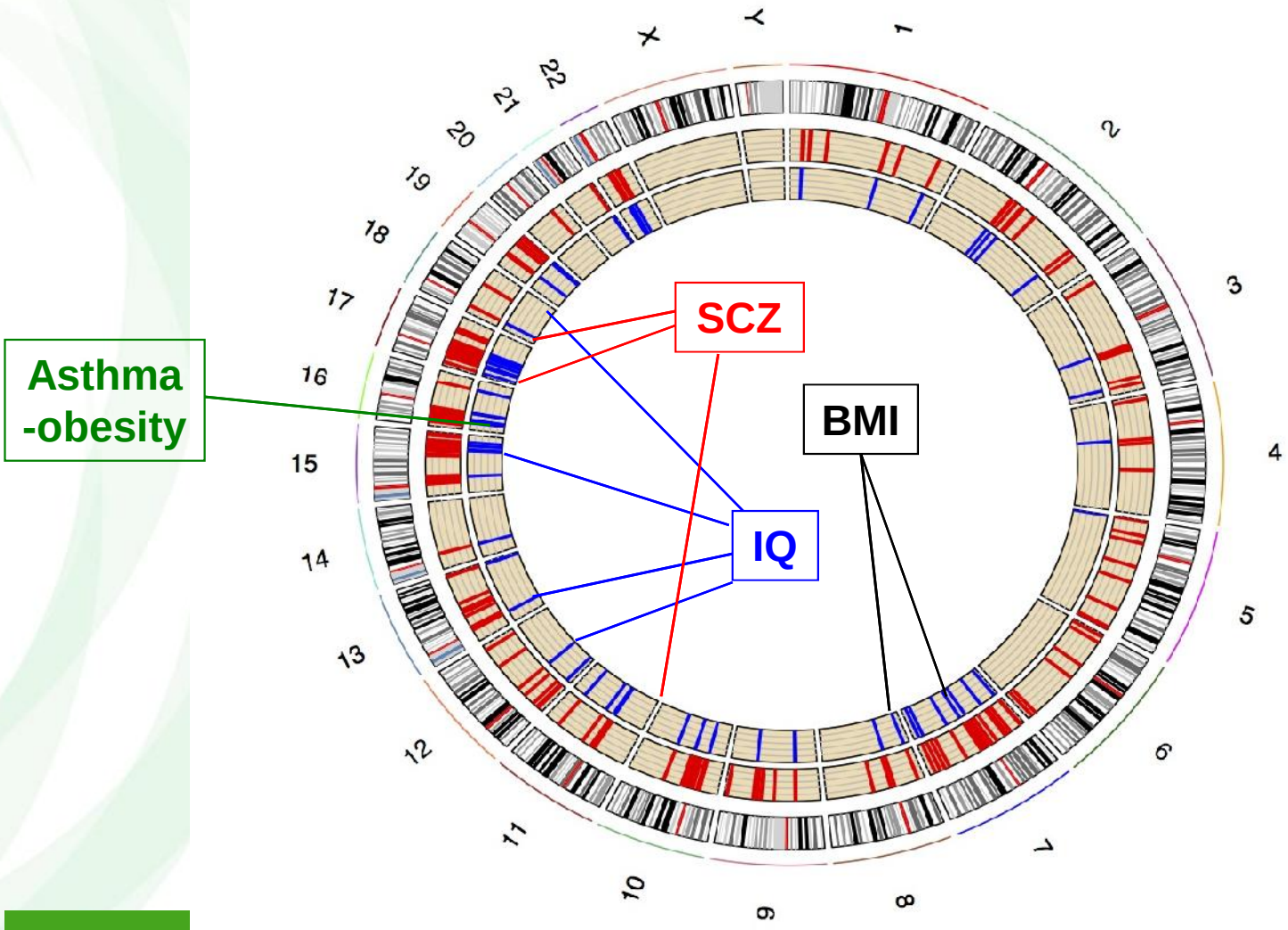
Step 4: Annotation and visualization



Controls

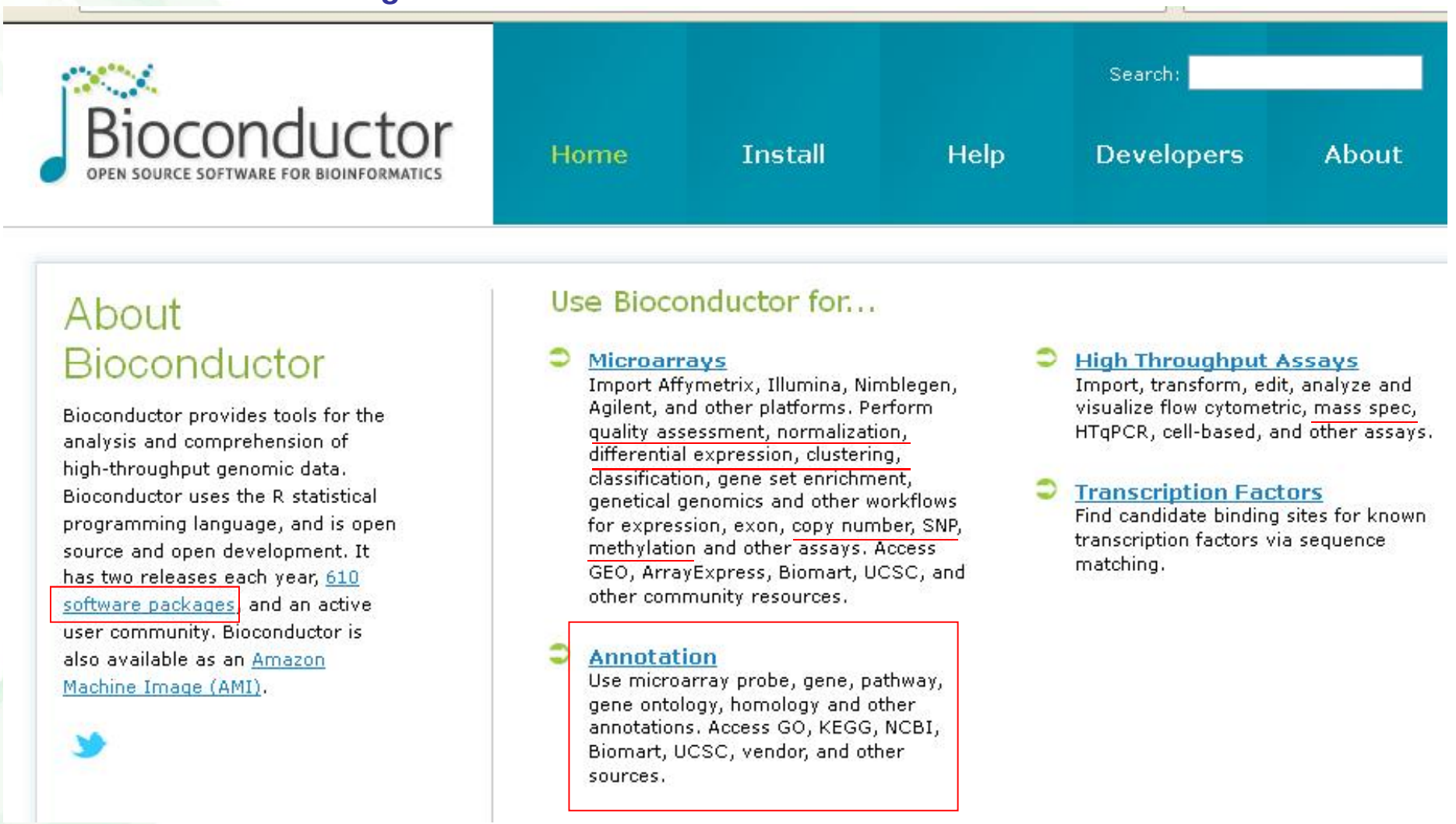
Cases

Step 4: Annotation and visualization



Software

www.bioconductor.org



Software

GENOMet Network (IP Juan R González, MTM2010-09526-E)

www.creal.cat/jrgonzalez/software.htm

SOFTWARE DEVELOPMENT - JUAN R GONZALEZ

We have developed some packages included in the R project in collaboration with other researches from different institutions. Some of these libraries are related to genetics and other ones to survival analysis with recurrent events.

