How do we prioritize variants in whole exome studies?

MDA course on NGS Data Analysis
Valencia, 28 Sep 2015

Francisco García
fgarcia@cipf.es

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

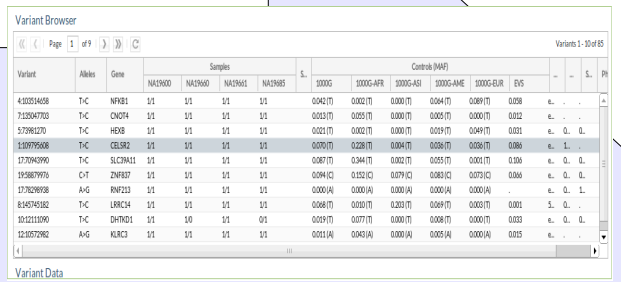
- We need a tool to **integrate this information and filter immediately** to select candidate variants related to the disease

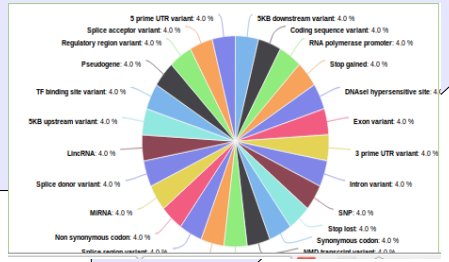
How does BiERapp work?

Filterings

VCF file multisample

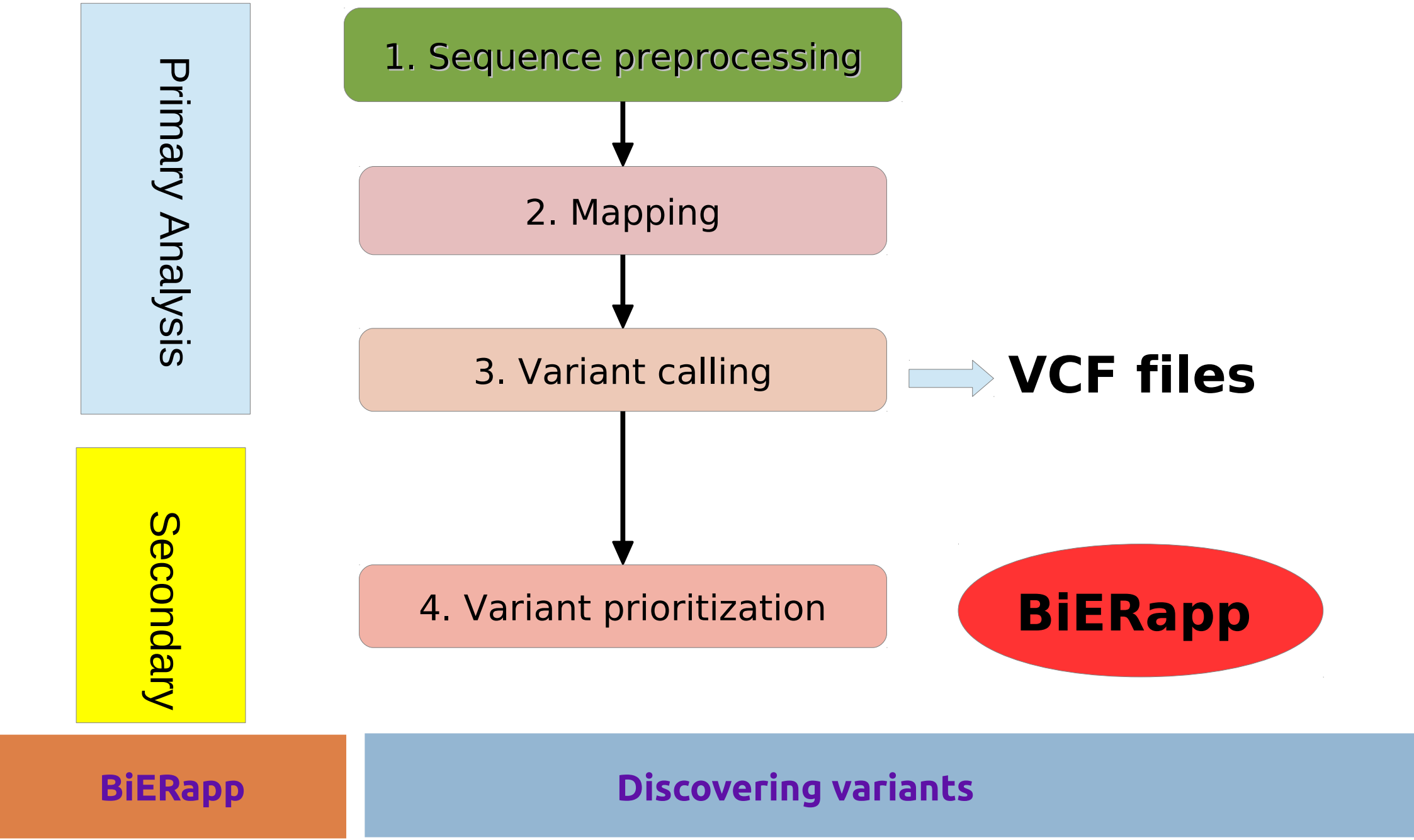
BiERapp





VARIANT CellBase

Input: VCF file



Input: VCF multisample

```
##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
```

