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)
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

EXPOSEMEn
dynamic
diet, metals, air pollution, stress…

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

TRANSCRIPTOME
gene expression (RNA)
tissue and time specific

PROTEOME
tissue and time specific

METABOLOME
tissue and time specific

DISEASOME (PHENOTYPE)

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

www.creal.cat
OMICS

Why OMICS?

To identify biomarkers
To understand molecular mechanisms

exposure
disease
Metabolomic data [serum/plasma]
- Mass-spectrometry 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
SNP array data (GWAS)

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

Can be predicted using existing GWAS data

Rahim, et al., Genome Biology 2008
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: 0, 1, 2, …., 2456, …., 34567, …
Generalized linear models
(Negative Binomial)

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)

[Diagram showing altered genes and functional set]
Step 3: Enrichment/pathway analysis

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

GO, KEGG, “Exposure-related”, ...
Step 4: Annotation and visualization
Step 4: Annotation and visualization

Z-value

SIM2 (GeneID: 6493)

Chromosome 21

38.06 mb 38.1 mb 38.12 mb

Gene Model

Methylation Profile

Sample Type

Sample8
Sample7
Sample6
Sample5
Sample4
Sample3
Sample2
Sample1

Controls

Cases
Step 4: Annotation and visualization

Asthma-obesity

SCZ

BMI

IQ

www.creal.cat
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, 610 software packages, and an active user community. Bioconductor is also available as an Amazon Machine Image (AMI).

Use Bioconductor for...

- **Microarrays**
  Import Affymetrix, Illumina, Nimblegen, Agilent, and other platforms. Perform quality assessment, normalization, differential expression, clustering, classification, gene set enrichment, genetical genomics, and other workflows for expression, exon, copy number, SNP, methylation, and other assays. Access GEO, ArrayExpress, Biomart, UCSC, and other community resources.

- **Annotation**
  Use microarray probe, gene, pathway, gene ontology, homology, and other annotations. Access GO, KEGG, NCBI, Biomart, UCSC, vendor, and other sources.

- **High Throughput Assays**
  Import, transform, edit, analyze, and visualize flow cytometric, mass spec, HTqPCR, cell-based, and other assays.

- **Transcription Factors**
  Find candidate binding sites for known transcription factors via sequence matching.
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 Binomial as particular cases. The testPT test is used to detect genes that are differentially expressed (DE).

The methods are described in the manuscript:

Esnloa 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 package (tweeDEseqCountData) that is available at Bioconductor (http://www.bioconductor.org/).
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
Bioinformatics & Data Analysis

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 ???
Amazon Elastic Compute Cloud (Amazon EC2)

Amazon Elastic Compute Cloud (Amazon EC2) es un servicio web que proporciona capacidad informática con tamaño modulable 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 armar 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 solo por la
Bioinformatics & Data Analysis

Human Resources

Luis Pérez-Jurado
Armand Gutiérrez

Benjamin Rodriguez-Santiago

Marta Puig
Mario Cáceres

Tonu Esko
Eva Reinmaa
Lili Milani
Andreas Mestpalu

www.creal.cat