Máster en Biotecnología Biomédica

Francisco García García
Bioinformatics & Biostatistics Unit. CIPF

WODA

WEB-BASED OMICS DATA ANALYSIS

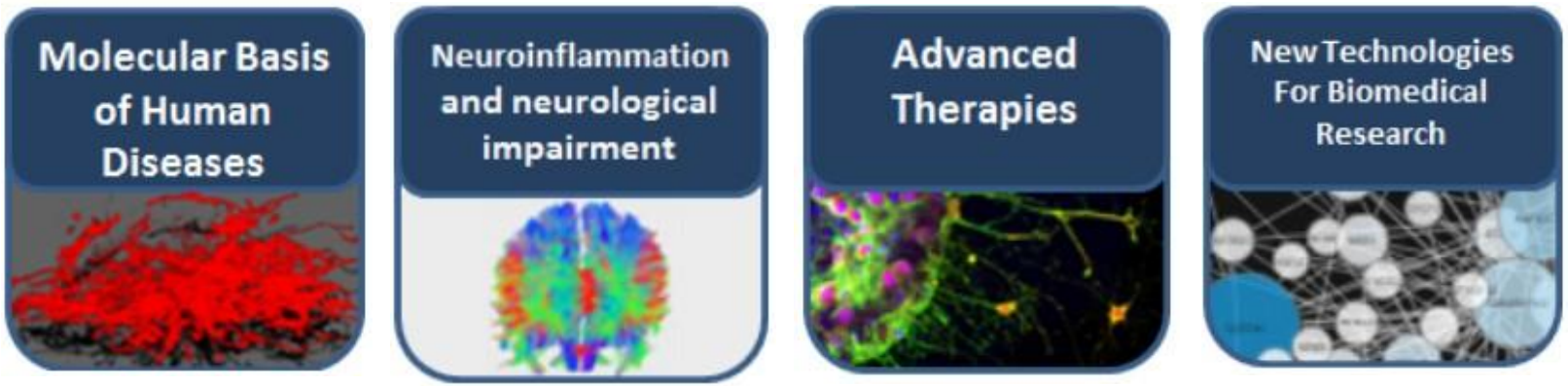
Unidad de
Bioinformática y
Bioestadística

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

UBB

The UBB-CIPF is a technical and scientific unit that aims to promote biomedical research from the interaction with the groups and services of our center.

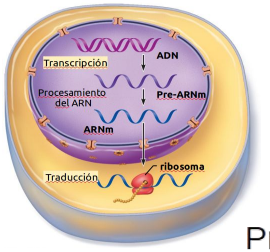
PRINCIPE FELIPE
CENTRO DE INVESTIGACION

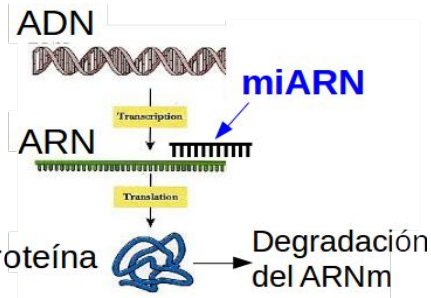


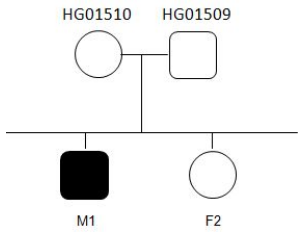
UBB team

http://bioinfo.cipf.es/ubb/

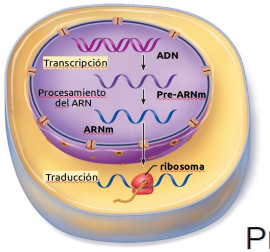
Why this Unit at CIPF?

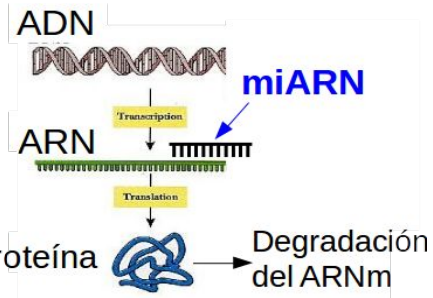


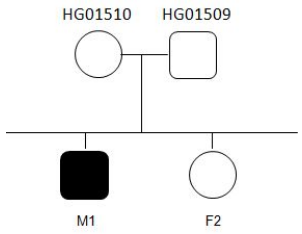


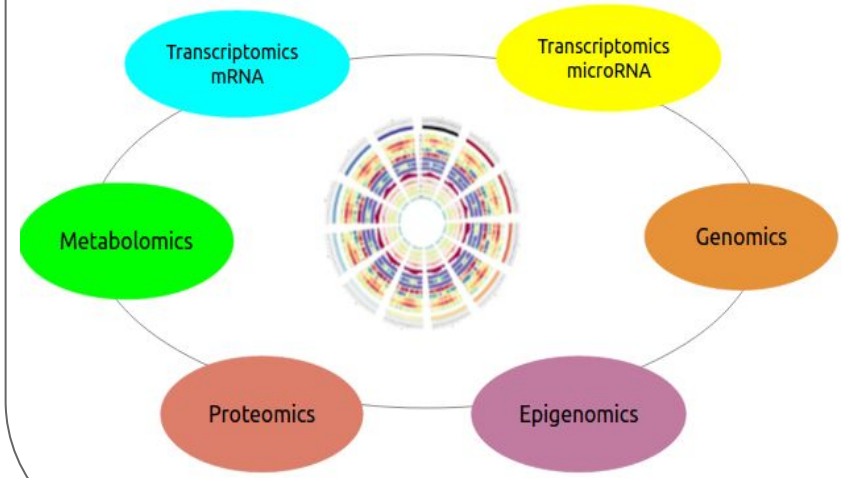


Why this Unit at CIPF?