One VCF (Variant Calling Format) **file for family or group**

Getting information

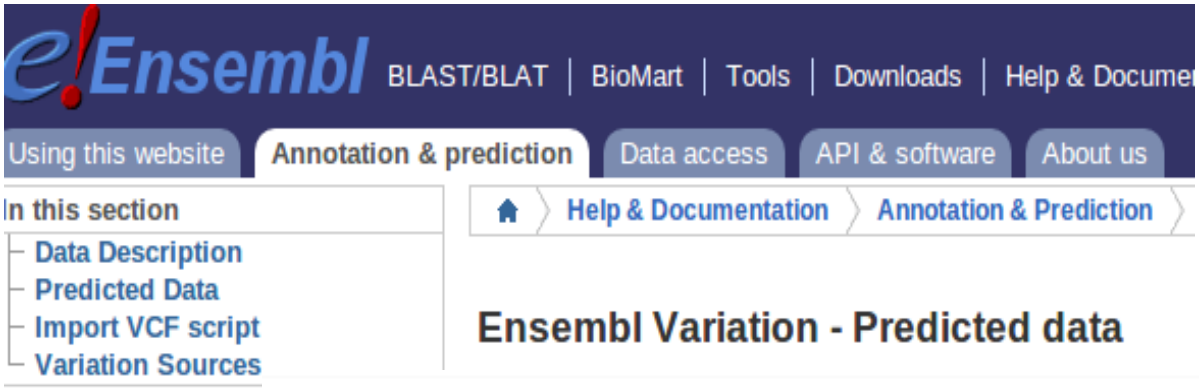
- **SIFT**
  - SIFT predicts whether an amino acid substitution affects protein function
  - **Interpretation**: 1 (tolerated) to 0 (not tolerated)

http://sift.jcvi.org/

- **PolyPhen**
  - Polymorphism Phenotyping is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein
  - **Interpretation**: 1 (probably damage) to 0 (benign)

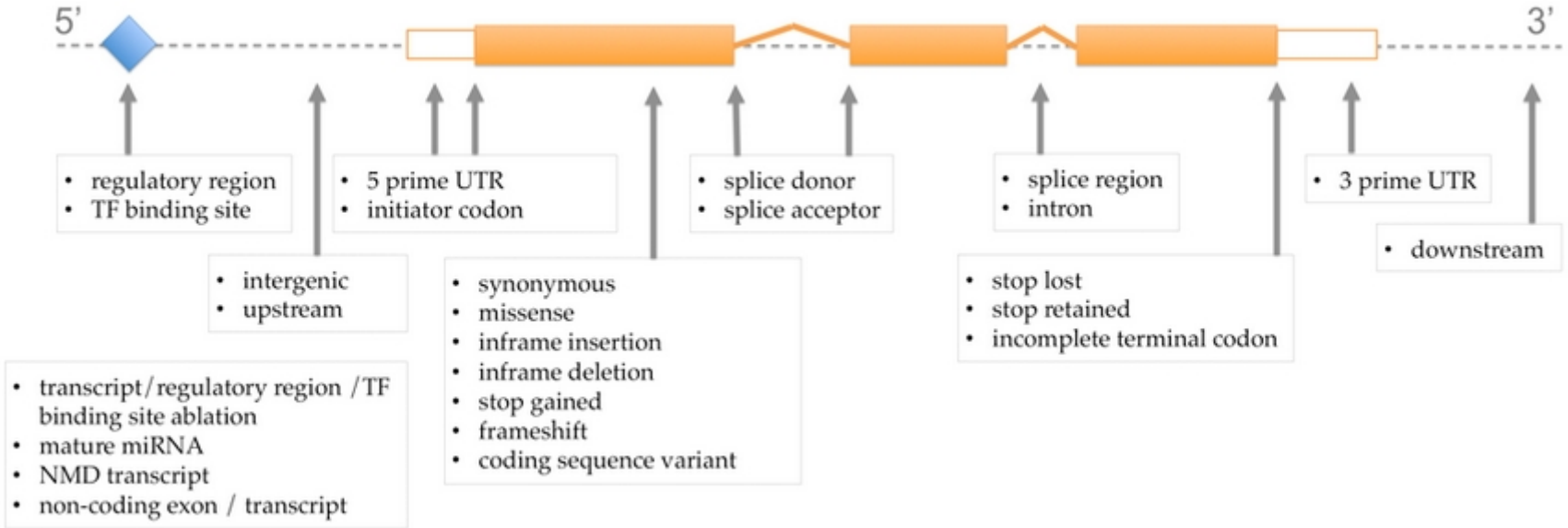
http://genetics.bwh.harvard.edu/pph2/index.shtml

