Prioritization of variants and genes: *BiERapp*

Álex Alemán

March 1st, 2016

GDA

International Course on

Genomic Data Analysis

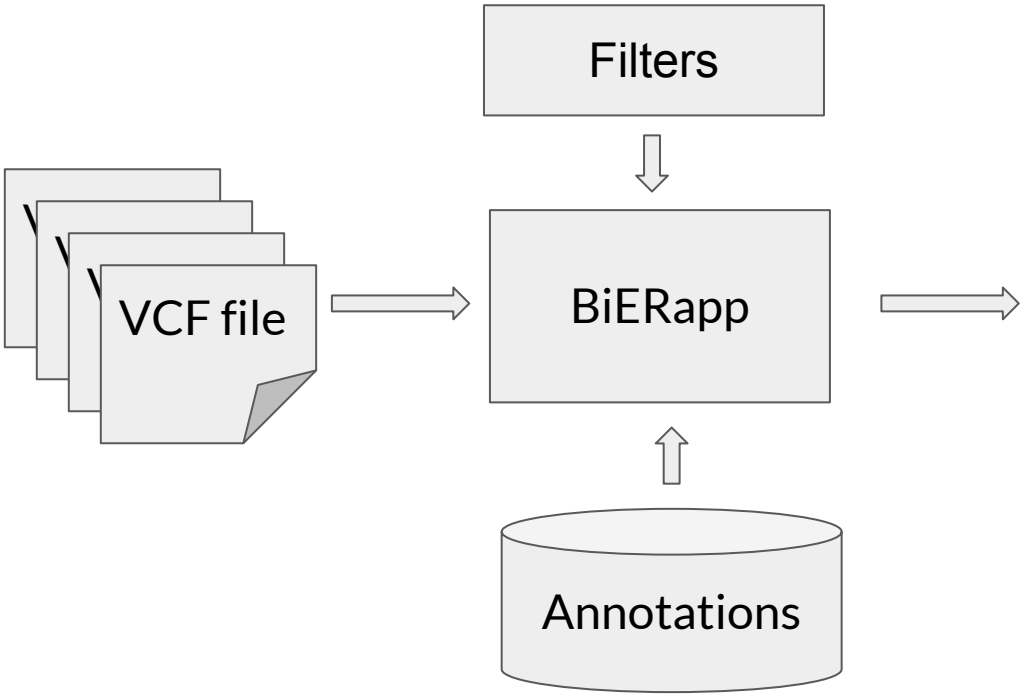
PRINCIPE FELIPE

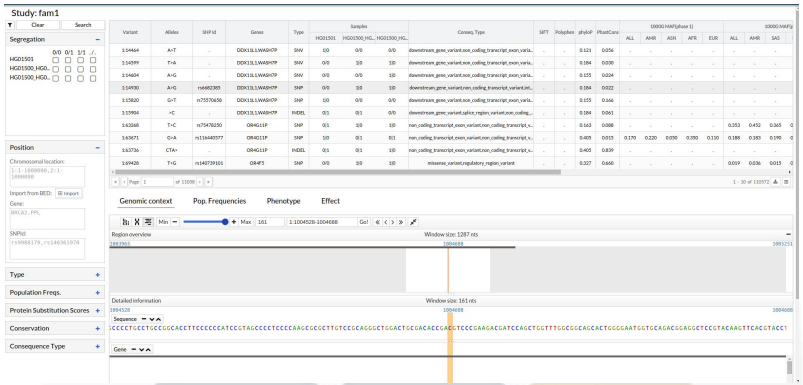
CENTRO DE INVESTIGACION

Introduction

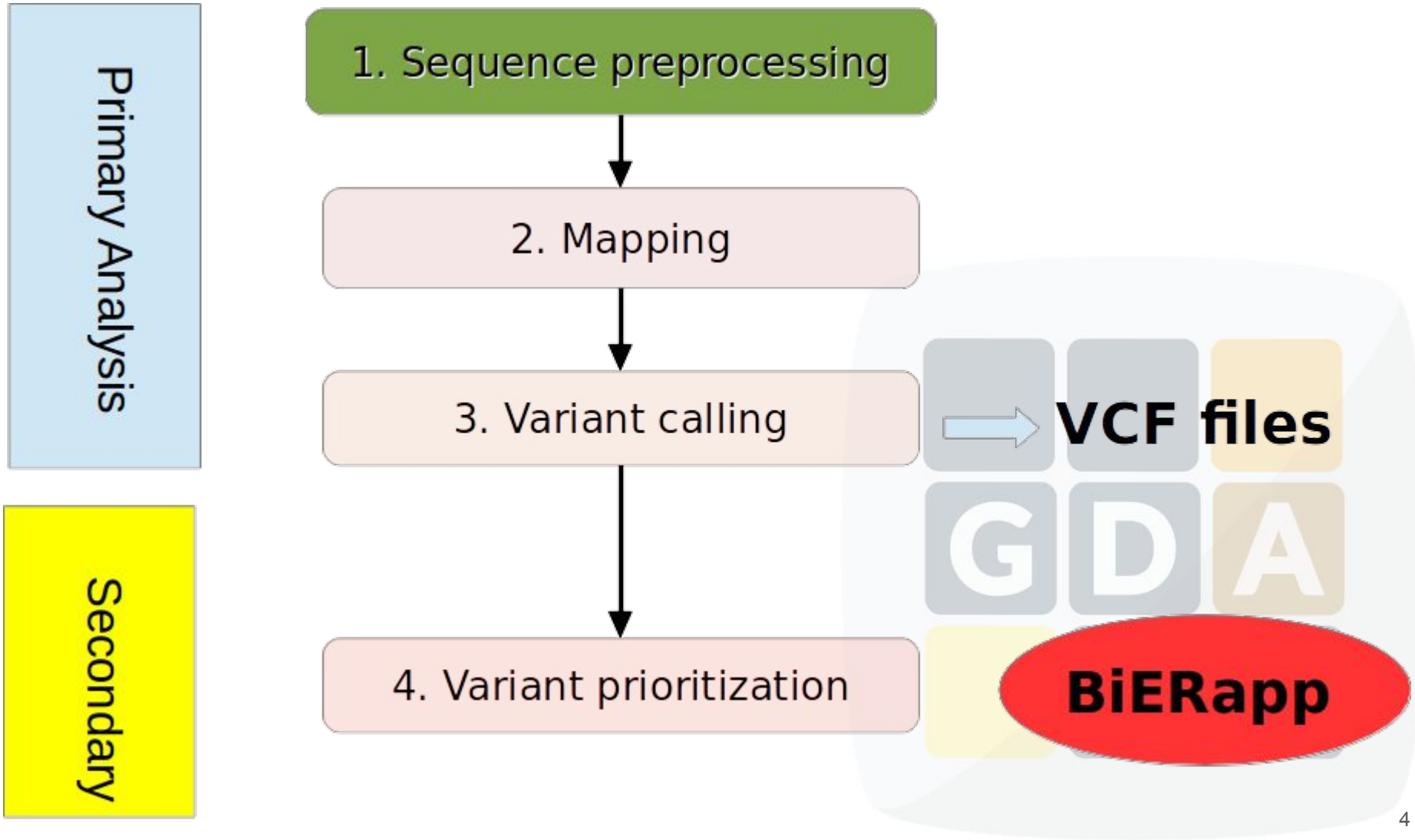
- Whole-exome sequencing has become a fundamental tool for the discovery of disease-related genes of familial diseases but there are difficulties to **find the causal mutation among the enormous background**.
- There are different scenarios, so we need **different and immediate strategies of prioritization**.
- 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?





Input: VCF file



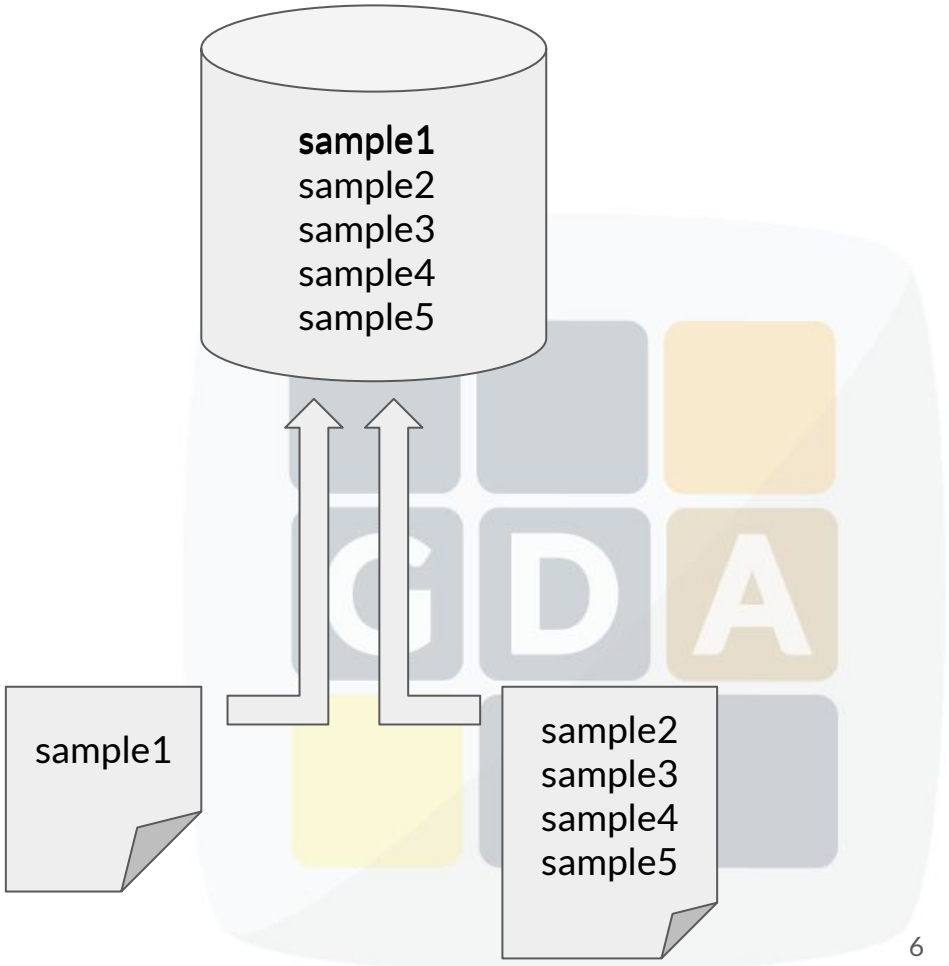
Input: VCF file

```
##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
```

- We can upload multiple VCF single/multi sample.
- You do not need to create a multi-sample file with all the samples.
- BiERapp will merge all the samples from those files in the database.

