Biological and Clinical Databases
CellBase and CIBERER Spanish Variant Server (CSVs)

Alicia Amadoz
March 1st, 2016
Outline

● Introduction
  ○ Genomic variation
  ○ Analysis pipeline
  ○ Databases and resources

● Biological Databases
  ○ dbSNP, COSMIC and HumsaVar

● Clinical Databases
  ○ ClinVar, OMIM and HGMD

● CellBase

● Catalogs of Human Genetic Variation
  ○ HapMap, 1000G, ESP, ExAC

● CIBERERER Spanish Variant Server (CSVs)

● Exercises
Introduction - Genomic variation

DNA → RNA → PROTEIN

transcription

reverse

translation

reverse

(Koonin 2012 Biology Direct)
Introduction - Genomic variation

Which genetic differences are associated to a given phenotype or disease?
Main types of genetic variation

- **Single nucleotide variant**
- **Small insertion**
- **Small deletion**

**SNP >1%**
Main types of genetic variation

- **SNP >1%**
- **CNVs >50bp**

**Sequence variants**
- Single nucleotide variant
- Small insertion
- Small deletion

**Structural variants**
- Deletion
- Duplication
- Inversion
- Translocation

**Chromosome 1**

**Chromosome 2**

Reference sequence: ATCGGGTCA...TCA

SNP example: ATCGGGTCA...ATCA

CNV example: ATCGGGTCA...GACGTCA
Introduction - Analysis pipeline

Primary Analysis

Sequence preprocessing

Mapping

Variant calling
Introduction - Analysis pipeline

Primary Analysis

1. Sequence preprocessing
2. Mapping
3. Variant calling

Which variants are related to a given phenotype or disease?

Secondary Analysis

4. Variant annotation
5. Variant prioritization

Methods?
Publicly available tools and databases?
Linkage disequilibrium (LD) analysis
Genome-wide association studies (GWAS)

LD is a non-random association between alleles at different sites.

Study a limited set of SNPs

In multi-factorial diseases is difficult to find causal variants.
Introduction - Analysis pipeline

Linkage disequilibrium (LD) analysis
Genome-wide association studies (GWAS)

Lack of well defined case-control groups
Insufficient sample size
Population stratification
Control for multiple testing

False positive rate
Assessment of variant function

Computational methods can estimate the functional effect of variants by using **comparative genomics** and knowledge of **protein biochemistry and structure**.

(Cooper et al. 2011 Nature Reviews Genetics)

Some tools to estimate functional impact

**Minor allele frequency (MAF)**

frequency of the less common allele of a SNP in a population

**Consequence type**

effect of the variant at transcriptional level

**SIFT**

amino acid substitution affects protein function (1=tolerated; 0=not tolerated)

http://sift.jcvi.org

**PolyPhen**

impact on the structure and function of a protein (1=probably damage; 0=benign)

http://genetics.bwh.harvard.edu/pph2

http://www.ensembl.org/info/genome/variation/predicted_data.html
Introduction - Databases and resources

- NAR Biological Database Collection
  
  http://www.oxfordjournals.org/our_journals/nar/database/c

  1685 databases

  ![Nucleic Acids Research]

  **Contents**

  Volume 44, Database issue, January 4, 2016

  NUCLEIC ACID SEQUENCE, STRUCTURE, AND REGULATION

  The 2016 database issue of *Nucleic Acids Research* and an updated molecular biology database collection

  D.I. Rigden, X.M. Fernández-Suárez

  M.Y. Galperin

  D1–D6

Major bioinformatics centers

- [NCBI](http://www.ncbi.nlm.nih.gov)
- [EMBL-EBI](http://www.ebi.ac.uk)
- [Swiss Institute of Bioinformatics](http://www.sib.ch)
Data repositories

- ArrayExpress (AE) http://www.ebi.ac.uk/arrayexpress
- The Sequence Read Archive (SRA) http://www.ncbi.nlm.nih.gov/sra
- European Nucleotide Archive (ENA) http://www.ebi.ac.uk/ena
Catalogs of human genetic variation