Genetics

Package tweeDEseq

tweeDEseq is an R package for analyzing RNAseq count data. It implements Poisson-Tweedie family of distributions to model count data distribution. This family includes Poisson and Negative Bionomial as particular cases. The testPT test is used to detect genes that are differentially expressed (DE).

The mehots are described in the manuscript

> Esnaola M, Puig P, Gonzalez D, Castelo R, Gonzalez JR. Gene-specific count data distributions are required in RNA-seq experiments with extensive replication. Submitted

The manuscript illustrates the performance of our proposed method using a real RNA-seq data set comprising 69 Nigerian. We have created an experimental data pacakge (tweeDEseqCountData) that is available at Bioconductor (http://www.bioconductor.org/).

Available R packa

Genetics

- tweeDEseq
- BayNet
- inveRsion
- MAD
- bayesGen
- CNVassoc
- R-GADA

Bioinformatics & Data Analysis

Software and statistical methods – SNP arrays

- SNPassoc (CRAN) – paper 190 cites [Bioinformatics] [SNPs]
- CNVassoc (CRAN) – 4th most viewed paper [BMC Genomics] [CNVs]
- R-GADA (R-forge) – Highly accessed [BMC Bioinformatics] [CNVs]
- inveRsion (Bioconductor) – Highly accessed [BMC Bioinformatics] [Inversions]
- invClust[Bioinformatics] [Inversions]
- BayesGen (CRAN) [Statistics in Medicine] [CNVs & SNPs]
- MLPAstats (CRAN) [BMC Bioinformatics] [CNVs]
- MAD (R-forge) [BMC Bioinformatics] [Mosaicisms]
  - Used to analyzed ~58,000 genomes [Nat Gen, 2012]

Software and statistical methods – Sequencing

- tweeDEseq (Bioconductor) [BMC Bioinformatics] [RNAseq]
- GRIAL [Bioinformatics] [Inversions]
- MADseq (Bioconductor) [In progress] [Mosaicisms]
- RASP (Bioconductor) [In progress] [Alternative Splicing]

Limitations

- Data storage
- Computing time

Infrastructure @ CREAL

3 workstations (one more by the end of the next month)

WK1: CPU with 2GHz and 32G of RAM memory (8 cores)

WK2: CPU with 2GHz and 32G of RAM memory (8 cores)

WK3: CPU with 2GHz and 64G of RAM memory (24 cores)