Input: VCF multisample

- Create a Study
- Upload a new single-sample file
- BiERapp stores the sample in the created study
- Now we upload a new multisample file with 4 more samples
- BiERapp merges these samples in the study



Getting information: SIFT & PolyPhen

- **SIFT**
  - **SIFT** predicts whether an amino acid substitution affects protein function
  - **Interpretation:** 1 (tolerated) to 0 (deleterious)

http://sift.jcvi.org/

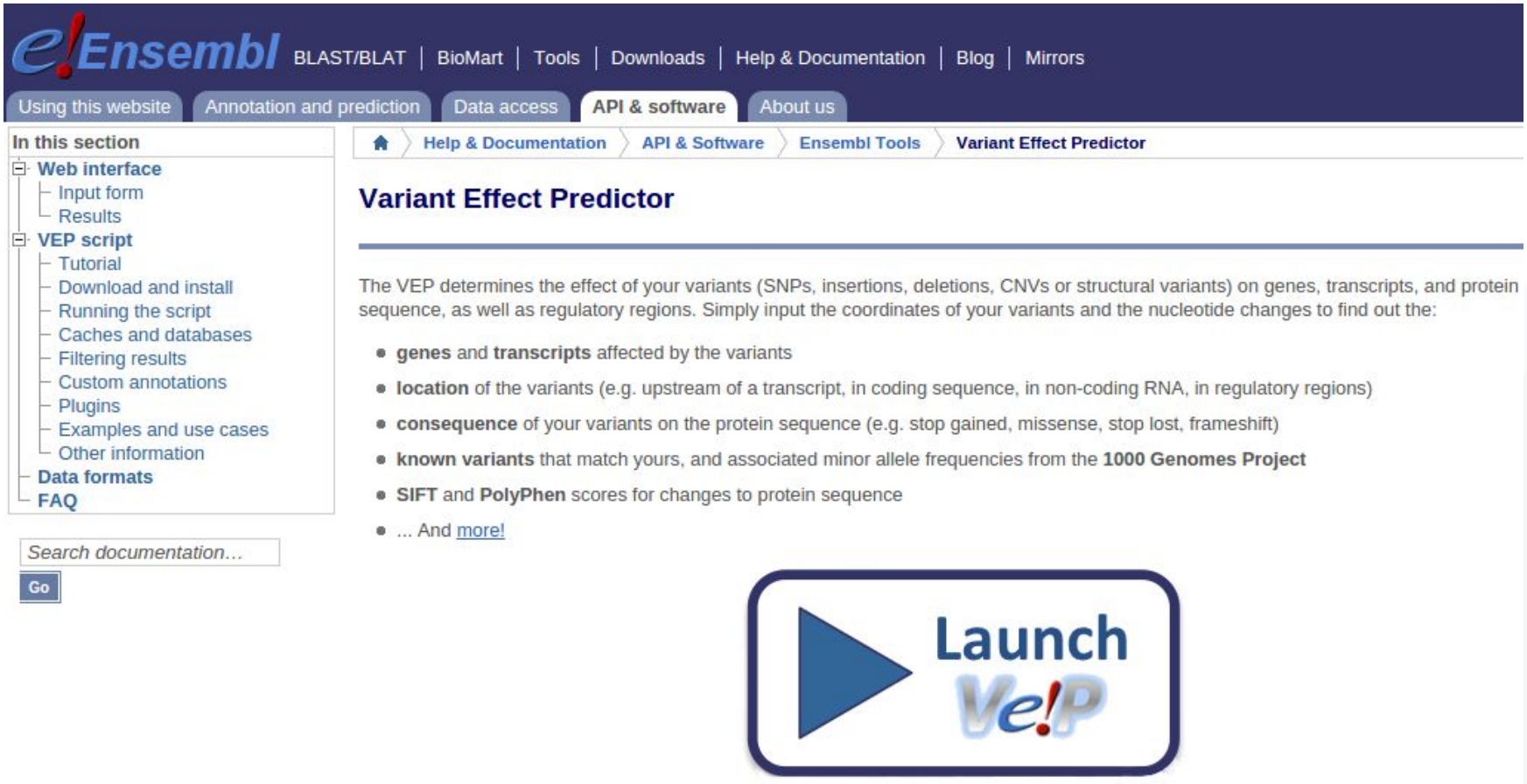
- **PolyPhen**
  - Polymorphism Phenotyping is a tool which preditcs possible impact of an amino acid substitution on the structure and function of a human protein.
  - **Interpretation:** 1 (probably damage) to 0 (bening)

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

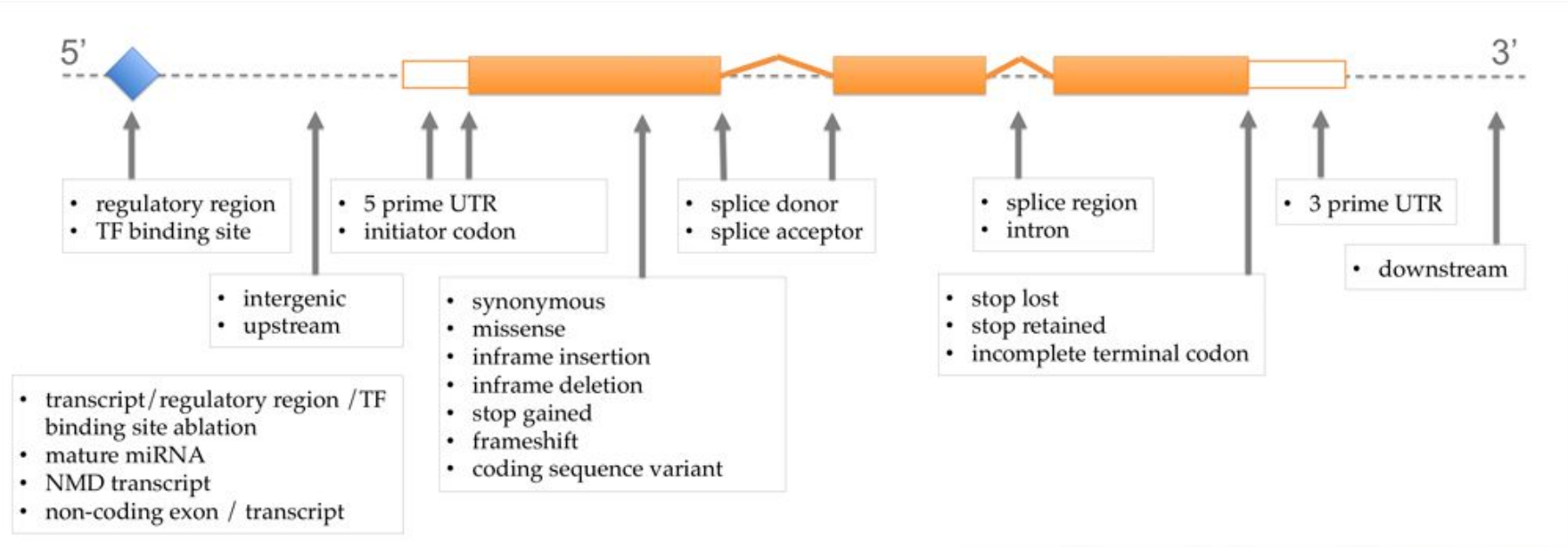
Getting information: Conservation

- **Phylop**
  - PhyloP scores measure evolutionary conservation at individual alignment sites. The scores are interpreted as follows compared to the evolution expected under neutral drift:
    - Positive scores -- Measure conservation, which is slower evolution than expected, at sites that are predicted to be conserved.
    - Negative scores -- Measure acceleration, which is faster evolution than expected, at sites that are predicted to be fast-evolving.
- **PhastCons**
  - PhastCons is a program for identifying evolutionarily conserved elements in a multiple alignment, given a phylogenetic tree.
  - PhastCons essentially does three things:
    - It produces base-by-base conservation scores (as displayed in the conservation tracks in the UCSC browser)
    - It produces predictions of discrete conserved elements (as displayed in the "most conserved" tracks in the browser)
    - It estimates free parameters.