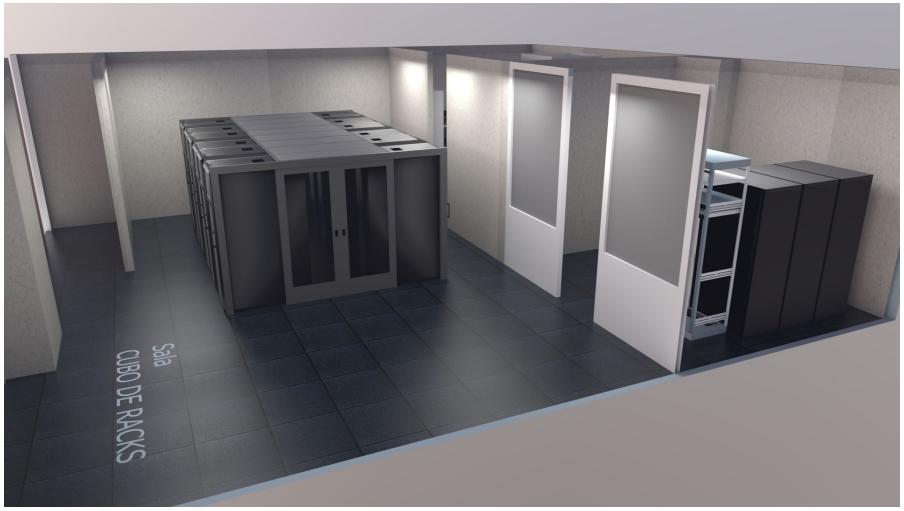
What activities do we do?

Bioinformatics &
Biostatistics Unit

UBB activities

1. Consulting
2. Teaching
3. Cluster coordination
4. Research

UBB activities



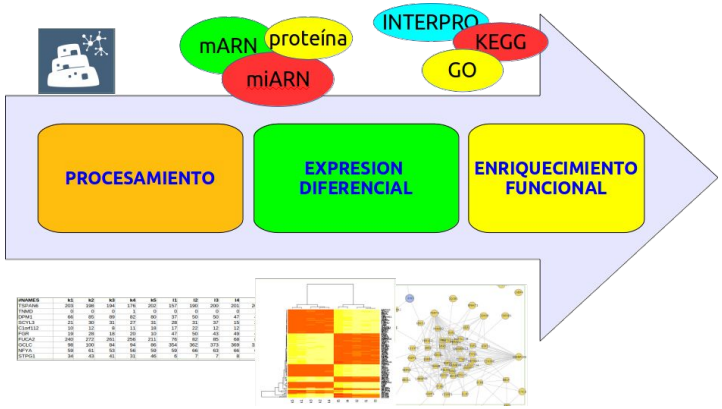
3. Cluster coordination

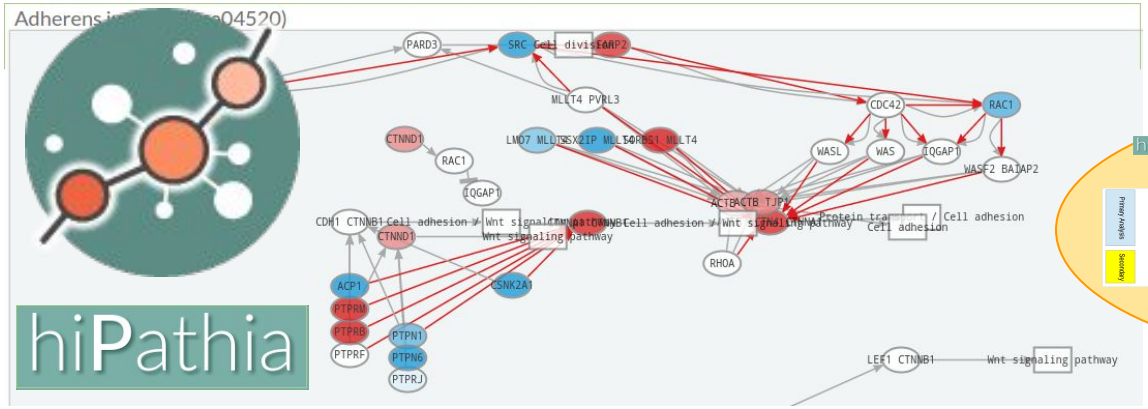
- Computing cluster 44 nodes
- 600 CPU’s
- RAM 11 TB
- 1 PB (130.000 DVDs)

UBB activities

4. Research activities

- Methods for omics data analysis
- Research projects





What is WODA?

A Practical course
Web-based resources
Free tools
Start point: processed or normalized data
Any laptop or pc