Getting information



Consequence type or effect

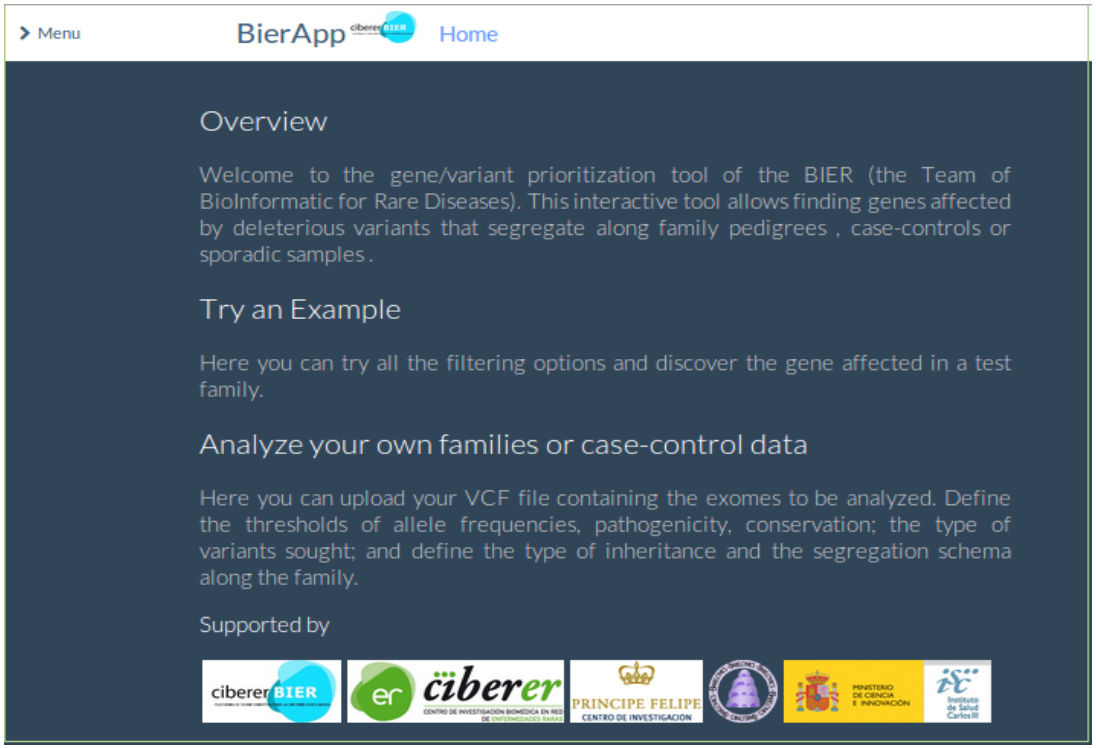
Ensembl Variation - Predicted data



http://www.ensembl.org/info/genome/variation/predicted_data.html

Tool interface

http://bierapp.babelomics.org/



Tool interface

Menu BierApp Home

Example 1000G (Short)

Filter

Clear Submit

Segregation

	0/0	0/1	1/1
NA19600:	☐	☑	☑
NA19660:	☐	☑	☑
NA19661:	☐	☐	☐
NA19685:	☐	☐	☐

MAF

1000G MAF <: 0.1

EVS MAF <:

1000G Populations

African MAF <:

American MAF <:

Asian MAF <:

European MAF <:

Position

Consequence Type

- ☐ 5KB_downstream_variant
- ☐ coding_sequence_variant
- ☐ RNA_polymerase_promoter
- ☐ stop_gained
- ☐ DNAseI_hypersensitive_site
- ☑ exon_variant
- ☐ 3_prime_UTR_variant
- ☐ intron_variant
- ☐ SNP
- ☐ stop_lost
- ☐ synonymous_codon
- ☐ NMD_transcript_variant
- ☐ CpG_island
- ☐ miRNA_target_site