Getting information: Effect

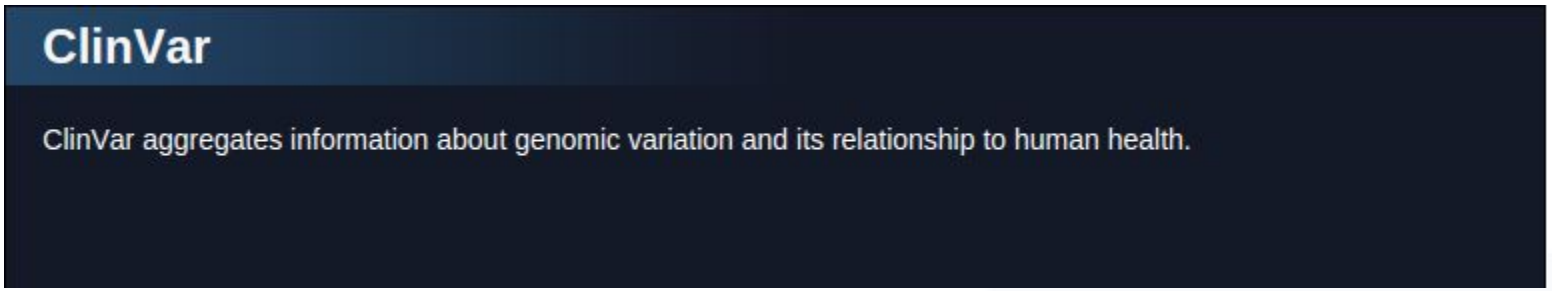


Getting information: Effect



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

Getting information: Phenotype



COSMIC
Catalogue of somatic mutations in cancer

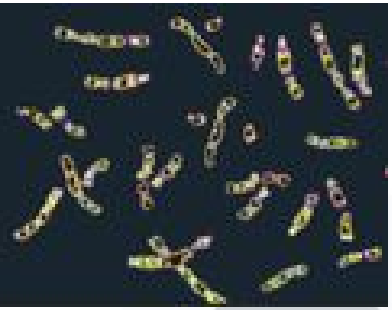
GWAS Catalog

The NHGRI-EBI Catalog of published genome-wide association studies

Getting information: Pop. Frequencies

1000 Genomes

A Deep Catalog of Human Genetic Variation

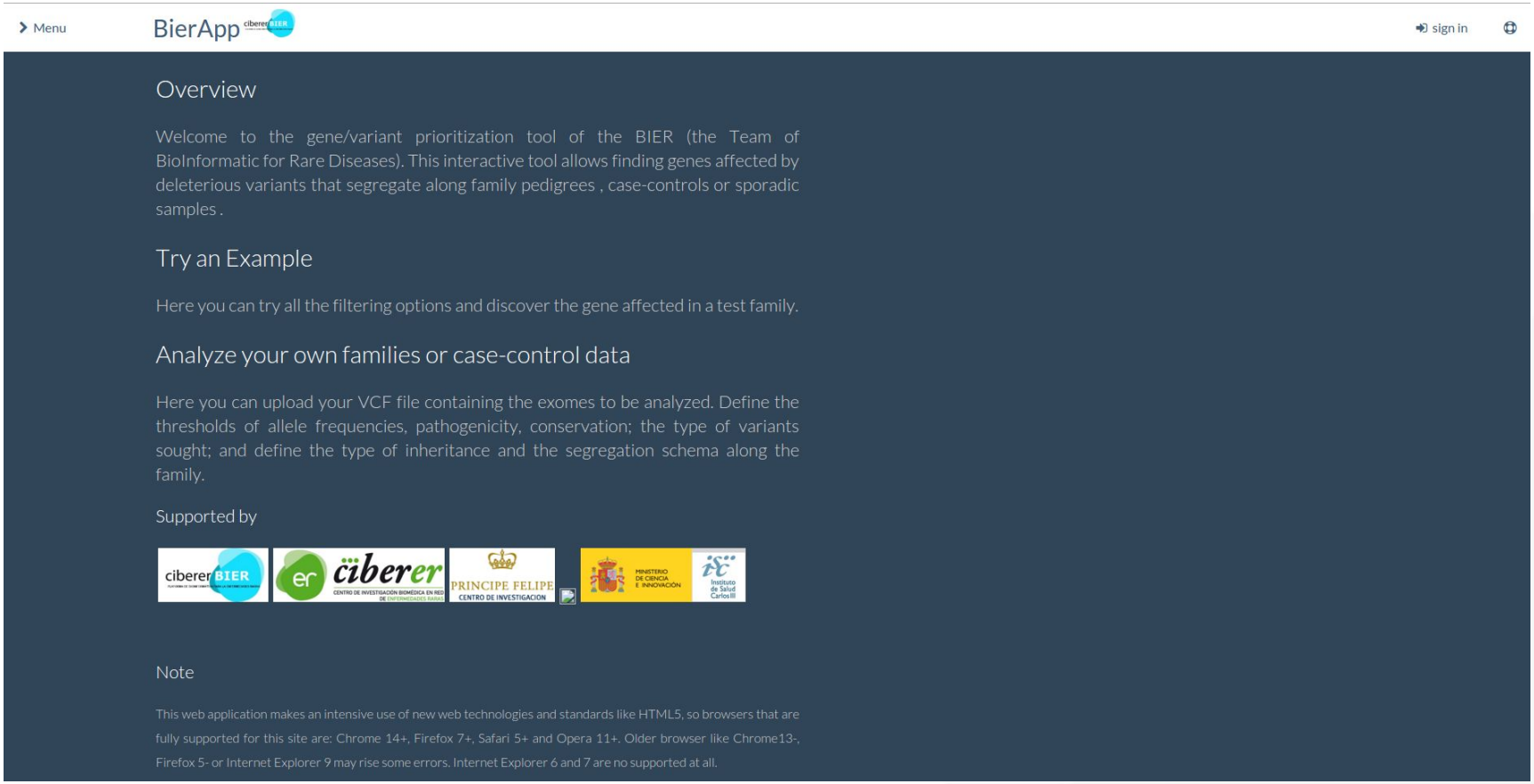


NHLBI Exome Sequencing Project (ESP)

