Introduction
Who we are

The Computational Genomics Department at the Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain, and...

...the INB, National Institute of Bioinformatics (Functional Genomics Node) and the BiER (CIBERER Network of Centers for Rare Diseases)

http://bioinfo.cipf.es
http://www.babelomics.org

@xdopazo @bioinfocipf
Source Pubmed. Query: "high-throughput sequencing"[Title/Abstract] OR "next generation sequencing"[Title/Abstract] OR "rna seq"[Title/Abstract]) AND year[Publication Date]
Tools used in the course

579 papers cite Babelomics (plus 835 FatiGO cites)

(source ISI Web of Knowledge, February 2016)

More than 150,000 experiments analysed during the last year.
More than 1000 experiments per day.
The Babelomics suite of programs becomes a classic. Now is cited as a method in the last edition of Molecular Cloning. The protocol 4 of chapter 8, Expression Profiling by Microarray and RNA-seq, contains a description on how to use Babelomics to analyze expression data.

High impact developments
Software development

Babelomics is the third most cited tool for functional analysis. Includes more than 30 tools for advanced, systems-biology based data analysis.

More than 150,000 experiments were analyzed in our tools during the last year.
The road of excess leads to the palace of wisdom

(William Blake, 28 November 1757 – 12 August 1827, poet, painter, and printmaker)

The introduction and popularisation of high-throughput techniques has drastically changed the way in which biological problems can be addressed and hypotheses can be tested.

But not necessarily the way in which we really address or test them...

Here you will learn how to do so using state-of-the-art methods and software.
Two main applications:

- Diagnostic
- Disease gene finding
Disease gene finding

Typically, an exome renders between 40 and 60K variants (and a genome about 1 million). Only one or a few among all of them are expected to be the causative factors of the disease.

The prioritization process is like a police investigation in which suspected are discarded by their alibies

Thus, through sequential heuristic filtering steps, unlikely candidates are discarded and a final, reduced list with one or a few candidate genes is (hopefully) produced
# Pipeline of data analysis

## Primary processing
- Initial QC
- FASTQ file
- Mapping
- BAM file
- Variant calling
- VCF File

## Secondary analysis (Heuristic filtering)
- Variant annotation
- Filtering by effect
- Filtering by MAF
- Filtering by family segregation

## Knowledge-based prioritization
- Proximity to other known disease genes
- Functional proximity
- Network proximity
- Burden tests
- Other prioritization methods

## Primary analysis

## Gene prioritization
**Pipeline of data analysis**

<table>
<thead>
<tr>
<th>Primary processing</th>
<th>Secondary analysis (Heuristic filtering)</th>
<th>Knowledge-based prioritization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial QC FASTQ file</td>
<td>Variant annotation</td>
<td>Proximity to other known disease genes</td>
</tr>
<tr>
<td>Mapping BAM file</td>
<td>Filtering by effect</td>
<td>Functional proximity</td>
</tr>
<tr>
<td>Variant calling VCF File</td>
<td>Filtering by MAF</td>
<td>Network proximity</td>
</tr>
<tr>
<td></td>
<td>Filtering by family segregation</td>
<td>Burden tests</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other prioritization methods</td>
</tr>
</tbody>
</table>

**Primary analysis** → **Gene prioritization**
Fastq file, up to hundreds of GB per run

**HPG suite**

*High-Performance Genomics*

- QC and preprocessing
- **HPG Aligner**, short read aligner
- SAM/BAM file
- QC and preprocessing
- **Variant calling analysis**
- VCF file
- QC and preprocessing
- **Variant VCF viewer**
  - HTML5+SVG Web based viewer
- **HPG Variant, Variant analysis**
  - Consequence type, GWAS, regulatory variants and system biology information

- QC stats, filtering and preprocessing options
- Double mapping strategy: Burrows-Wheeler Transform (*GPU Nvidia CUDA*) + Smith-Waterman (*CPU OpenMP+SSE/AVX*)
- GATK and SAM mPileup HPC Implementation.
- Statistics genomic tests
- Consequence type, GWAS, regulatory variants and system biology information

More info at:

http://bioinfo.cipf.es/docs/compbio/projects/hpg/doku.php

Other analysis (HPC4Genomics consortium)

- RNA-seq
- DNA assembly
- Methyl-seq
- Copy Number
- Structural variation
- Transcript isoform
- Etc.
### Pipeline of data analysis

#### Primary processing
- Initial QC
  - FASTQ file
- Mapping
  - BAM file
- Variant calling
  - VCF File

#### Secondary analysis (Heuristic filtering)
- Variant annotation
- Filtering by effect
- Filtering by MAF
- Filtering by family segregation

#### Knowledge–based prioritization
- Proximity to other known disease genes
- Functional proximity
- Network proximity
- Burden tests
- Other prioritization methods