Disk space

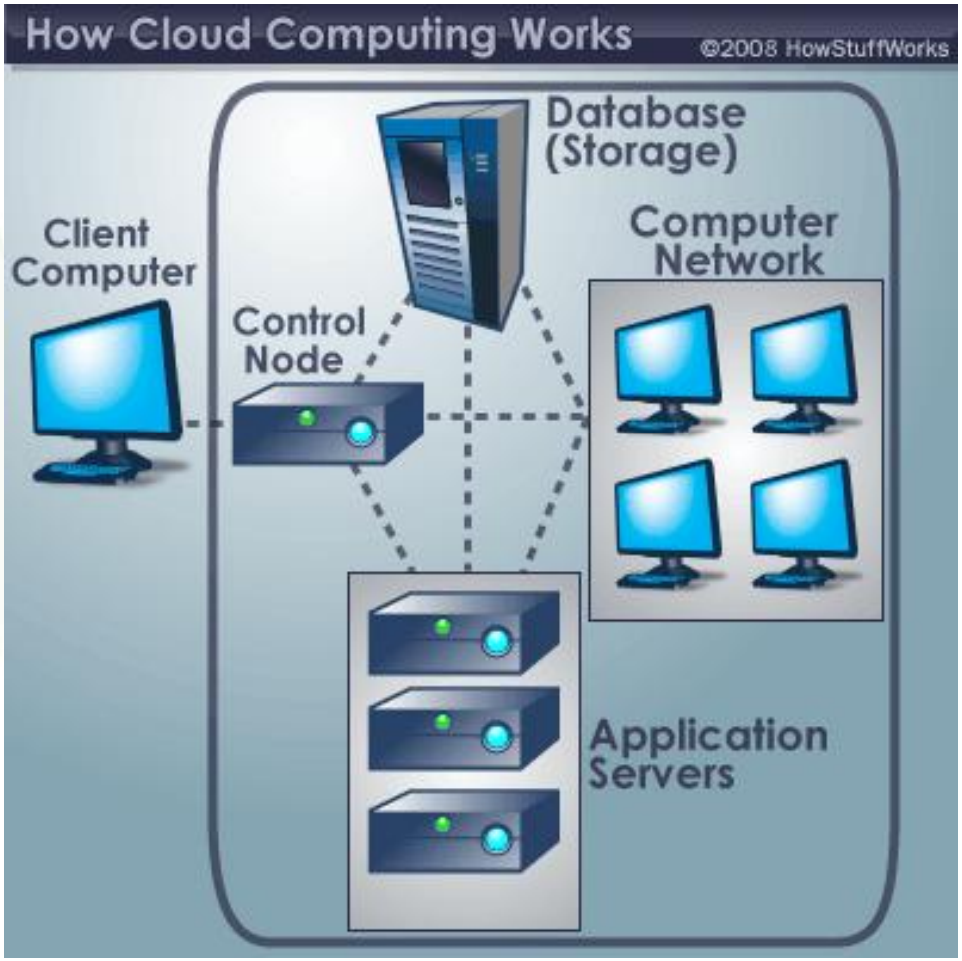
1Tb (WK1) + 2Tb (WK1) + 6Tb (WK1)

+ External device 14Tb

Limitations

Cloud computing ???

Challenges


How Cloud Computing Works
©2008 HowStuffWorks
Database (Storage)
Client Computer
Control Node
Computer Network
Application Servers

Challenges


Amazon Elastic Compute Cloud (Amazon EC2) - Mozilla Firefox
Herramientas Ayuda
Documentation Forums
amazon web services
Inscribase
Cuenta / Consola
Español
Productos y soluciones
AWS Product Information
Desarrolladores
Soporte
Detalles de Amazon EC2
Visión general de EC2
Preguntas más frecuentes sobre EC2
Fijación de precios de EC2
SLA de Amazon EC2
Tipos de instancias de EC2
Opciones de compra de instancias de EC2
Amazon Elastic Compute Cloud (Amazon EC2)
Amazon Elastic Compute Cloud (Amazon EC2) es un servicio web que proporciona capacidad informática con tamaño modificable en la nube. Está diseñado para facilitar a los desarrolladores recursos informáticos escalables y basados en web.
La sencilla interfaz de servicios web de Amazon EC2 permite obtener y configurar su capacidad con una fricción mínima. Proporciona un control completo sobre sus recursos informáticos y permite ejecutarse en el entorno informático acreditado de Amazon. Amazon EC2 reduce el tiempo necesario para obtener y arrancar nuevas instancias de servidor en minutos, lo que permite escalar rápidamente la capacidad, ya sea aumentándola o reduciéndola, según cambien sus necesidades. Amazon EC2 cambia el modelo económico de la informática, al permitir pagar sólo por la
Es fácil contratarla,
pague solo por el consumo realizado
Registrarse ahora

Bioinformatics & Data Analysis

Human Resources

Luis Pérez-Jurado
Armand Gutiérrez

Tonu Esko
Eva Reinmaa
Lili Milani
Andreas Mestpalu

Benjamín Rodríguez-Santiago

Marta Puig
Mario Cáceres