- The International HapMap project
- The 1000 genome project
  [http://www.1000genomes.org](http://www.1000genomes.org)
- Exome Sequencing Project (ESP)
  [http://evs.gs.washington.edu/EVS](http://evs.gs.washington.edu/EVS)
- Exome Aggregation Consortium (ExAC)
- The CIBERER Spanish Variant Server (CSVs)
  [http://csvs.babelomics.org](http://csvs.babelomics.org)
## dbSNP


### Chromosome location
### Functional consequence
### Global MAF
### Validation status

<table>
<thead>
<tr>
<th>Organism</th>
<th>dbSNP Build</th>
<th>Genome Build</th>
<th>Number of Submissions (ss#s)</th>
<th>Number of RefSNP Clusters (rs#s) ( # validated)</th>
<th>Number of (rs#s) in gene</th>
<th>Number of (ss#s) with genotype</th>
<th>Number of (ss#s) with frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>146</td>
<td>38.2</td>
<td>538,341,120</td>
<td>150,482,731 (100,135,281)</td>
<td>67,339,846</td>
<td>73,617,935</td>
<td>130,173,026</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>146</td>
<td>7.1</td>
<td>293,824,705</td>
<td>99,547,043 (47,585,089)</td>
<td>49,248,999</td>
<td>10,202</td>
<td>968</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>146</td>
<td>38.4</td>
<td>135,736,870</td>
<td>80,443,437 (16,396,141)</td>
<td>37,738,633</td>
<td>24,843,897</td>
<td>77</td>
</tr>
<tr>
<td>Zea mays</td>
<td>146</td>
<td>1.1</td>
<td>71,409,972</td>
<td>54,251,824 (9,388,953)</td>
<td>12,798,179</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>146</td>
<td>2.3</td>
<td>17,902,170</td>
<td>9,505,665 (0)</td>
<td>3,103,829</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>146</td>
<td>3.3</td>
<td>7,436,723</td>
<td>5,332,639 (1,041,856)</td>
<td>2,759,092</td>
<td>2,551,456</td>
<td>43</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>146</td>
<td>6.1</td>
<td>6,473,263</td>
<td>5,075,460 (5,042,280)</td>
<td>1,645,109</td>
<td>6,425,787</td>
<td></td>
</tr>
<tr>
<td>Salmo salar</td>
<td>146</td>
<td>1.1</td>
<td>1,351,351</td>
<td>1,038,504 (60,176)</td>
<td>691,675</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Brassica napus</td>
<td>146</td>
<td>1.1</td>
<td>914,257</td>
<td>901,510 (0)</td>
<td>156,347</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cicer arietinum</td>
<td>146</td>
<td>1.2</td>
<td>873,342</td>
<td>519,066 (0)</td>
<td>134,145</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>146</td>
<td>1.1</td>
<td>29</td>
<td>29 (0)</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total: 11 Organisms</td>
<td></td>
<td></td>
<td>1,074,263,802</td>
<td>407,698,208 (179,649,776)</td>
<td>195,615,871</td>
<td>107,749,367</td>
<td>130,174,124</td>
</tr>
</tbody>
</table>

November 2015
COSMIC
http://cancer.sanger.ac.uk/cosmic

Mutation type
Tissue distribution
Somatic status
Pathways affected
Curated publications
HumsaVar
http://www.uniprot.org/docs/humsavar

Amino acid change
Type of variant
Disease name

Description: Human polymorphisms and disease mutations: index
Name: humsavar.txt
Release: 2016_02 of 17-Feb-2016

Important note: variants classification is intended for research purposes only, not for clinical and diagnostic use. The label disease variant is assigned according to literature reports on probable disease-association that can be based on theoretical reasons. Therefore this label must not be considered as a definitive proof for the pathogenic role of a variant.

Statistics for single amino acid variants:

Disease variants: 27606
Polymorphisms: 36285
Unclassified variants: 7375
Total: 73266
ClinVar


a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence.