Variant Browser

Page 1 of 9 — Variants 1 - 10 of 85

Variant	Alleles	Gene	Samples				...	Controls (MAF)						...	...	...	Ph
			NA19600	NA19660	NA19661	NA19685		1000G	1000G-AFR	1000G-ASI	1000G-AME	1000G-EUR	EVS				
4:103514658	T>C	NFKB1	1/1	1/1	1/1	1/1		0.042 (T)	0.002 (T)	0.000 (T)	0.064 (T)	0.089 (T)	0.058	e...	.	.	
7:135047703	T>C	CNOT4	1/1	1/1	1/1	1/1		0.013 (T)	0.055 (T)	0.000 (T)	0.005 (T)	0.000 (T)	0.012	e...	.	.	
5:73981270	T>C	HEXB	1/1	1/1	1/1	1/1		0.021 (T)	0.002 (T)	0.000 (T)	0.019 (T)	0.049 (T)	0.031	e...	0...	0...	
1:109795608	T>C	CELSR2	1/1	1/1	1/1	1/1		0.070 (T)	0.228 (T)	0.004 (T)	0.036 (T)	0.036 (T)	0.086	e...	1...	.	
17:70943990	T>C	SLC39A11	1/1	1/1	1/1	1/1		0.087 (T)	0.344 (T)	0.002 (T)	0.055 (T)	0.001 (T)	0.106	e...	0...	0...	
19:58879976	C>T	ZNF837	1/1	1/1	1/1	1/1		0.094 (C)	0.152 (C)	0.079 (C)	0.083 (C)	0.073 (C)	0.066	e...	0...	0...	
17:78298938	A>G	RNF213	1/1	1/1	1/1	1/1		0.000 (A)	0.000 (A)	0.000 (A)	0.000 (A)	0.000 (A)	.	e...	0...	1...	
8:145745182	T>C	LRRC14	1/1	1/1	1/1	1/1		0.068 (T)	0.010 (T)	0.203 (T)	0.069 (T)	0.003 (T)	0.001	5...	0...	.	
10:12111090	T>C	DHTKD1	1/1	1/0	1/1	0/1		0.019 (T)	0.077 (T)	0.000 (T)	0.008 (T)	0.000 (T)	0.033	e...	0...	0...	

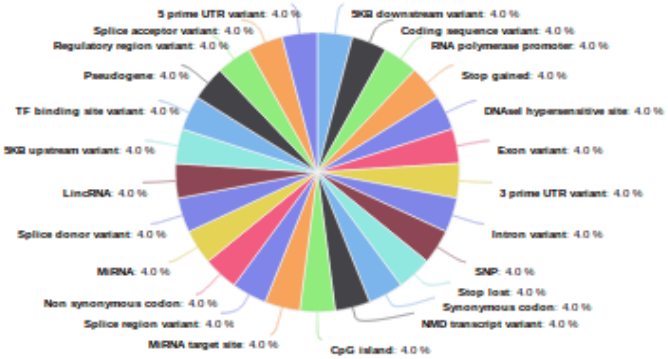
Variant Data

Genomic Context | Effect & Annotation | Study Summary

Effects

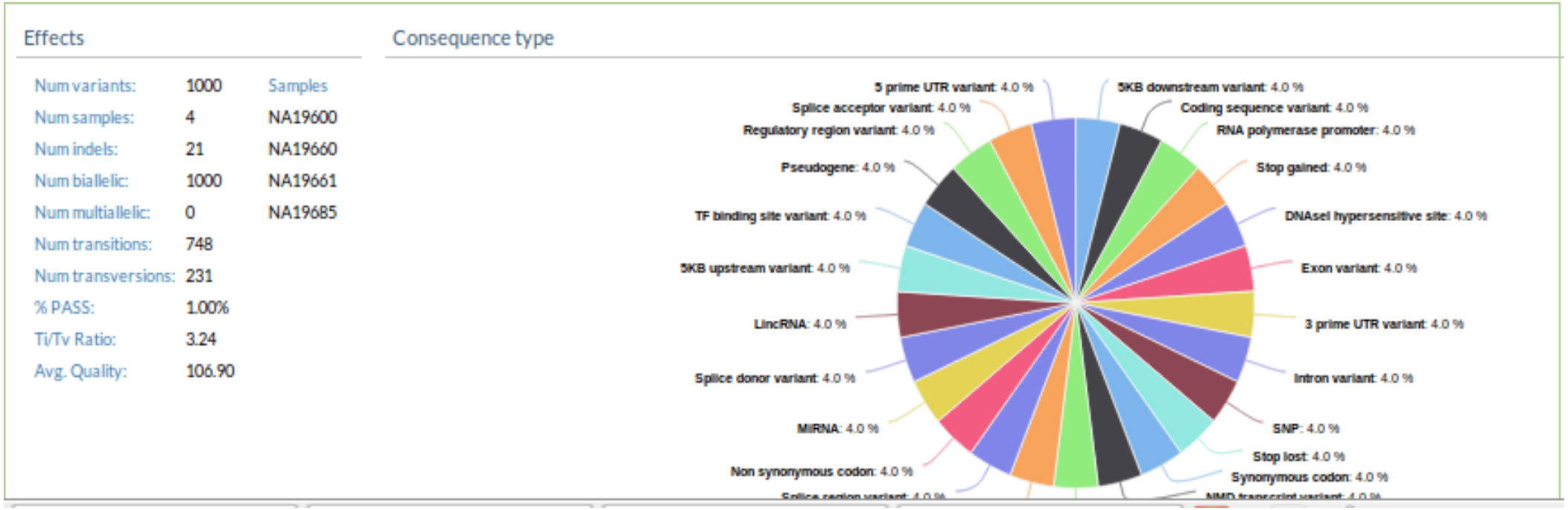
		Samples
Num variants:	1000	NA19600
Num samples:	4	NA19660
Num indels:	21	NA19661
Num biallelic:	1000	NA19685
Num multiallelic:	0	
Num transitions:	748	
Num transversions:	231	
% PASS:	100%	
Ti/Tv Ratio:	3.24	
Avg. Quality:	106.90	