Programming skills
Raw data processing
Powerful computational infrastructure

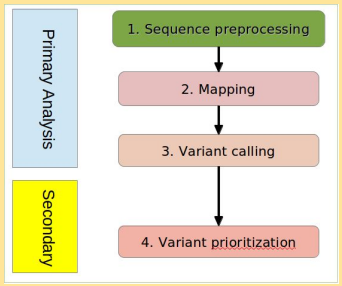
Toolbox

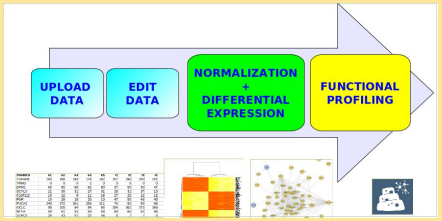
Omics toolbox

hiPathia

Babelomics 5

PathAct





TEAM

CIBERER Copy Number Variation

BierApp

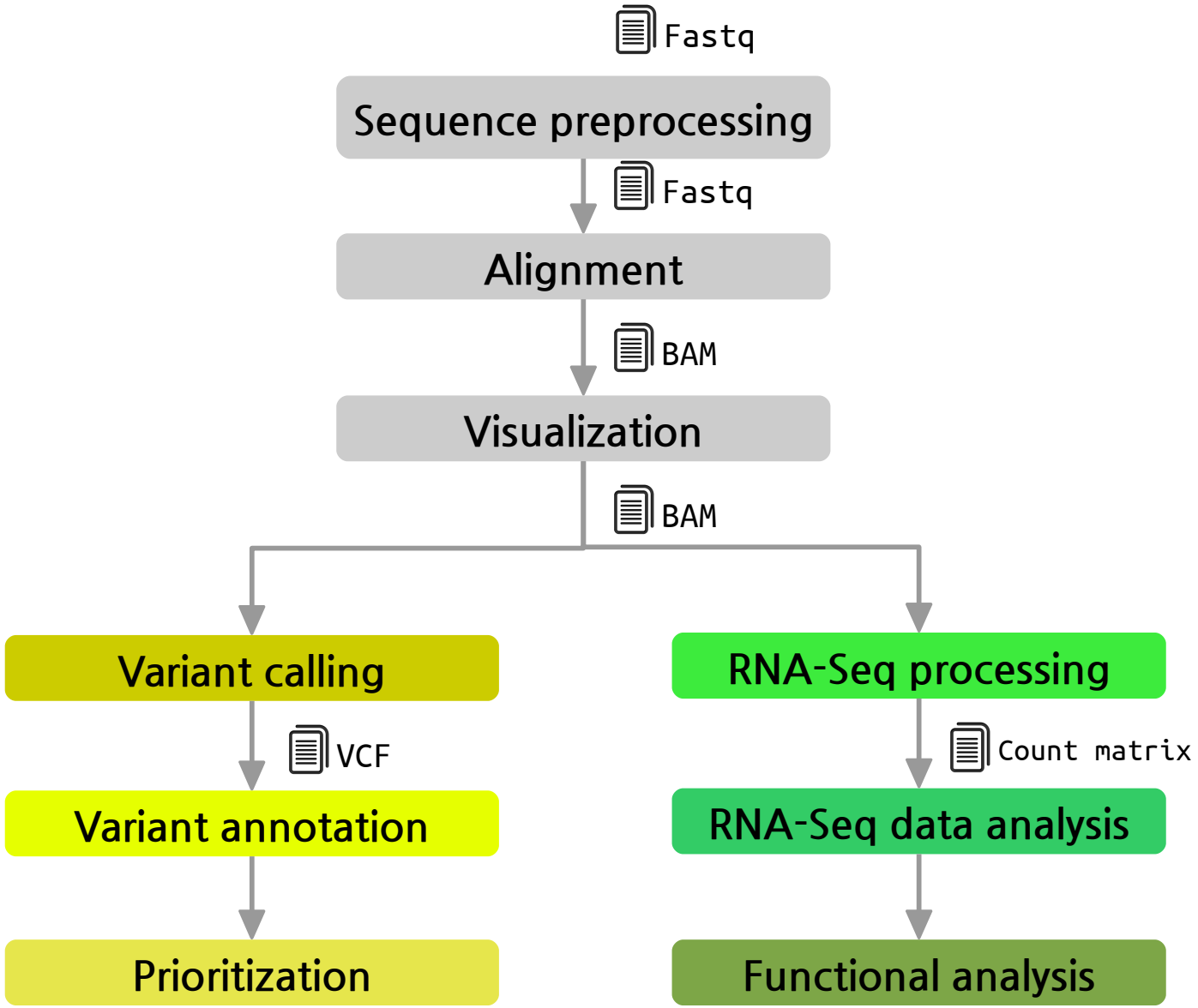
CSVS

Web tools to analyze omic data

fgardos@gmail.com

Máster en Biotecnología Biomédica. UPV

NGS Data Analysis Pipeline



Fastq format

- We could say “it is a fasta with **qualities**”:
  - 1. Header (like the fasta but starting with “@”)
  - 2. Sequence (string of nt)
  - 3. “+” and sequence ID (optional)
  - 4. Encoded quality of the sequence

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

BAM/SAM format

```
@PG     ID:HPG-Aligner  VN:1.0
@SQ     SN:20   LN:63025520

HWI-ST700660_138:2:2105:7292:79900#2@0/1        16      20      76703   254     76=     *       0       0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
GIJGJLGGFLILGGIEIFEKEDELIGLJIHJFIKKFELFIKLFFGLGHKKGJLFIIGKFFEFFEFGKCKFHHCCCF    AS:i:254       NH:i:1  NM:i:0

HWI-ST700660_138:2:2208:6911:12246#2@0/1        16      20      76703   254     76=     *       0       0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
HHJFHLGFFLILEGIKIEEMGEDLIGLHIHJFIKKFELFIKLEFGKGHEKHJLFHIGKFFDFFEFGKDKFHHCCCF    AS:i:254       NH:i:1  NM:i:0

HWI-ST700660_138:2:1201:2973:62218#2@0/1        0       20      76655   254     76M     *       0       0
AACCCCAAAAATGTTGGAAGAATAATGTAGGACATTGCAGAAGACGATGTTTAGATACTGAAAGGGACATACTTCT
FEFFGHHHGGHFKCCJKFHIGIFFIFLDEJKGJGGFKIHLFIJGIEGFLDEDFLFGEIIMHHIKL$BBGFFJIEHE    AS:i:254       NH:i:1  NM:i:1

HWI-ST700660_138:2:1203:21395:164917#2@0/1      256     20      68253   254     4M1D72M *       0       0
NCACCCATGATAGACCAGTAAAGGTGACCACTTAAATTCCTTGCTGTGCAGTGTTCTGTATTCCTCAGGACACAGA
#4@ADEHFJFFEJDHJGKEFIHGHBGFHHFIICEIIFFKKIFHEGJEHHGLELEGKJMFGGGLEIKHLFGKIKHDG    AS:i:254       NH:i:3  NM:i:1

HWI-ST700660_138:2:1105:16101:50526#6@0/1       16      20      126103  246     53M4D23M        *       0       0
AAGAAGTGCAAACCTGAAGAGATGCATGTAAAGAATGGTTGGGCAATGTGCGGCAAAGGGACTGCTGTGTTCCAGC
   FEHIGGHIGIGJI6FCFHJIFFLJJCJGJHGFKKKKGIJKHFFKIFFFKHFLKHGKJLJGKILLEFFLIHJIEIIB    AS:i:368       NH:i:1  NM:i:4
```