Exome Variant Server

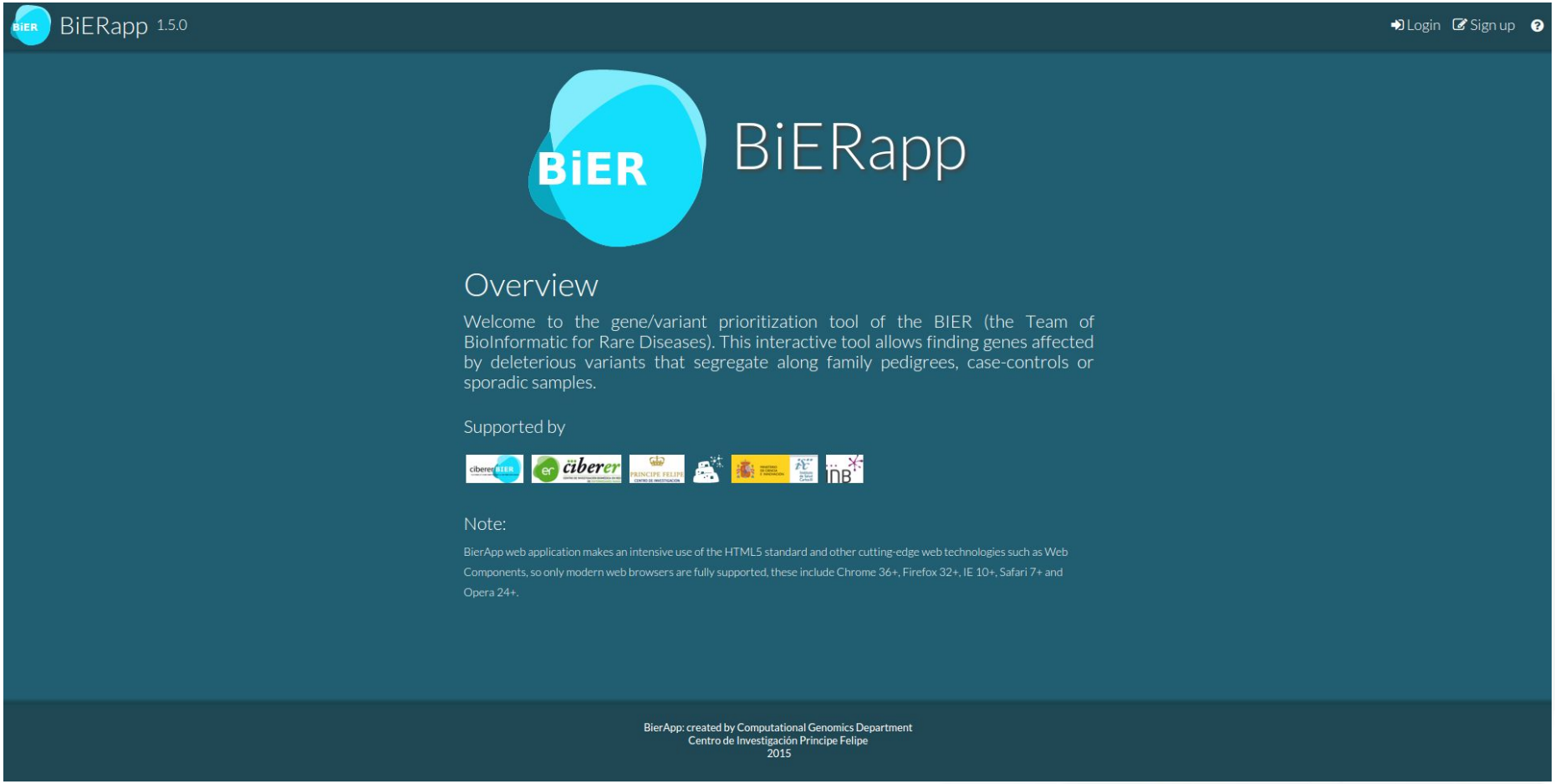
Tool interface: Official release

http://bierapp.babelomics.org/

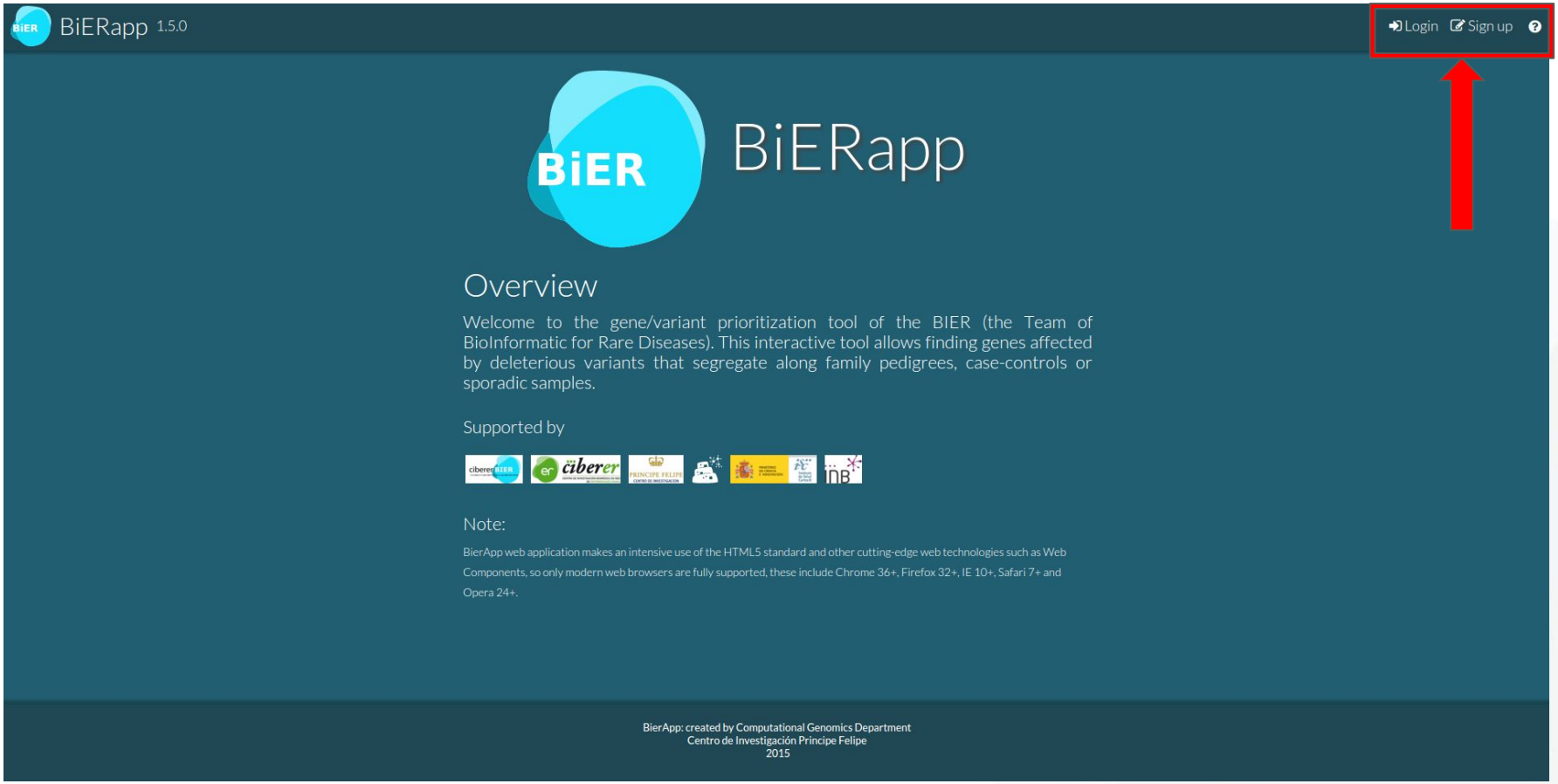


Tool interface: Beta

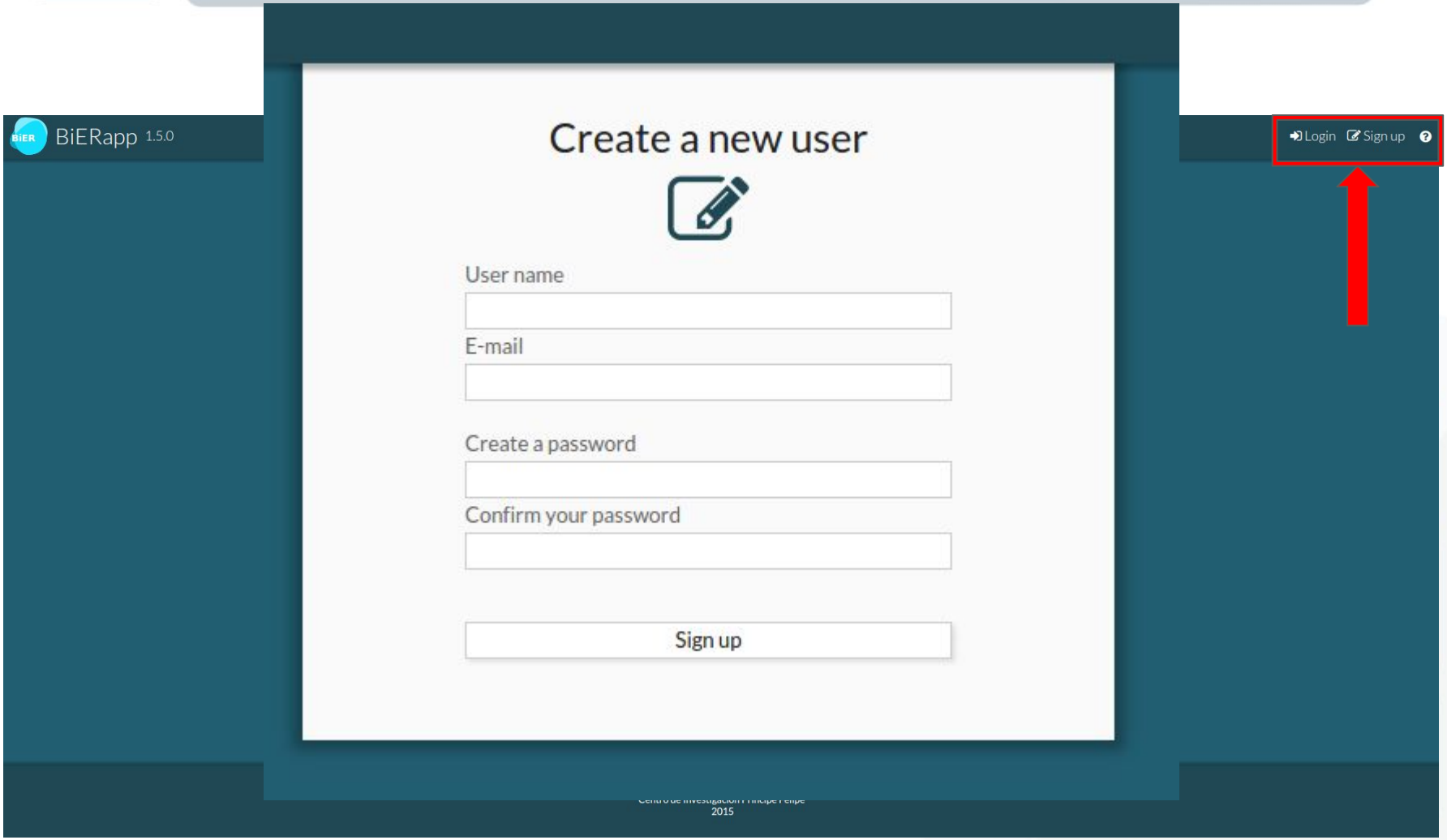
http://bierapp.babelomics.org/beta