### Gene prioritization
3-Methylglutaconic aciduria (3-MGA-uria) is a heterogeneous group of syndromes characterized by an increased excretion of 3-methylglutaconic and 3-methylglutaric acids. WES with a consecutive filter approach is enough to detect the new mutation in this case.
The BiERapp

An interactive web tool that implements different heuristic filters for disease variant/gene prioritization
Filters include family pedigree segregation, population frequencies, pathogenic indexes, etc.
In less than two years, the CIBERER initiative rendered 36 new mutations in known disease genes and 27 new mutations in 13 new disease genes.
## Pipeline of data analysis

### Primary processing
- Initial QC
  - FASTQ file
- Mapping
  - BAM file
- Variant calling
  - VCF File

### Secondary analysis (Heuristic filtering)
- Variant annotation
- Filtering by effect
- Filtering by MAF
- Filtering by family segregation

### Knowledge-based prioritization
- Proximity to other known disease genes
- Functional proximity
- Network proximity
- Burden tests
- Other prioritization methods

### Gene prioritization
Knowledge-based prioritization

Network analysis

Four new loci associations discovered by pathway-based and network analyses of the genome-wide variability profile of Hirschsprung’s disease

SNPs validated in independent cohorts

CHRNA7 (rs2175886 p = 0.000607)
IQGAP2 (rs950643 p = 0.0003585)
DLC1 (rs1454947 p = 0.007526)
NGS for diagnostic

A web tool for the design and management of panels of genes for targeted enrichment and massive sequencing for clinical applications

If no diagnostic variants appear, then secondary findings are studied

New filter based on local population variant frequencies

http://team.babelomics.org
Uses of BiERapp and TEAM

Implementation of tools in the IT4I Supercomputing Center (Czech Republic)

The pipelines of primary and secondary analysis developed by the Computational Genomics Department of the CIPF in close collaboration with the Bull Chair of computational genomics has proven its efficiency in the analysis of more than 1000 exomes in a joint collaborative project of the CIBERER and the MGP. A first pilot implementation has been done in the IT4I supercomputing center, which aims to centralize the analysis of genomics data in the country.
# Course Schedule

**GDA2016: 29 Feb - 4 Mar**

<table>
<thead>
<tr>
<th>Monday 29</th>
<th>Tuesday 1</th>
<th>Wednesday 2</th>
<th>Thursday 3</th>
<th>Friday 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00-9:30</td>
<td>Registration</td>
<td>Biological and Clinical Databases, CSVS</td>
<td>Panel of genes: design and analysis for clinical applications, TEAM</td>
<td>Differential Expression Analysis, Babelomics 5</td>
</tr>
<tr>
<td>9:30-10:00</td>
<td>Course presentation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:00-10:30</td>
<td>Introduction to NGS Technologies for Genomic Analysis</td>
<td>Differential Expression Analysis</td>
<td>Differential Expression Analysis, Babelomics 5</td>
<td>Pathways Analysis</td>
</tr>
<tr>
<td>10:30-11:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:00-11:30</td>
<td>Coffee break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:30-12:00</td>
<td>Introduction to Linux</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00-12:30</td>
<td>Prioritization of variants and genes: BIErapp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:30-13:00</td>
<td>Primary Analysis: Quality control for raw data</td>
<td>Differential Expression Analysis</td>
<td>Differential Expression Analysis, Babelomics 5</td>
<td>Pathways Analysis</td>
</tr>
<tr>
<td>13:00-13:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:30-14:00</td>
<td>Lunch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00-14:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:30-15:00</td>
<td>Primary Analysis: Mapping NGS Reads and visualization for Genomics Studies</td>
<td>Prioritization of variants and genes: BIErapp and Network tools</td>
<td>Course presentation</td>
<td>Functional Profiling from Babelomics 5, FatGO and Gene Set Analysis</td>
</tr>
<tr>
<td>15:00-15:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:30-16:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00-16:30</td>
<td>Coffee break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00-17:00</td>
<td>Primary Analysis: Variant Calling SNPs and INDELS</td>
<td>Bring your own data!</td>
<td>Primary Analysis for RNA-Seq data: quality control, mapping and quantification</td>
<td>Functional Profiling from Babelomics 5, Network Analysis</td>
</tr>
<tr>
<td>17:00-17:30</td>
<td>Variant Annotation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:30-18:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21:00</td>
<td>SOCIAL DINNER</td>
<td>SOCIAL DINNER</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**GENOMICS: 29 February, 1 and 2 March**

**TRANSCRIPTOMICS: 2, 3 and 4 March**
Social

GDA group in Linked-in
https://www.linkedin.com/groups/1934338

Babelomics group in twitter
@xdopazo
@bioinfocipf
@babelomics

http://bioinfo.cipf.es
http://www.babelomics.org

And the social dinner (and mascletás)... Networking is very important for your career. Keep in touch with fellows and instructors