SAM Specification:
http://samtools.sourceforge.net/SAM1.pdf

VCF format

```
#fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

http://www.1000genomes.org/

Counts

Gene

Sample

Ensembl	Gene.Name	T1	T2	T3	T4	T5	WT1	WT2	WT3	WT4	WT5	WT6
ENSMUSG00000000134	Tfe3	312	295	333	258	392	257	344	223	423	277	389
ENSMUSG00000000142	Axin2	165	171	138	166	203	170	172	119	203	147	178
ENSMUSG00000000148	Brat1	213	196	207	224	350	204	268	143	300	177	288
ENSMUSG00000000149	Gna12	684	684	613	545	900	496	672	426	1023	583	797
ENSMUSG00000000154	Slc22a18	3	2	3	2	2	3	3	2	1	1	3
ENSMUSG00000000157	Itgb2l	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000159	Igsf5	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000167	Pih1d2	15	19	6	10	9	5	5	5	7	6	6
ENSMUSG00000000168	Dlat	899	777	967	756	1116	777	1047	614	1155	894	1126
ENSMUSG00000000171	Sdhd	1055	1003	1047	914	1430	939	1192	766	1390	916	1412
ENSMUSG00000000182	Fgf23	1	0	3	1	0	2	0	2	2	0	0
ENSMUSG00000000183	Fgf6	0	0	0	0	0	0	0	1	0	0	0
ENSMUSG00000000184	Ccnd2	1961	1978	1804	1779	2090	1655	2148	1585	2504	1895	2274
ENSMUSG00000000194	Gpr107	784	733	667	615	889	654	818	483	1034	627	1015
ENSMUSG00000000197	Nalcn	1120	1009	1047	917	1356	1129	1202	758	1625	1127	1044