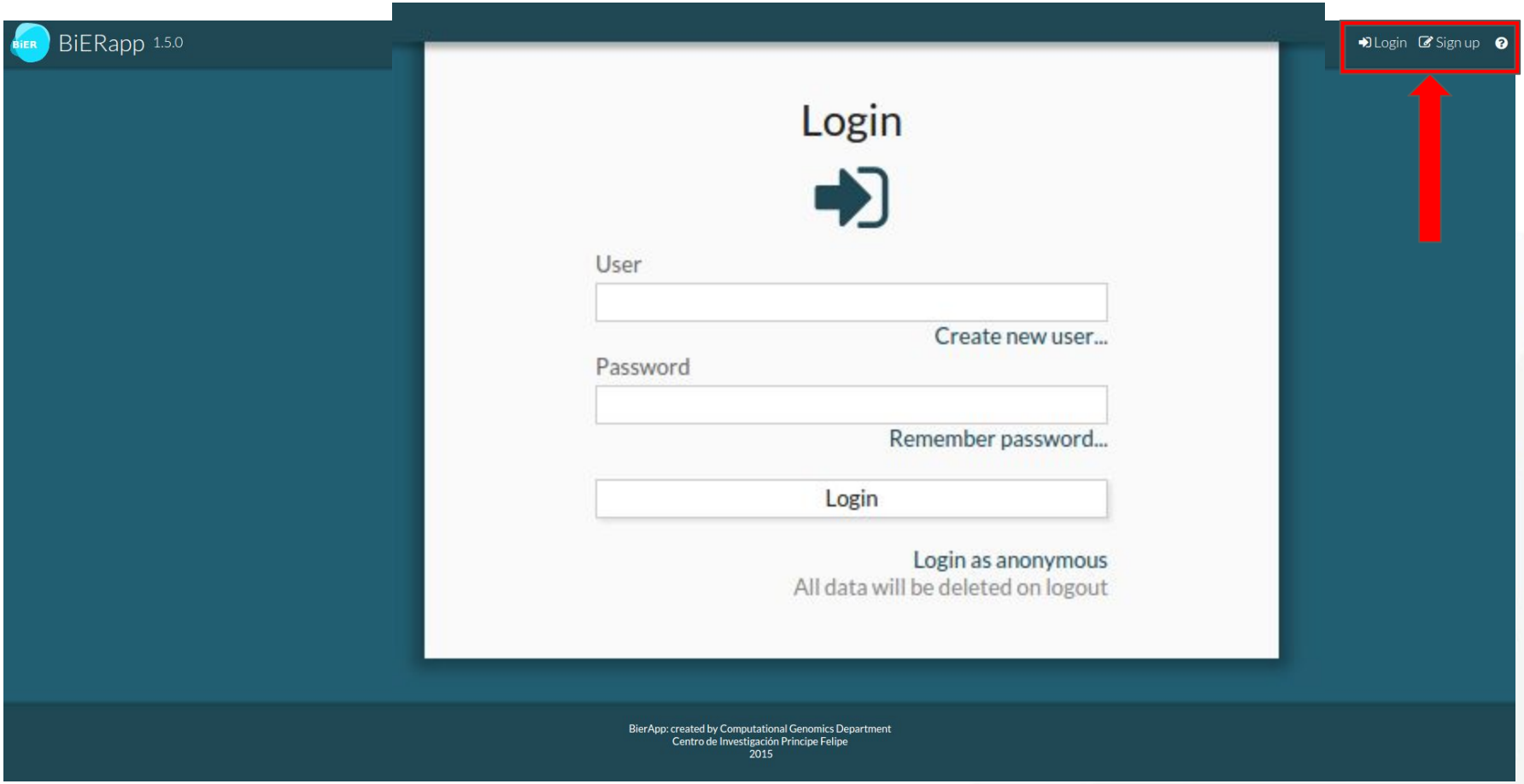
Tool interface: Sign up /Log in



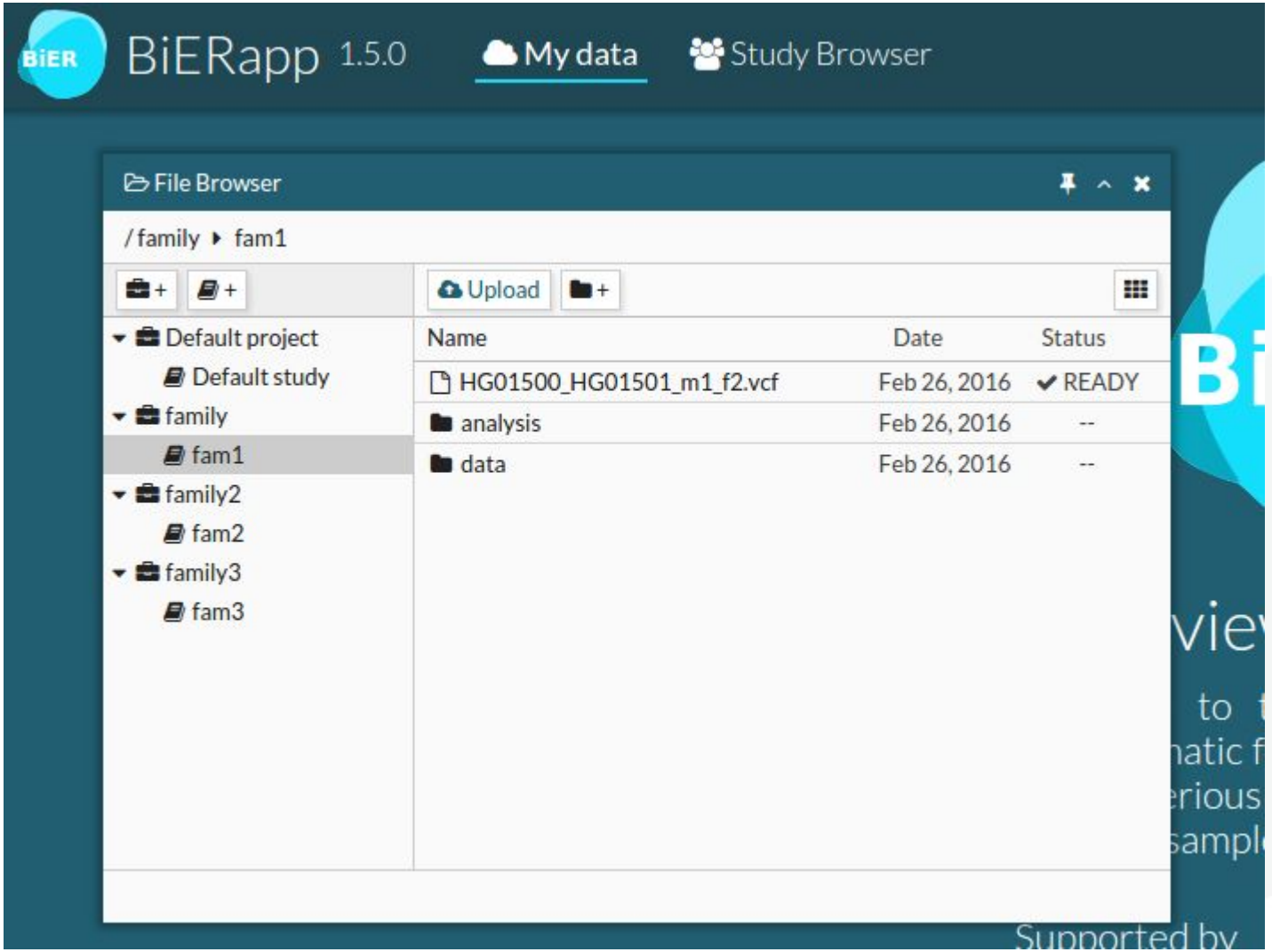
Tool interface: Sign up /Log in



Tool interface: Sign up /Log in

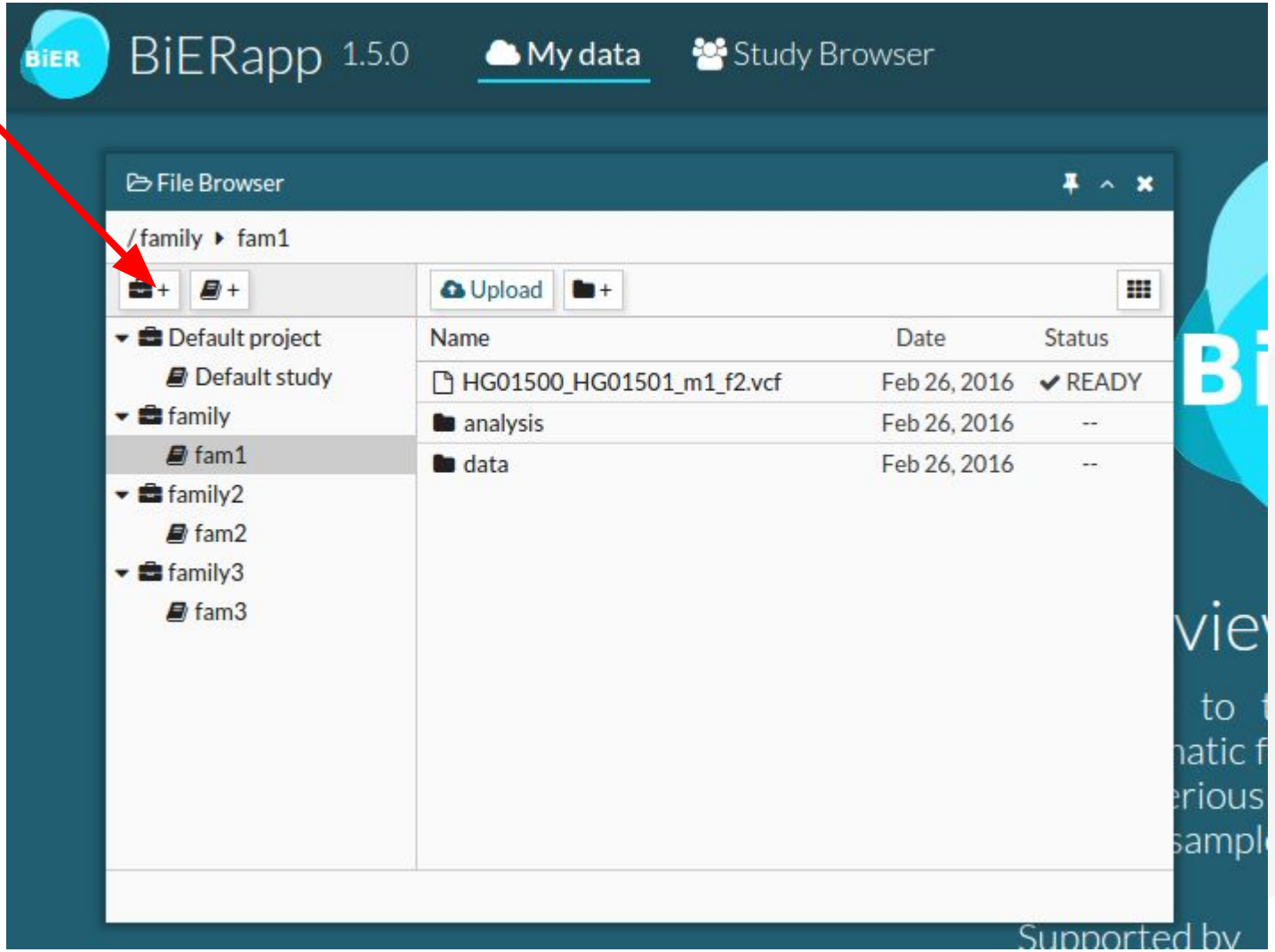


Tool interface: Project/Study