Consequence type



Results

1. Summary. Description about number or variants, INDELs... Also a distribution of consequences types.



Results

2. List of candidate variants.
We can order this list by several criteria.

Variant Browser

Page 1 of 9 | Variants 1 - 10 of 85

Variant	Alleles	Gene	Samples				S...	Controls (MAF)						...	...	S...	Ph
			NA19600	NA19660	NA19661	NA19685		1000G	1000G-AFR	1000G-ASI	1000G-AME	1000G-EUR	EVS				
4:103514658	T>C	NFKB1	1/1	1/1	1/1	1/1		0.042 (T)	0.002 (T)	0.000 (T)	0.064 (T)	0.089 (T)	0.058	e...	.	.	
7:135047703	T>C	CNOT4	1/1	1/1	1/1	1/1		0.013 (T)	0.055 (T)	0.000 (T)	0.005 (T)	0.000 (T)	0.012	e...	.	.	
5:73981270	T>C	HEXB	1/1	1/1	1/1	1/1		0.021 (T)	0.002 (T)	0.000 (T)	0.019 (T)	0.049 (T)	0.031	e...	0...	0...	
1:109795608	T>C	CELSR2	1/1	1/1	1/1	1/1		0.070 (T)	0.228 (T)	0.004 (T)	0.036 (T)	0.036 (T)	0.086	e...	1...	.	
17:70943990	T>C	SLC39A11	1/1	1/1	1/1	1/1		0.087 (T)	0.344 (T)	0.002 (T)	0.055 (T)	0.001 (T)	0.106	e...	0...	0...	
19:58879976	C>T	ZNF837	1/1	1/1	1/1	1/1		0.094 (C)	0.152 (C)	0.079 (C)	0.083 (C)	0.073 (C)	0.066	e...	0...	0...	
17:78298938	A>G	RNF213	1/1	1/1	1/1	1/1		0.000 (A)	0.000 (A)	0.000 (A)	0.000 (A)	0.000 (A)	.	e...	0...	1...	
8:145745182	T>C	LRRC14	1/1	1/1	1/1	1/1		0.068 (T)	0.010 (T)	0.203 (T)	0.069 (T)	0.003 (T)	0.001	5...	0...	.	
10:12111090	T>C	DHTKD1	1/1	1/0	1/1	0/1		0.019 (T)	0.077 (T)	0.000 (T)	0.008 (T)	0.000 (T)	0.033	e...	0...	0...	
12:10572982	A>G	KLRC3	1/1	1/1	1/1	1/1		0.011 (A)	0.043 (A)	0.000 (A)	0.005 (A)	0.000 (A)	0.015	e...	.	.	

Variant Data

Results

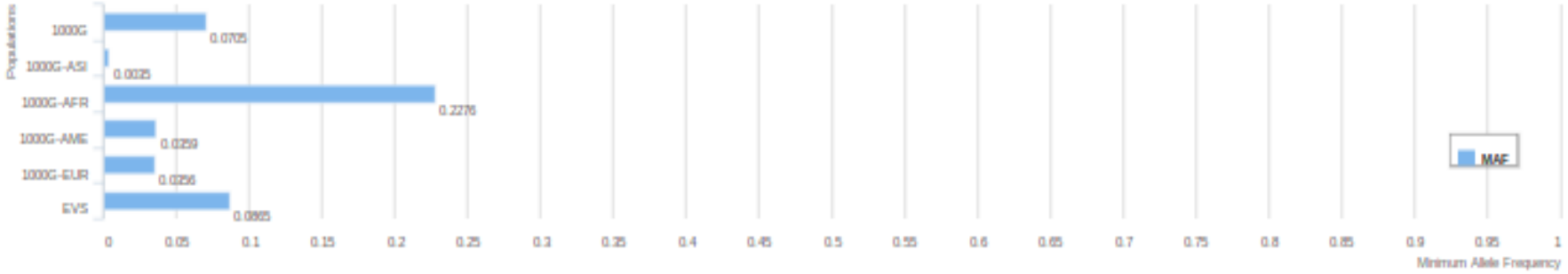
3. Effects for each transcript where we detected a candidate variant.