Transcriptomic Studies

BIER

PRINCIPE FELIPE
CENTRO DE INVESTIGACION
Computational • Genomics

er ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED
DE ENFERMEDADES RARAS

fgardos@gmail.com

Máster en Biotecnología Biomédica. UPV

RNA-Seq Data Analysis Pipeline

Primary

1. Sequence preprocessing
2. Mapping
3. Quantification

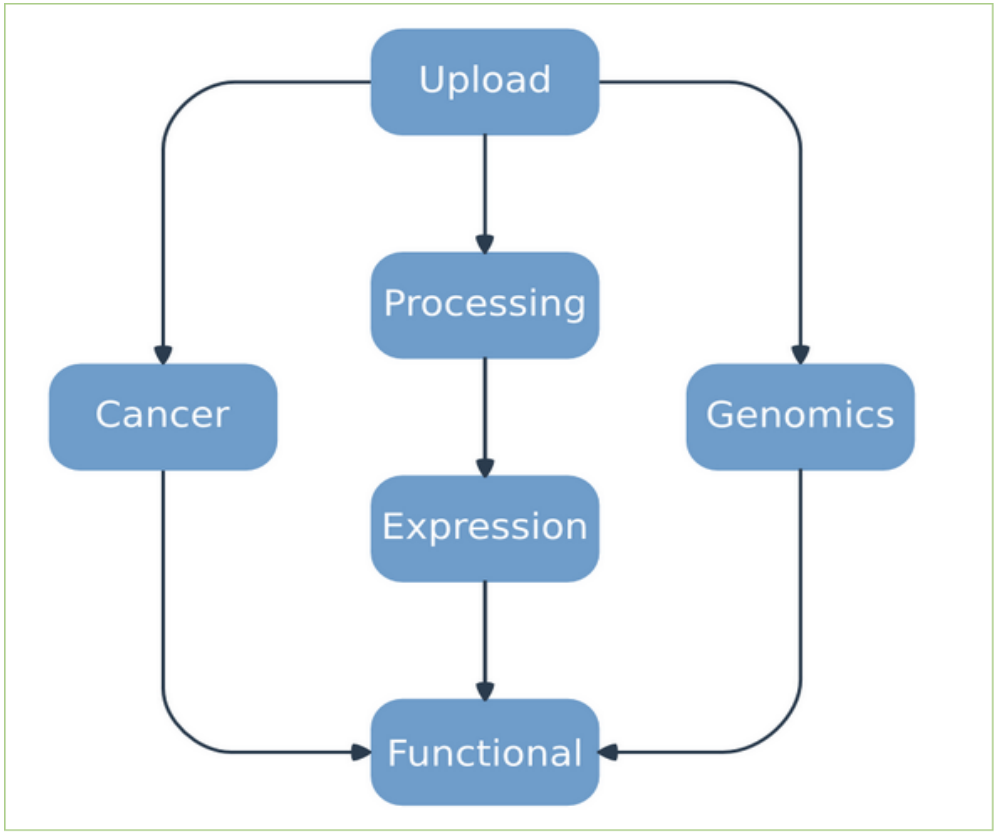
Secondary

4. Normalization
5. Differential expression
6. Functional Profiling

Babelomics 5

Babelomics 5

GENE EXPRESSION, GENOME VARIATION AND FUNCTIONAL PROFILING ANALYSIS SUITE



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

Babelomics 5

Analyzing omics data + functional profiling

Differential Expression

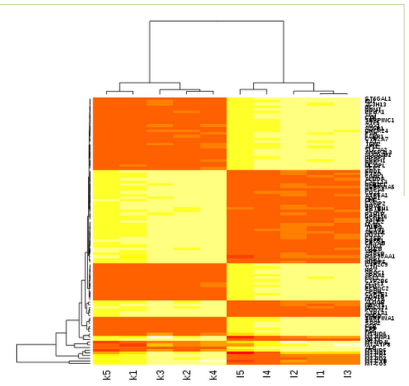
UPLOAD DATA

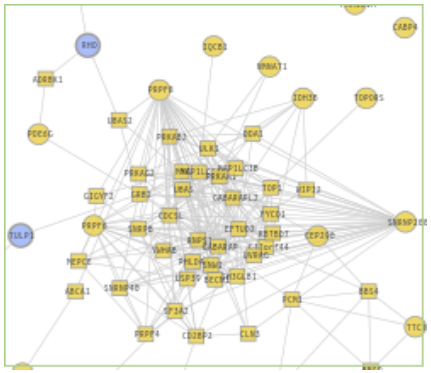
EDIT DATA

NORMALIZATION + DIFFERENTIAL EXPRESSION

FUNCTIONAL PROFILING

#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4
TSPAN6	203	198	194	176	202	157	190	200	201
TNMD	0	0	0	1	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47
SCYL3	21	30	31	27	31	28	31	37	15
C1orf112	10	12	8	11	18	17	22	12	12
FGR	19	28	18	20	10	47	50	43	49
FUCA2	240	272	261	256	211	76	82	85	68