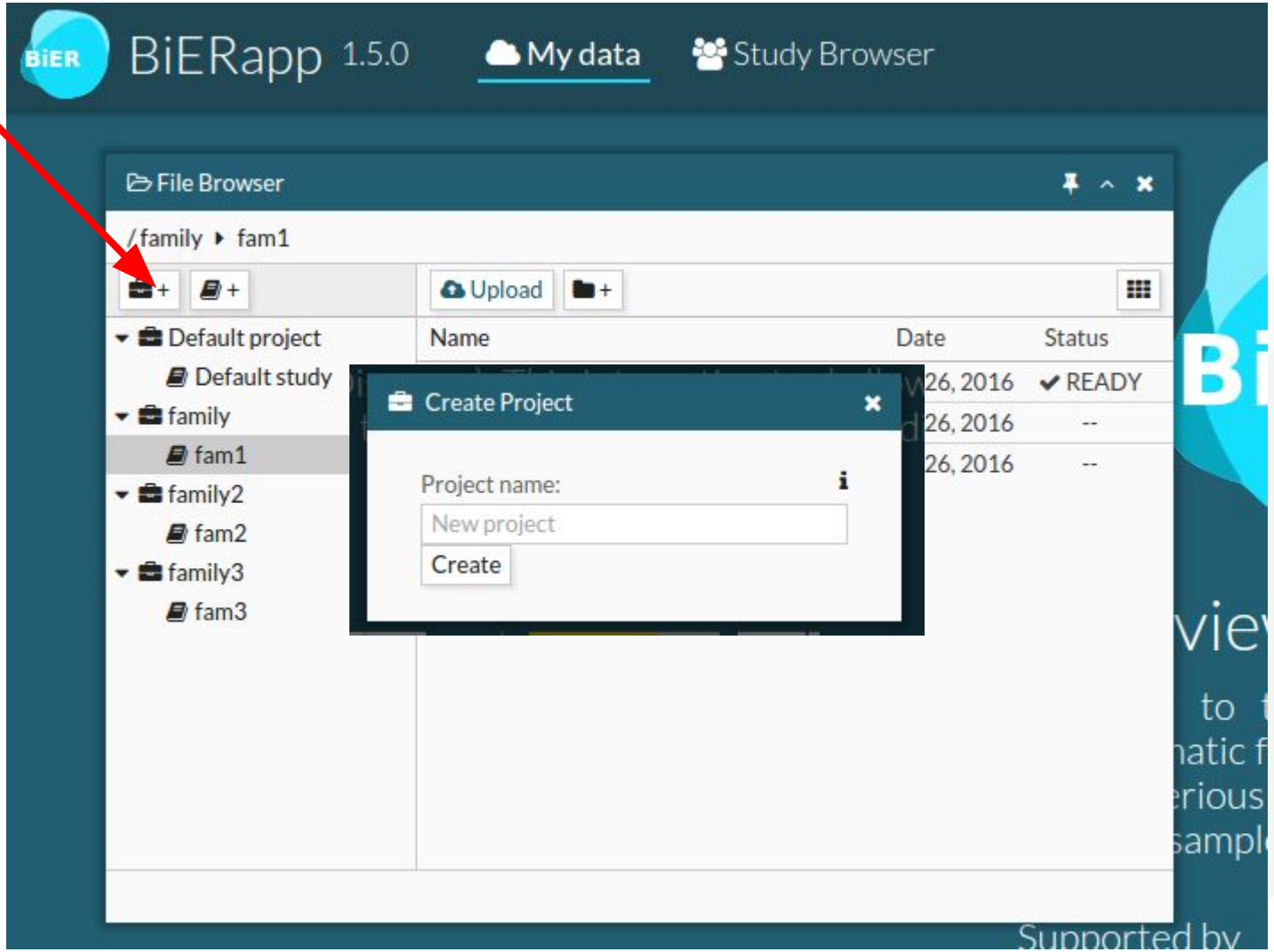
Tool interface: Project/Study

New Project



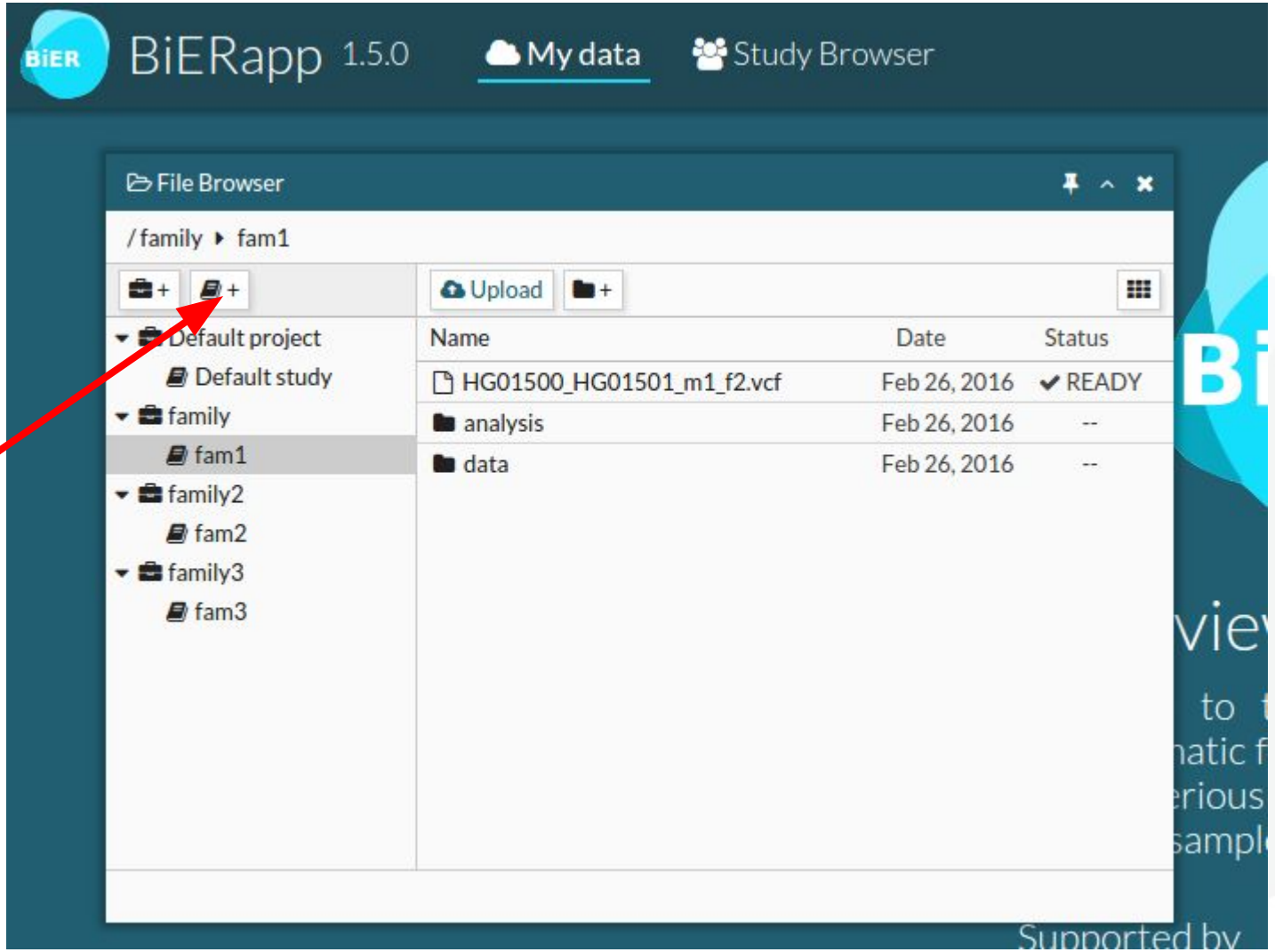
Tool interface: Project/Study

New Project



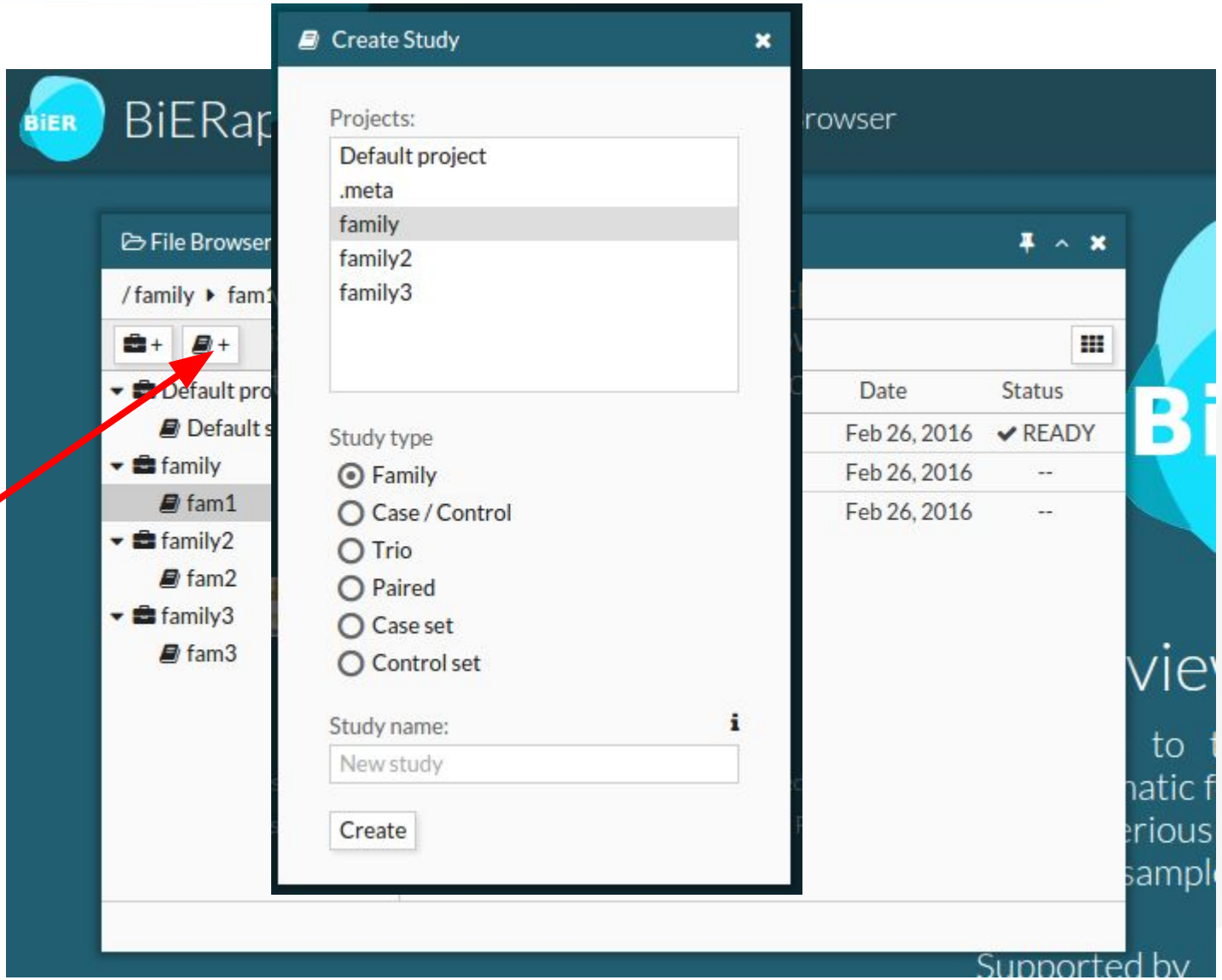
Tool interface: Project/Study

New Study