The plot shows MAFs for different groups (1000 Genomes, Exome Variant Server)

Effects

	Position chr:start...	SNP Id	Conseq. Type	Aminoacid Change	Gene (EnsemblId)	Transcript Id	Feature Id	Feature Name	Feature Type	Feature Biotype
1	1:109792641-109...		non_synonymous...	S/P-TCA/CCA (9...	CELSR2 (ENSG00...	ENST00000271332	ENSE00000958214	CELSR2	exon	protein_coding
2	1:109788400-109...		regulatory_regio...	.	.		H3K27me3		regulatory_region	
3	1:109792400-109...		regulatory_regio...	.	.		H3K36me3		regulatory_region	
4	1:109795608-109...	rs437444	SNP (SO:0000694)	.	.		rs437444		snp	

Population Frequencies



Results

4. Visualization of candidate variants from GenomeMaps

Remarks

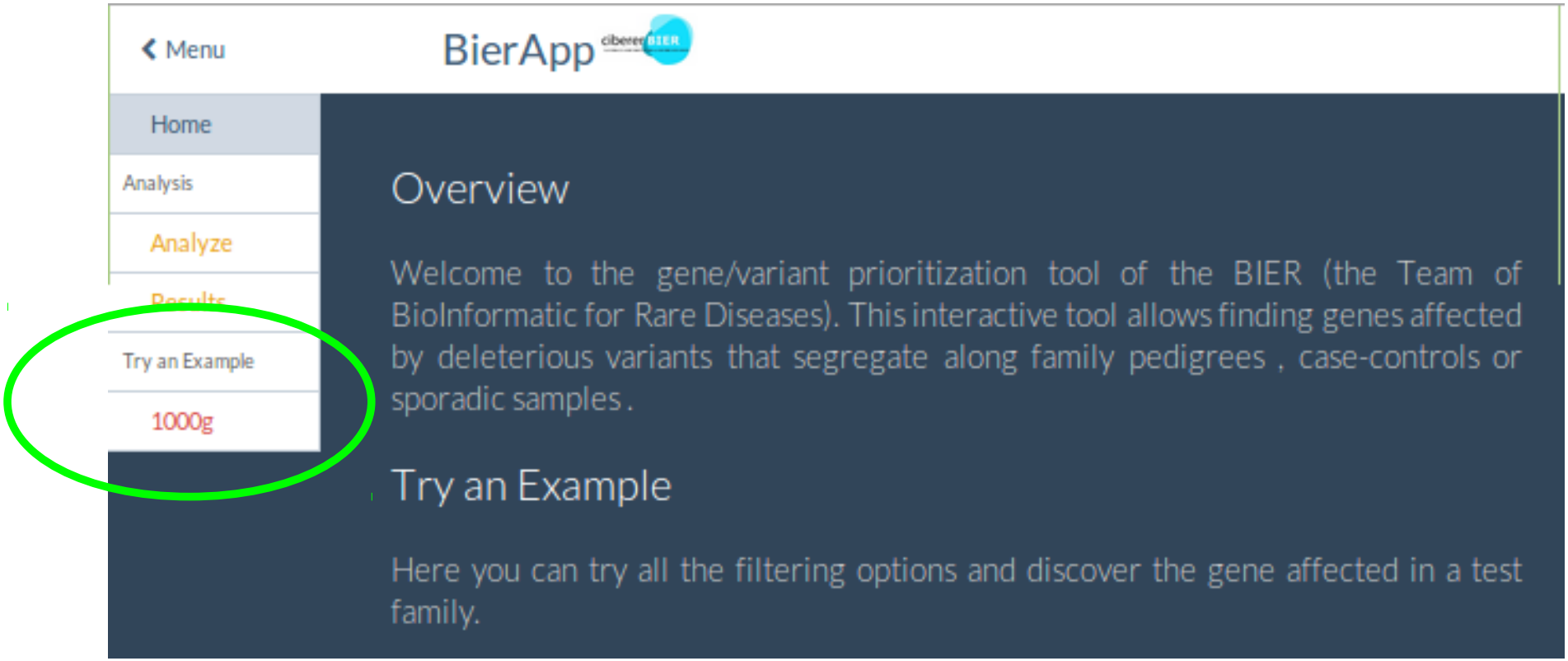
- The proposed web-based interactive framework has **great potential to detect disease-related variants** in familial diseases as demonstrated by its successful use in several studies

- **The use of the filters is interactive** and the results are almost instantaneously displayed in a panel that includes the genes affected, the variants and specific information for them