Clinical significance

Review status (gold stars)
OMIM
http://www.omim.org

Known mendelian disorders
3,519 genes with phenotype-causing mutation

Phenotype
Inheritance

OMIM Entry Statistics

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene description</td>
<td>15,199</td>
</tr>
<tr>
<td>Gene and phenotype, combined</td>
<td>85</td>
</tr>
<tr>
<td>Phenotype description, molecular basis known</td>
<td>4,664</td>
</tr>
<tr>
<td>Phenotype description or locus, molecular basis unknown</td>
<td>1,634</td>
</tr>
<tr>
<td>Other, mainly phenotypes with suspected mendelian basis</td>
<td>1,808</td>
</tr>
<tr>
<td>Totals</td>
<td>23,390</td>
</tr>
</tbody>
</table>
Clinical databases - HGMD

HGMD
http://www.hgmd.cf.ac.uk

Highly curated repository

Codon change
Amino acid change
Phenotype
References
CellBase

https://github.com/opencb/cellbase/wiki

A scalable and high-performance database that integrates biological information from well-known data sources such as Ensembl and Uniprot among others.

Data sources

Core features: genes, transcripts, exons, proteins (UniProt), etc.

Regulatory: TFBSs, miRNAs, regulatory regions, PWMs, conserved regions, etc.

Functional annotation: OBO ontologies (Gene ontology, disease ontology, etc.), InterPro, etc.

Variation: dbSNP, HapMap, 1000 Genomes project, COSMIC, protein variants, etc.

Systems biology: IntAct, Reactome, gene co-expression, etc.

(Bleda et al. 2012 Nucl. Acids Res.)
CellBase

Java RESTful Web Services API design

**Structure**
ws.bioinfo.cipf.es/cellbase/rest/{version}/{species}/{category}/{subcategory}/id/{resource}?{filters}

**Categories**
- **genomic**

- **feature**
  - Example: ws.bioinfo.cipf.es/cellldb/rest/{version}/{species}/feature/gene/BRCA2,BCL2/transcript
  - Example: ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/id/BRCA2/xref?dbname=go

- **regulatory**
  - Example: ws.bioinfo.cipf.es/cellbase/rest/{version}/{species}/regulatory/{subcategory}/id/{resource}
  - Example: ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/regulatory/tf/USF1/tfbs
  - Example: ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/regulatory/mirna_gene/hsa-mir-149/disease

- **network**
  - Example: ws.bioinfo.cipf.es/cellbase/rest/{version}/{species}/network/{subcategory}/id/{resource}
  - Example: ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/network/pathway/list
  - Example: ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/network/pathway/Triacylglycerol%20biosynthesis/image

Server

Client

TXT or JSON

- **Programmatic access**
  - CLI client has been implemented

- **Web browser access**

- **Usage in web applications**

(Bleda et al. 2012 Nucl. Acids Res.)
## CellBase

### CellBase API

http://bioinfo.hpc.cam.ac.uk/cellbase/webservices

http://HOST/cellbase/rest/{version}/{species}/{category}/{subcategory}/{id}/{resource}