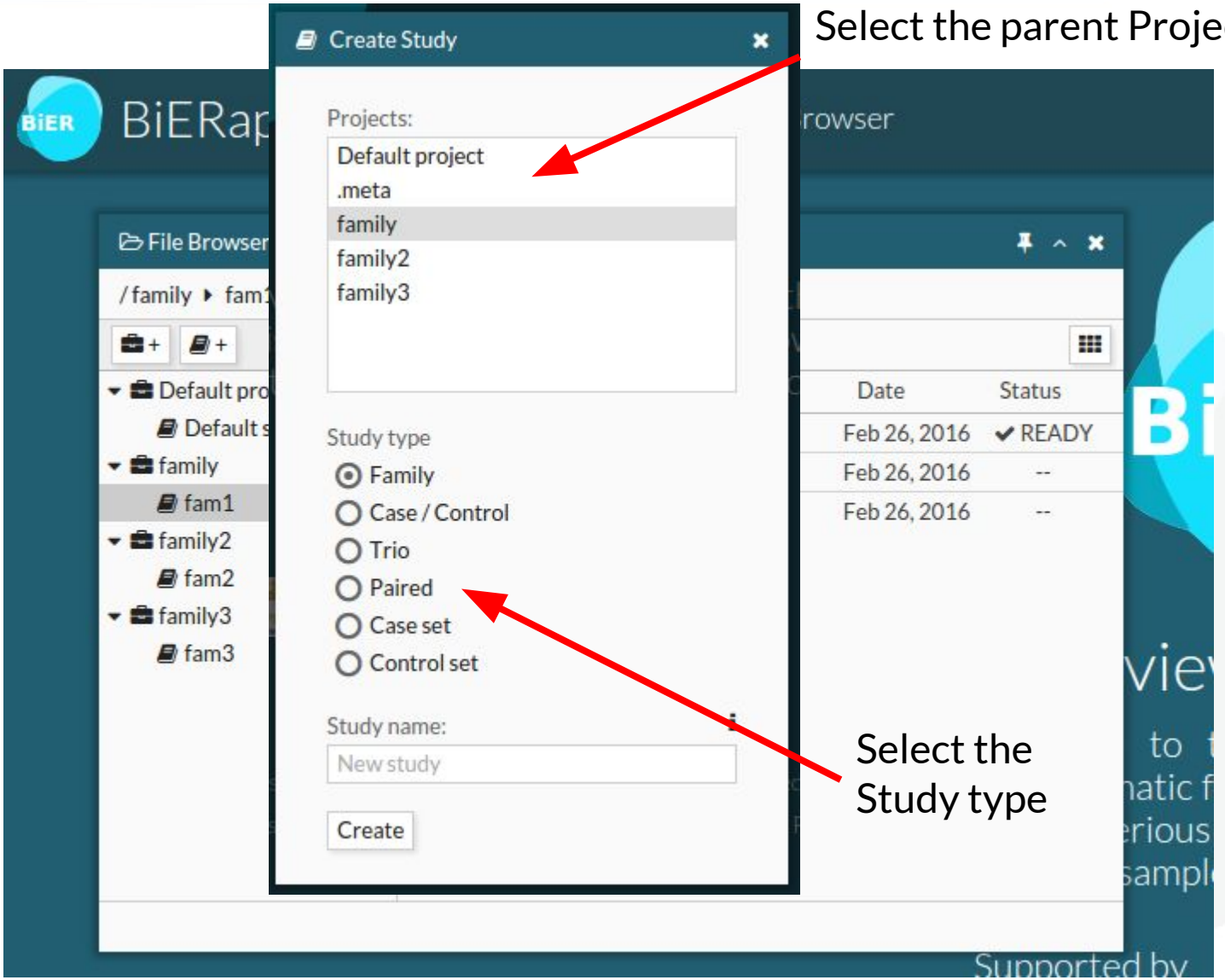


Tool interface: Project/Study

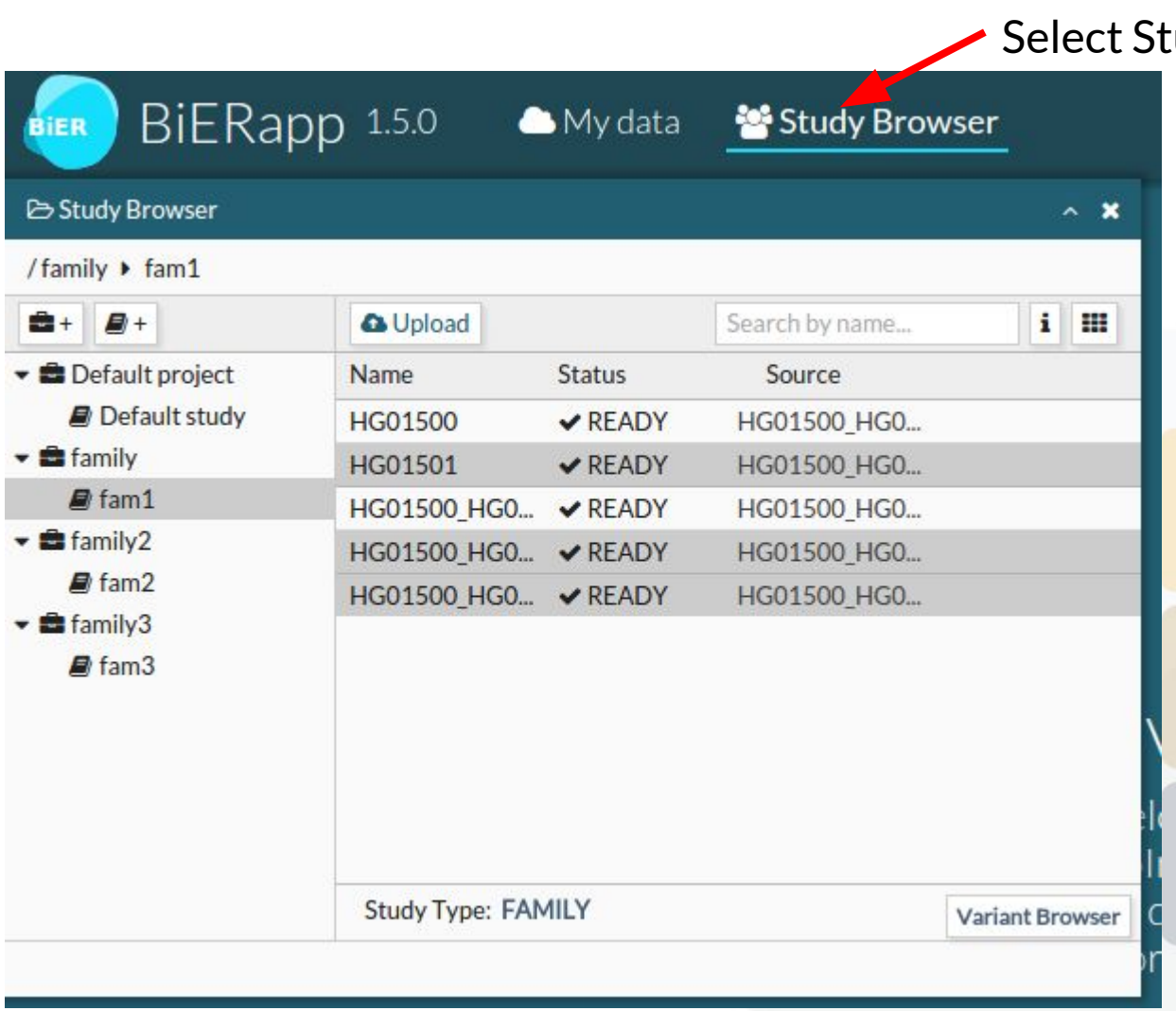
New Study



Tool interface: Project/Study



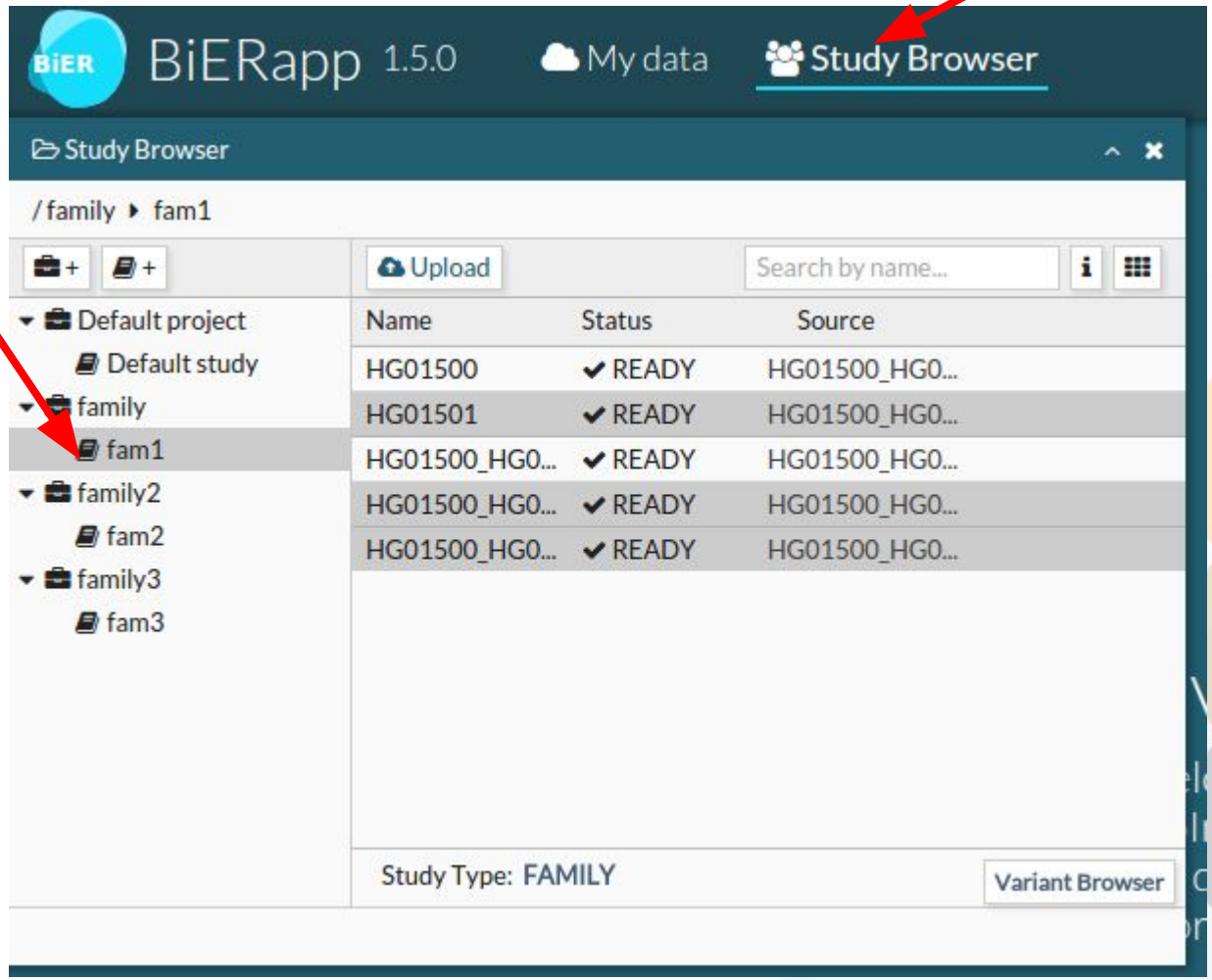
Tool interface: Study Browser



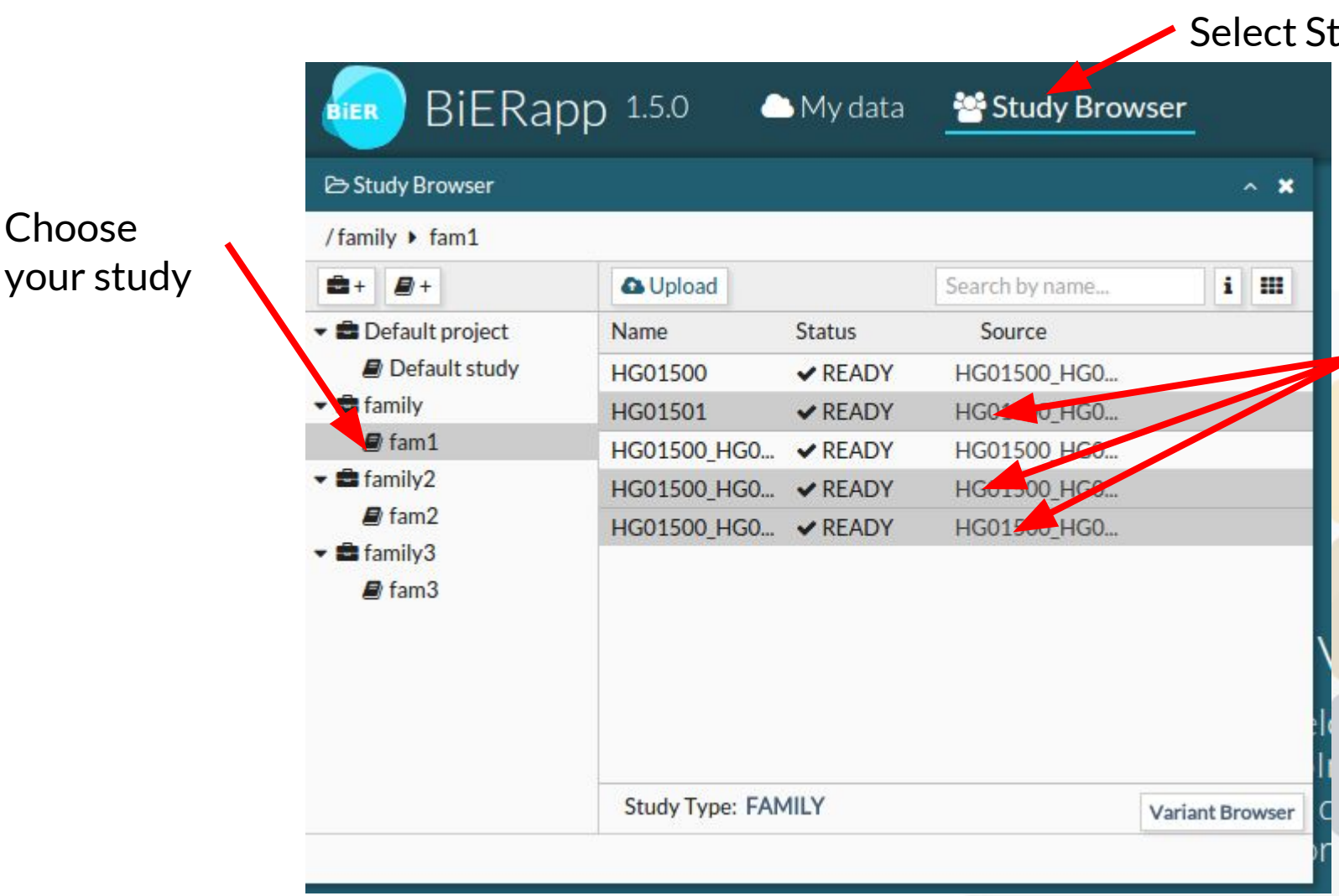