- Candidate variants are **new knowledge useful for future diagnostic**

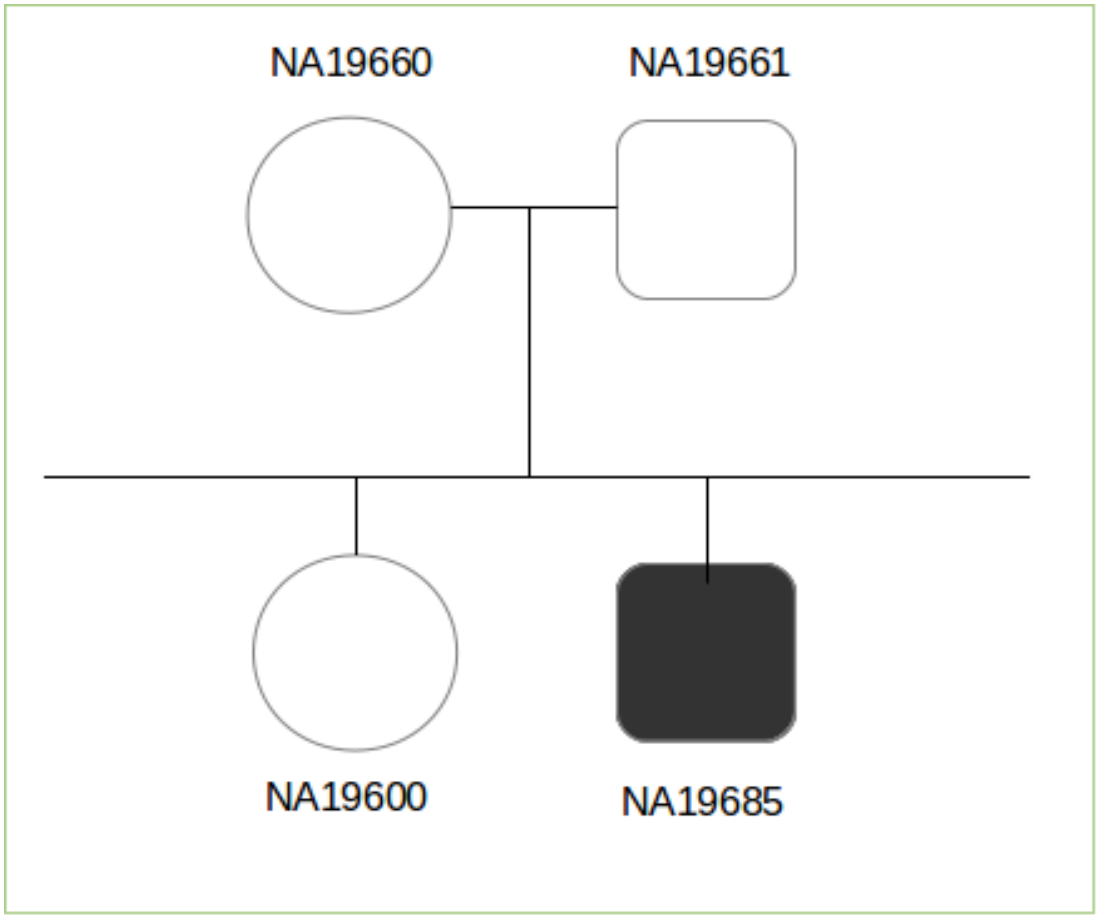
Hands on

http://bierapp.babelomics.org/



Hands on

Pedigree

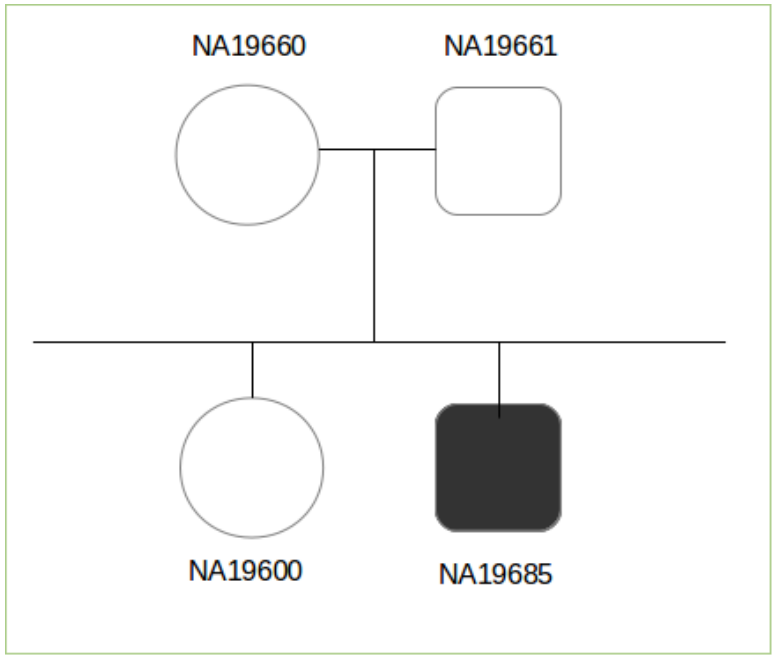


Hands on

Case 1.