GCLC	98	100	84	94	86	354	362	373	369
NFYA	59	61	53	56	59	59	66	63	66
STPG1	34	43	41	31	46	6	7	7	8

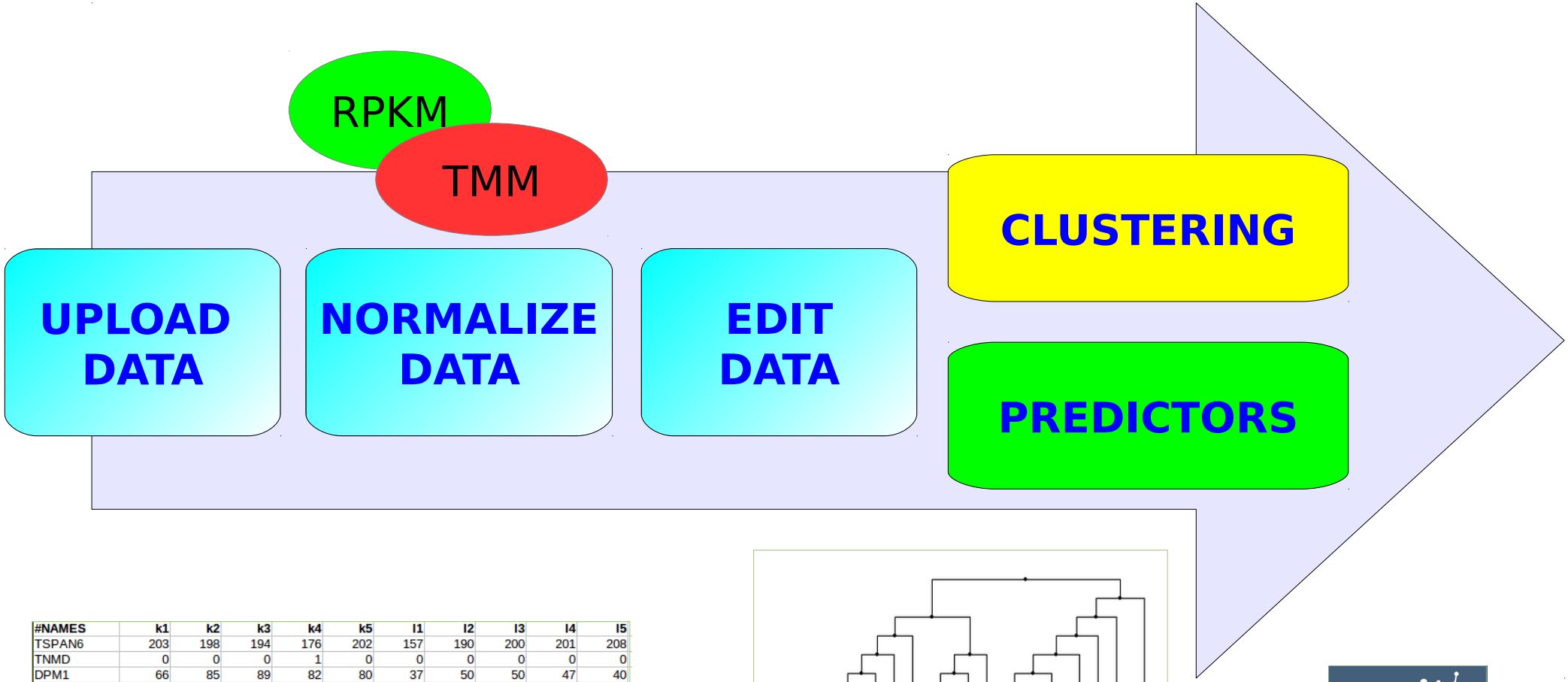




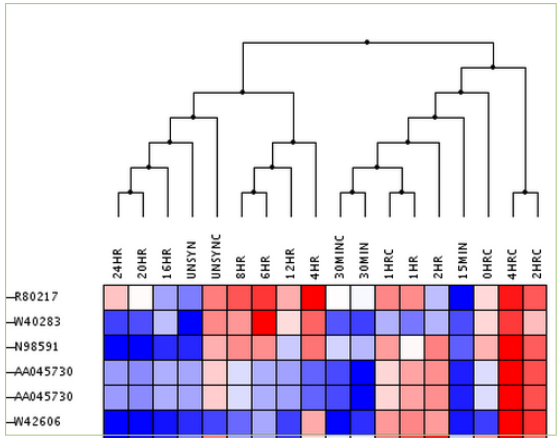
Babelomics 5

Analyzing omics data + functional profiling

Supervised and Unsupervised Classification

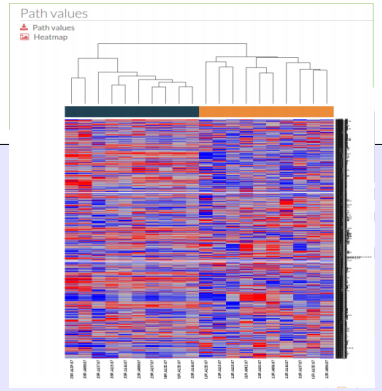


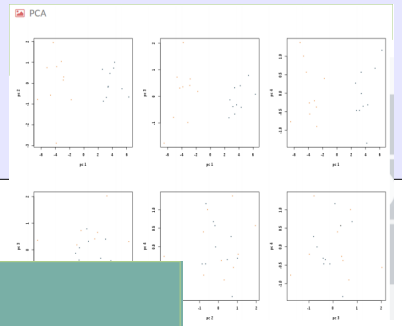
#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4	l5
TSPAN6	203	198	194	176	202	157	190	200	201	208
TNMD	0	0	0	1	0	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47	40
SCYL3	21	30	31	27	31	28	31	37	15	21
C1orf112	10	12	8	11	18	17	22	12	12	19
FGR	19	28	18	20	10	47	50	43	49	48
FUCA2	240	272	261	256	211	76	82	85	68	83
GCLC	98	100	84	94	86	354	362	373	369	326
NFYA	59	61	53	56	59	59	66	63	66	62
STPG1	34	43	41	31	46	6	7	7	8	7

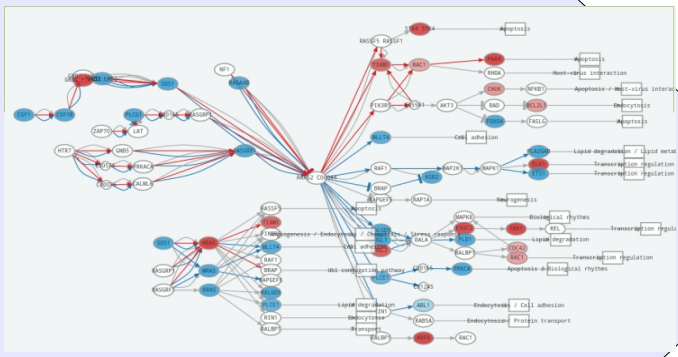


Signaling Pathways Analysis

#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4	l5
TSPAN6	203	198	194	176	202	157	190	200	201	208
TNMD	0	0	0	1	0	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47	40
SCYL3	21	30	31	27	31	28	31	37	15	21
C1orf112	10	12	8	11	18	17	22	12	12	19
FGR	19	28	18	20	10	47	50	43	49	48
FUCA2	240	272	261	256	211	76	82	85	68	83