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcategory</th>
<th>Identifier format</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomic</td>
<td>Region</td>
<td>chr:start-end</td>
<td>gene, transcript, snp, sequence, reverse, tfbs, mirna_target, regulatory</td>
</tr>
<tr>
<td></td>
<td>Variant</td>
<td>chr:position:new allele</td>
<td>consequence_type</td>
</tr>
<tr>
<td></td>
<td>Position</td>
<td>chr:position</td>
<td>gene, snp, mutation, functional</td>
</tr>
<tr>
<td>Feature</td>
<td>Gene</td>
<td>All gene ID formats</td>
<td>info, sequence, transcript, tfbs, mirna_target, protein_feature, snp, mutation</td>
</tr>
<tr>
<td></td>
<td>Transcript</td>
<td>Ensembl or RefSeq ID</td>
<td>info, sequence, exon</td>
</tr>
<tr>
<td></td>
<td>Snp</td>
<td>dbSNP or Ensembl ID</td>
<td>info, consequence_type, population_frequency, phenotype, xref</td>
</tr>
<tr>
<td></td>
<td>Exon</td>
<td>Ensembl ID</td>
<td>info, sequence, region, transcript</td>
</tr>
<tr>
<td></td>
<td>Protein</td>
<td>UniProt or Ensembl ID</td>
<td>info, gene, sequence, transcript, feature, xref, variant</td>
</tr>
<tr>
<td></td>
<td>Id</td>
<td>All possible IDs</td>
<td>Xref</td>
</tr>
<tr>
<td>Regulatory</td>
<td>mirna_gene</td>
<td>miRBase gene ID</td>
<td>info, gene, mirna_mature, target, disease</td>
</tr>
<tr>
<td></td>
<td>mirna_mature</td>
<td>miRBase mature ID</td>
<td>info, gene, mirna_gene, target, disease</td>
</tr>
<tr>
<td>Network</td>
<td>Pathway</td>
<td>none</td>
<td>info, tfbs, gene, protein, pwm</td>
</tr>
<tr>
<td></td>
<td>Interactome</td>
<td>Reactome pathway name</td>
<td>info, subpathway, element, gene, protein, image</td>
</tr>
</tbody>
</table>

(Bleda et al. 2012 Nucl. Acids Res.)
Catalogs of Human Genetic Variation

- HapMap
  
  [Link to HapMap website]

Genetic similarities and differences in human beings.

A haplotype is a sequence of alleles that are usually inherited together.

**Combination of different alleles at 20 nearby SNPs**

**Minimum SNP set to identify a haplotype**
Catalogs of Human Genetic Variation

- 1,000 Genomes Project

[http://www.1000genomes.org](http://www.1000genomes.org)
Genetic variants with frequencies of at least 1%
2,504 samples, 26 populations

Polymorphic variants within sampled populations

(The 1000 Genomes Project Consortium 2015 Nature)
● NHLBI Exome Sequencing Project (ESP)

http://evs.gs.washington.edu/EVS

Samples from heart, lung and blood disorders

6,500 exomes
Catalogs of Human Genetic Variation

- **ExAC**
  

  60,706 unrelated individuals

---

**Contributing projects**
- 1000 Genomes
- Bulgarian Trios
- Finland-United States Investigation of NIDDM Genetics (FUSION)
- GoT2D
- Inflammatory Bowel Disease
- METabolic Syndrome In Men (METSIM)
- Jackson Heart Study
- Myocardial Infarction Genetics Consortium:
  - Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
  - Ottawa Genomics Heart Study
  - Pakistan Risk of Myocardial Infarction Study (PROMIS)
  - Precocious Coronary Artery Disease Study (PROCARDIS)
  - Registre Gironi del COR (REGICOR)
- NHLBI-Go Exome Sequencing Project (ESP)
- National Institute of Mental Health (NIMH) Controls
- SIGMA-T2D
- Sequencing in Suomi (SISu)
- Swedish Schizophrenia & Bipolar Studies
- T2D-GENES
- Schizophrenia Trios from Taiwan
- The Cancer Genome Atlas (TCGA)
- Tourette Syndrome Association International Consortium for Genomics (TSAICG)
Genetic variability of Spanish population.

Useful for filtering polymorphisms and local variations.

578 unrelated Spanish individuals.

http://csvs.babelomics.org

Overview
Welcome to the CIBERER Spanish Variant Server. CSVS was created to provide information about the variability of the Spanish population to the scientific/medical community. It is useful for filtering polymorphisms and local variations in the process of prioritizing candidate disease genes. CSVS currently stores information on 578 unrelated Spanish individuals. We accept submissions from WES or WGS. See the protocol for sending samples.

Submission Protocol

Supported by

[Logos of supporting institutions]
Subpopulations

MGP (267)
IBS (107)
18 subgroups by disease type
Thanks for your attention

Any questions?