- De novo variants

  How many variants?



Case 2.

- Recessive heritage

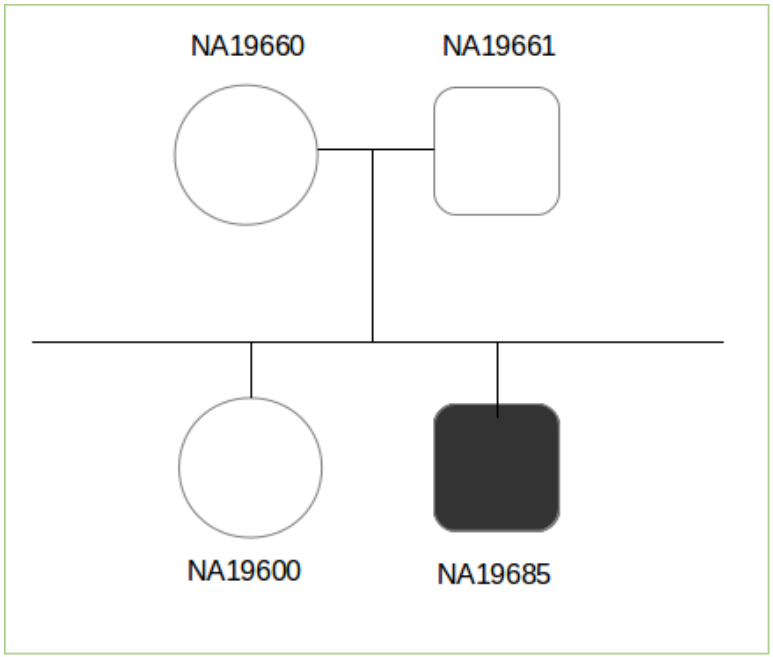
  How many variants?

Hands on

Case 3.

- Recessive heritage
- Rare disease (MAF < 0.1)

How many variants?



Case 4.

- Variants in mother and daughter at the same time

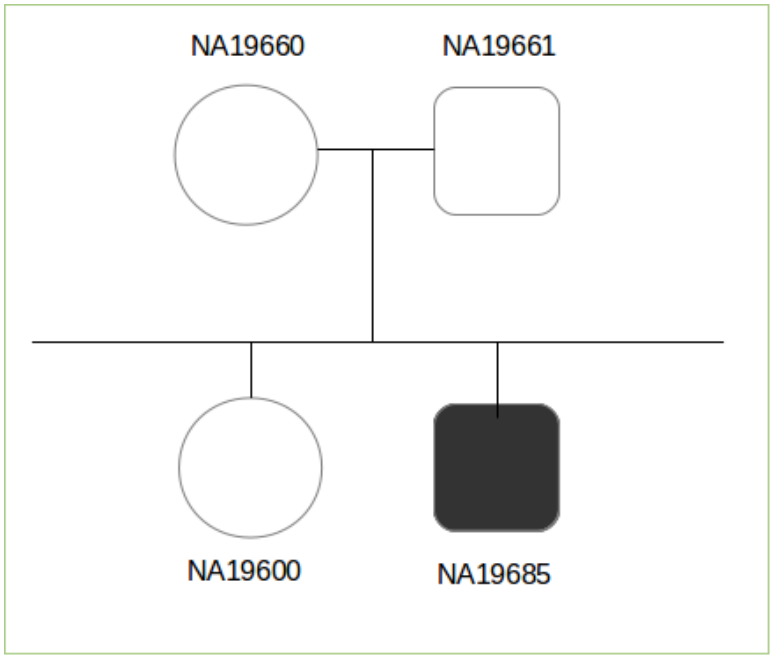
How many variants?

Hands on

Case 5.

- Variants in mother and daughter at the same time
- Only in chromosome 4

How many variants?



Case 6.

- Variants in mother and daughter at the same time
- Only in these genes: HEXB,NFKB1,KLRC3

How many variants?

More information

Nucleic Acids Research Advance Access published May 6, 2014
Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku407

A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies

Alejandro Alemán[1,2], Francisco Garcia-Garcia[1], Francisco Salavert[1,2], Ignacio Medina[1] and Joaquín Dopazo[1,2,3,*]

[1]Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia 46012, Spain,
[2]Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras (CIBERER), Valencia 46010, Spain and
[3]Functional Genomics Node, (INB) at CIPF, Valencia 46012, Spain

BIER

BiERapp Tutorial:

http://bierapp.babelomics.org/