GCLC	98	100	84	94	86	354	362	373	369	326
NFYA	59	61	53	56	59	59	66	63	66	62
STPG1	34	43	41	31	46	6	7	7	8	7







hiPathia

HIGH THROUGHPUT PATHWAY INFERENCE ANALYSIS

http://hipathia.babelomics.org/

Genomic Variation Studies

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

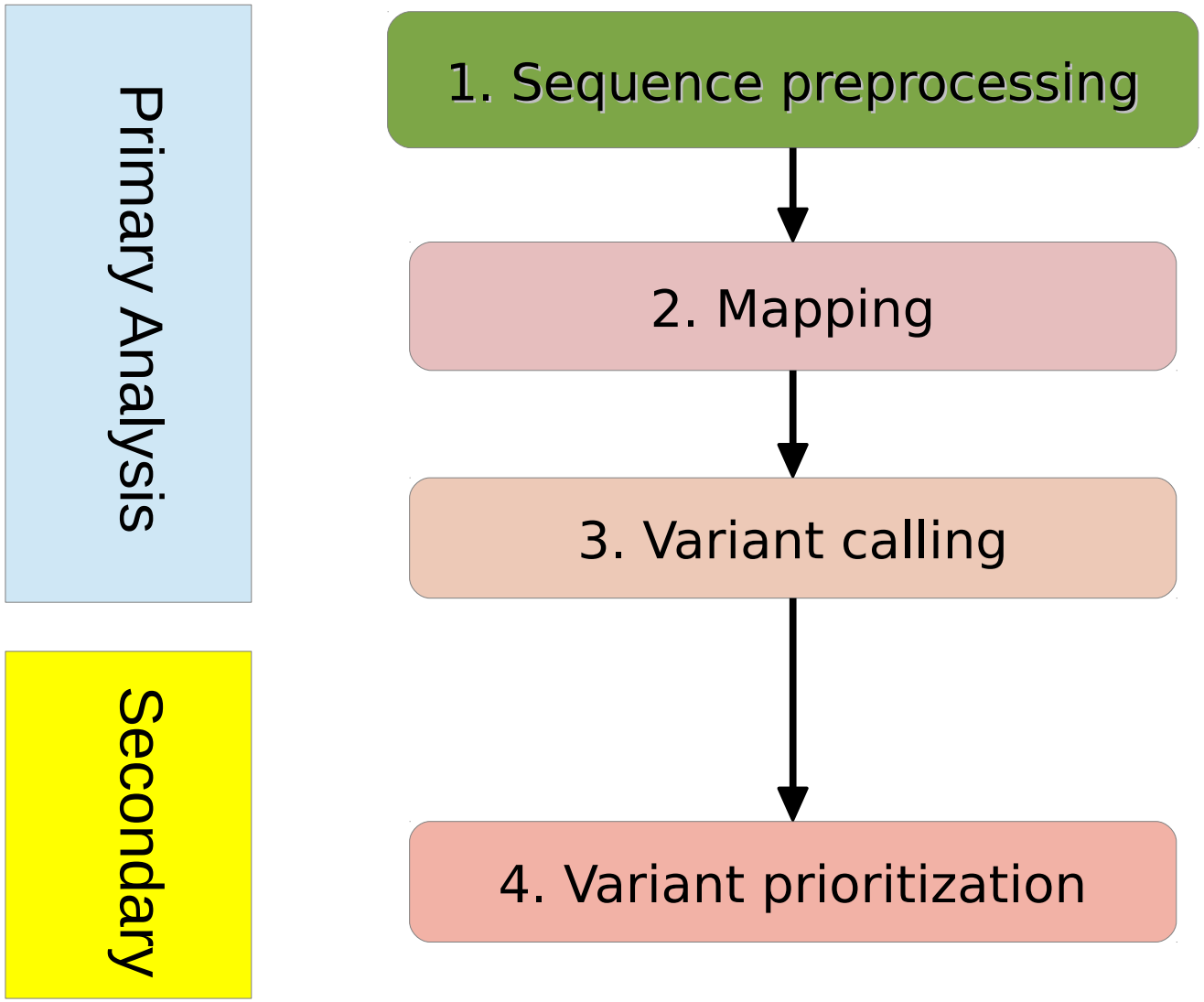
Computational • Genomics

ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED
DE ENFERMEDADES RARAS

fgardos@gmail.com

Máster en Biotecnología Biomédica. UPV

Genomics Data Analysis Pipeline



How do we prioritize variants in whole exome studies?

http://courses.babelomics.org/bierapp/

Introduction

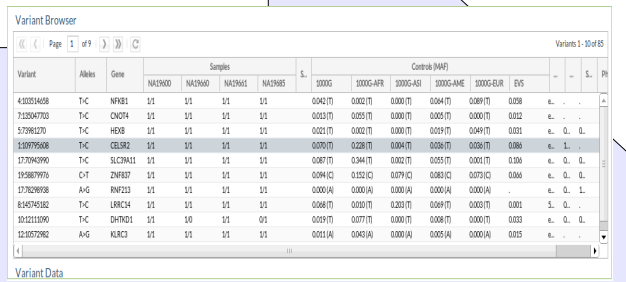
- Whole-exome sequencing has become a fundamental tool for the discovery of disease-related genes of familial diseases but there are difficulties to **find the causal mutation among the enormous background**
- There are different scenarios, so we need **different and immediate strategies of prioritization**
- Vast amount of **biological knowledge available** in many databases
- We need a tool to **integrate this information and filter immediately** to select candidate variants related to the disease

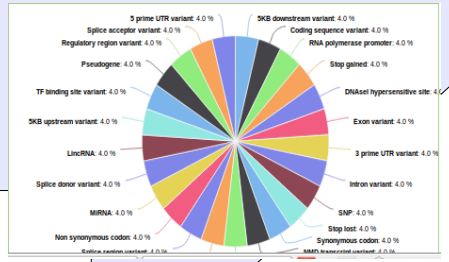
How does BiERapp work?

Filterings

VCF file multisample

BiERapp

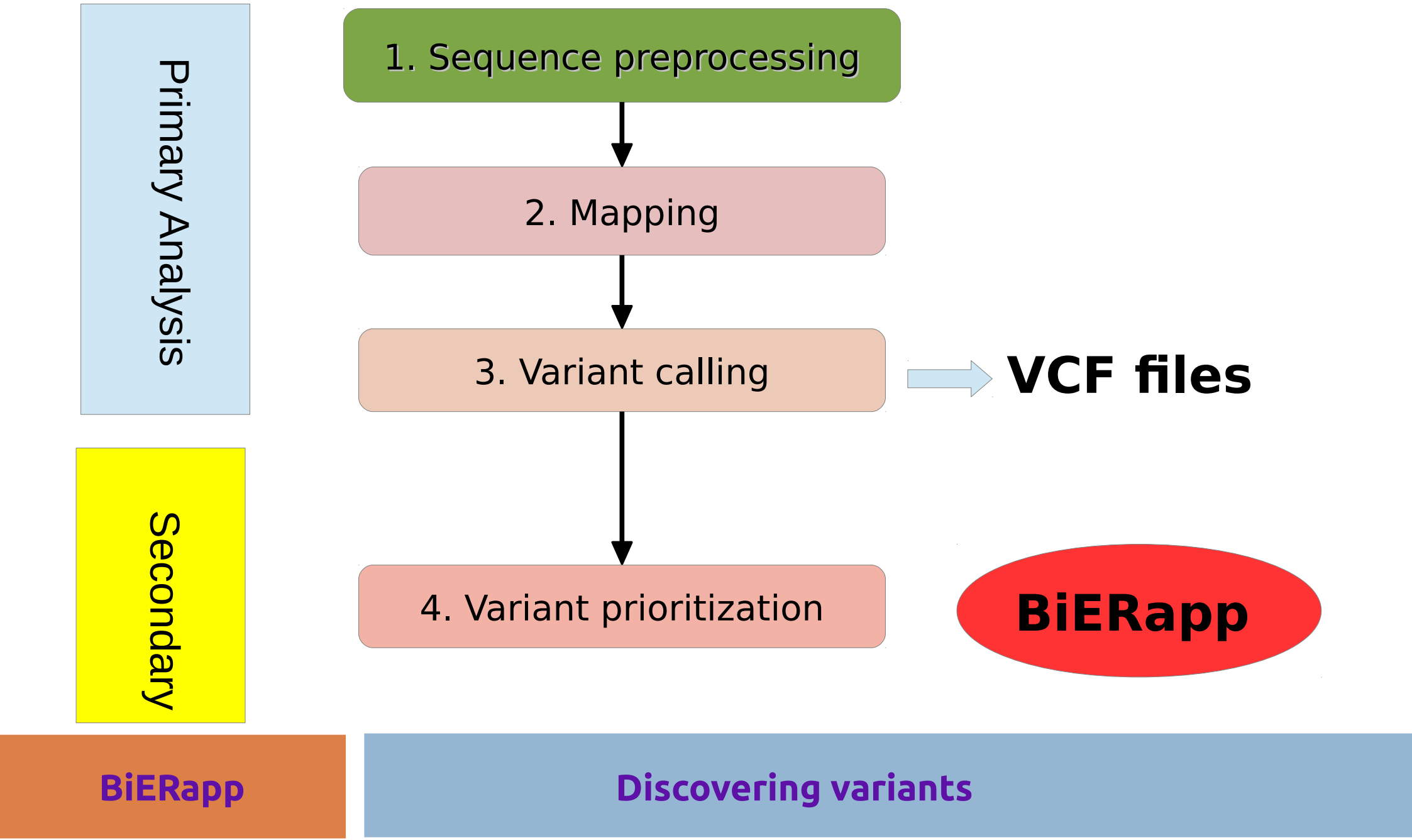




VARIANT CellBase

BiERapp | **Discovering variants**

Input: VCF file



Can I interpret sequencing data for diagnostic?

http://courses.babelomics.org/team/

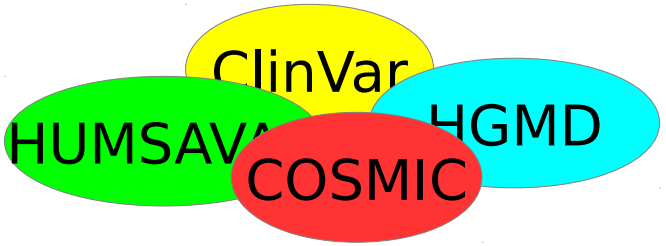
TEAM

Targeted Enrichment Analysis and Management

Gene panel

Sequencing data

Biological knowledge



TEAM

Diagnostic

TEAM | **Targeted Enrichment Analysis and Management**

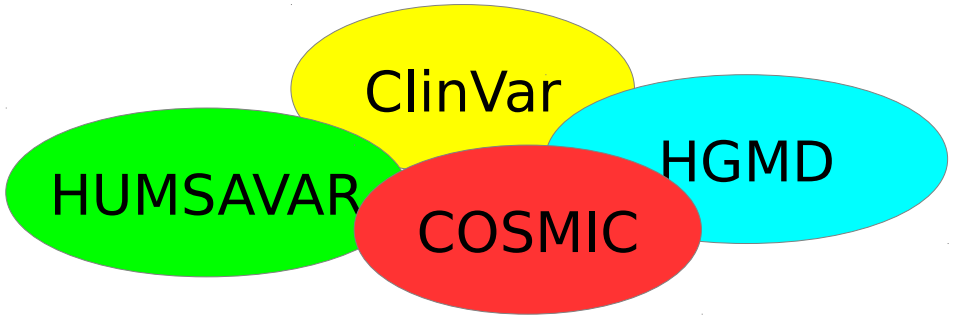
Gene panel

1. VCF files
2. Gene panel

TEAM

Diagnostic: Retinitis

	Chromosome	Position	SNP Id	Ref	Alt	Gene	Con
1	3	129247734	.	T	C	.	exon ator
2	3	129247734	.	T	C	RHO	exon ator
3	3	129247734	.	T	C	RHO	exon ator
4	3	129247734	.	T	C	RHO	exon ator



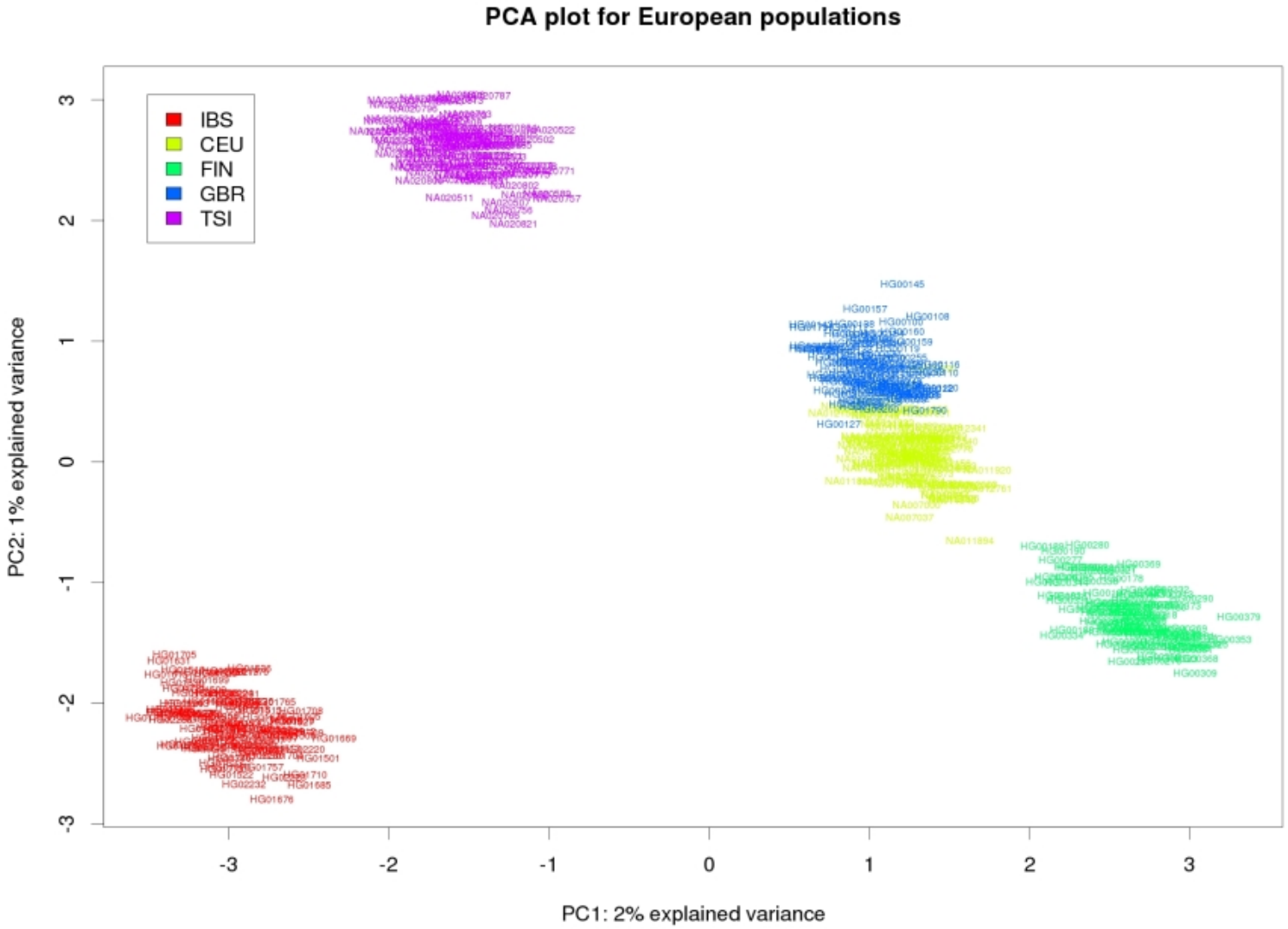
TEAM — **Targeted Enrichment Analysis and Management**

CSVS:
CIBERER Spanish Variant Server

Repositorio de frecuencias de variantes
en la población española

http://csvs.babelomics.org/

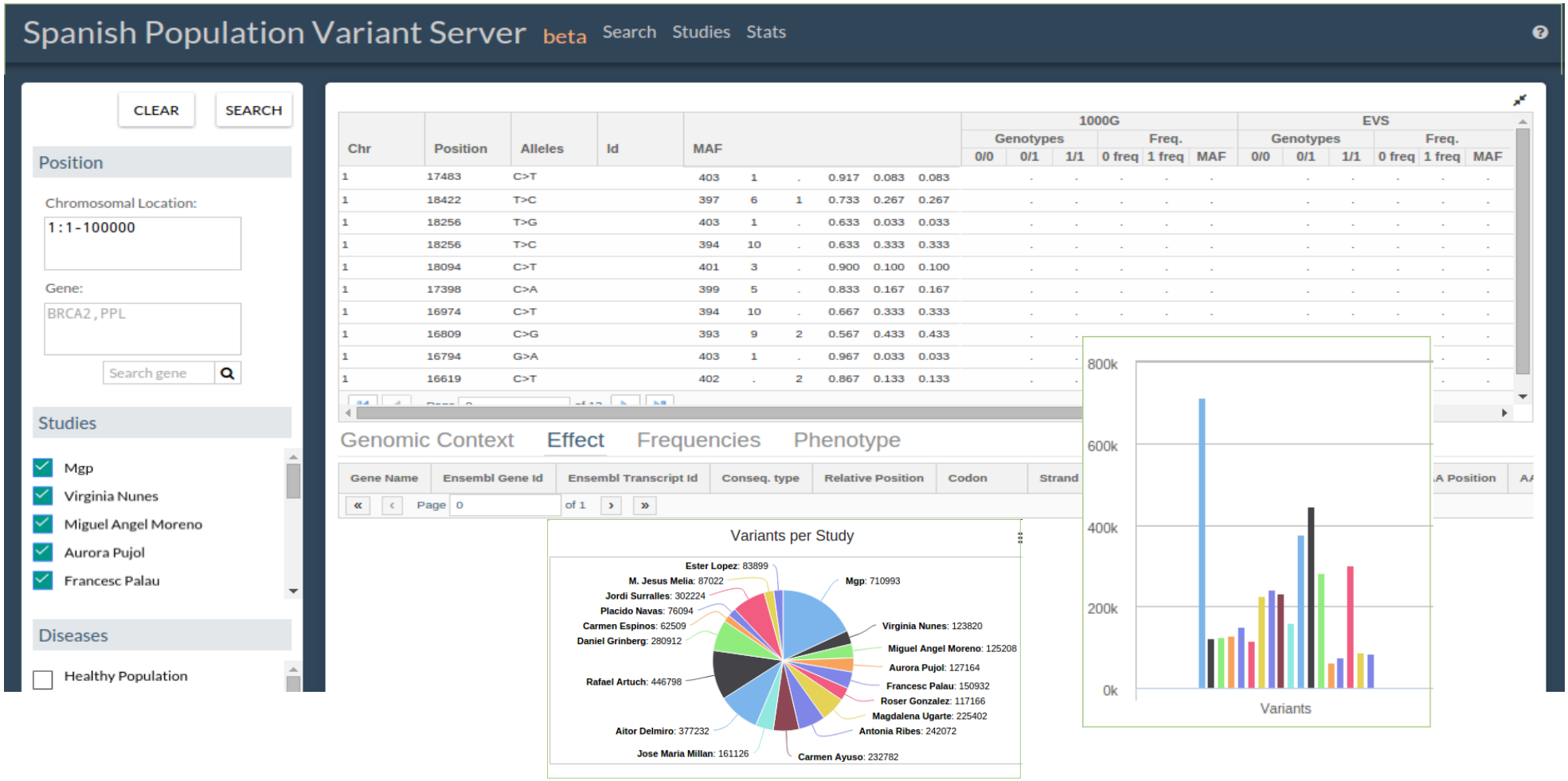
CIBERER Spanish Variant Server



CSVS

Local genetic variability

Tool interface



http://csvs.babelomics.org/