Tool interface: Study Browser

Select Study Browser

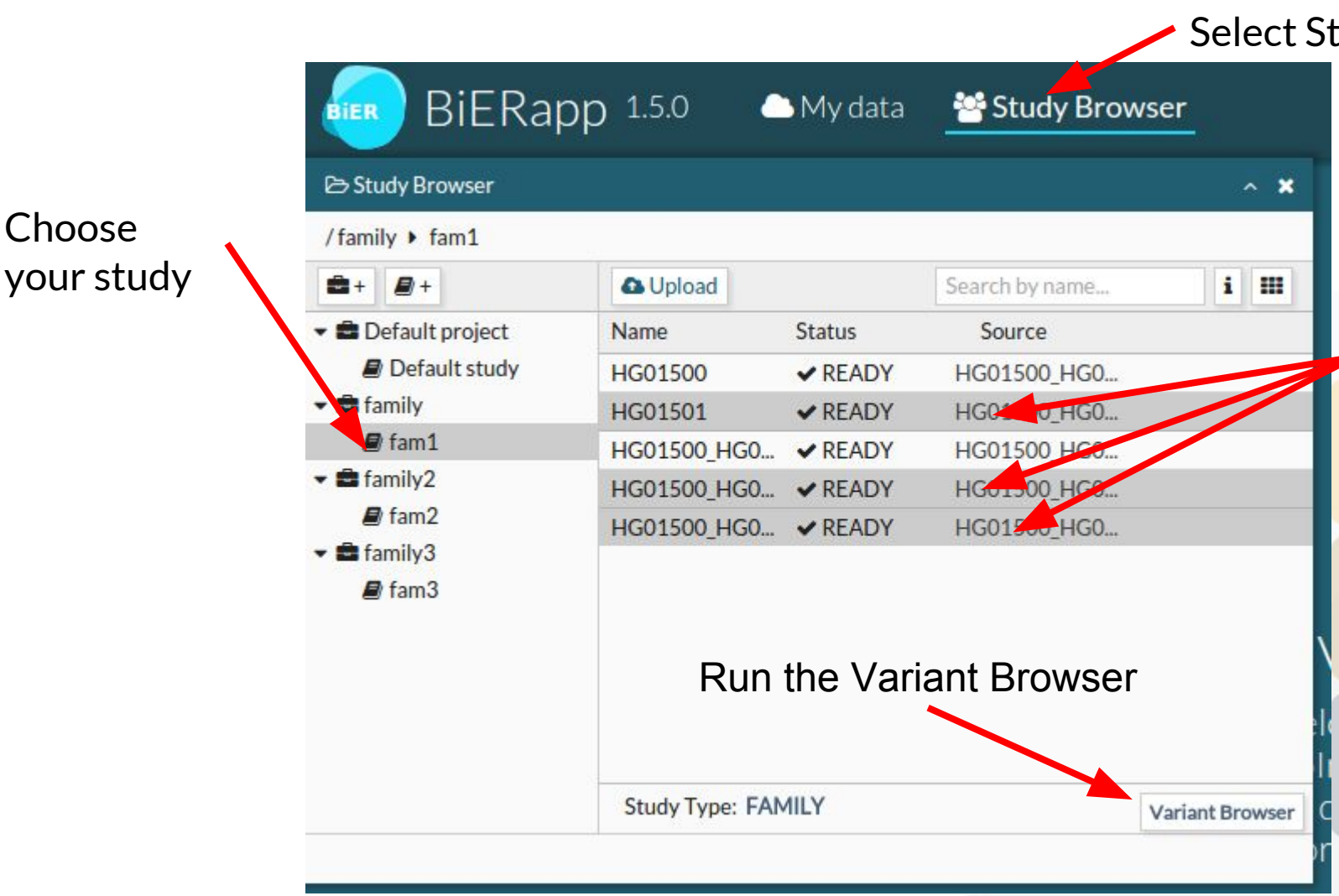
Choose your study



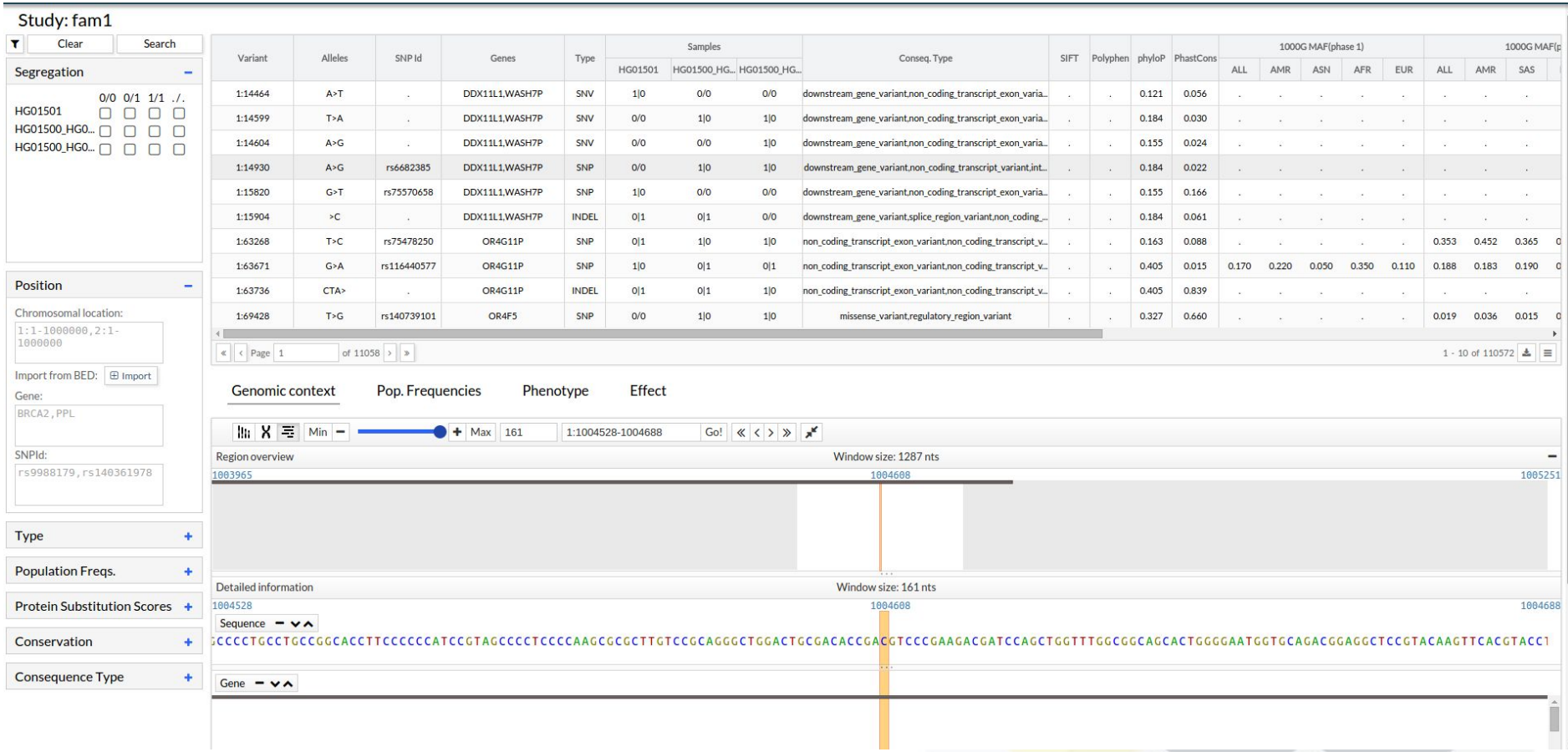
Tool interface: Study Browser



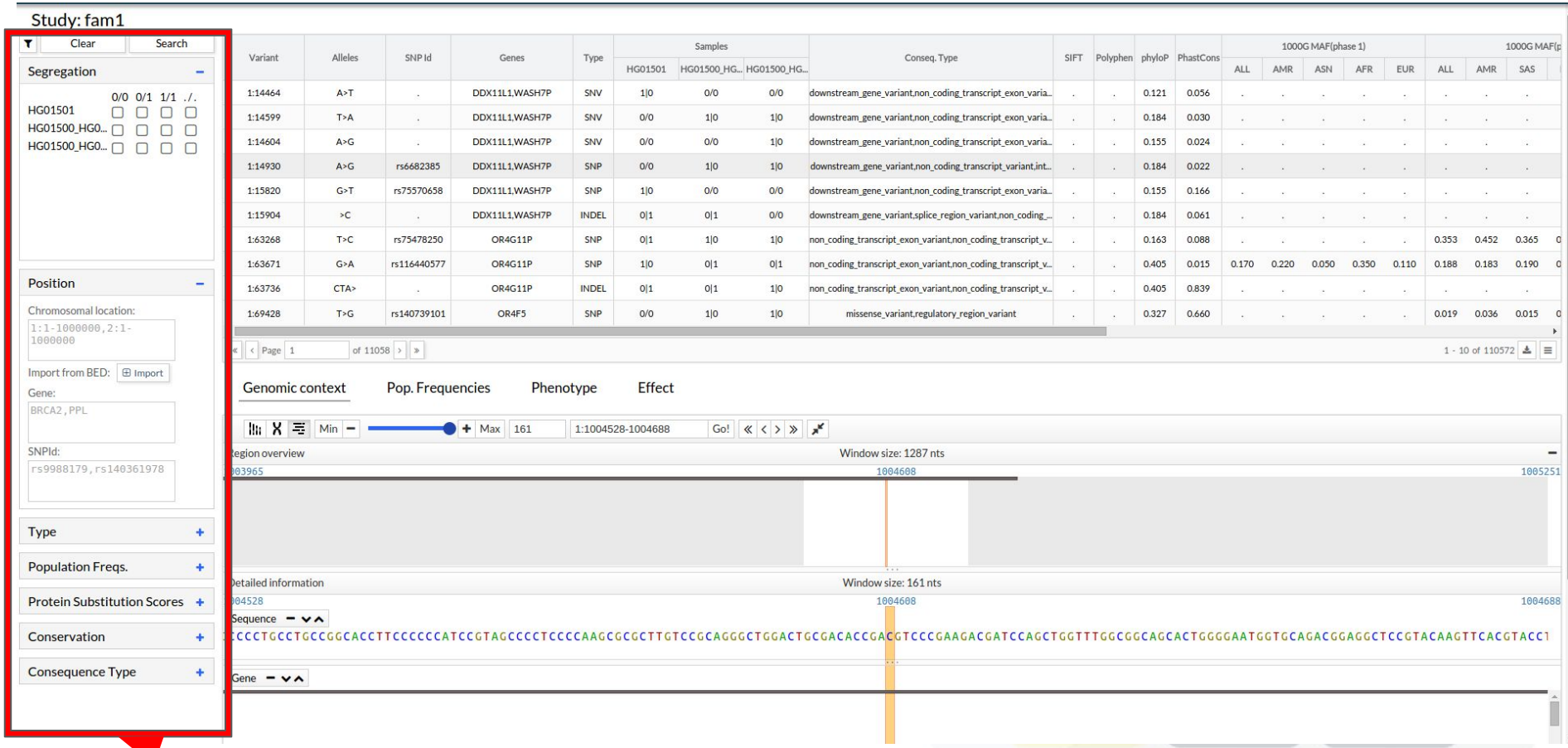
Tool interface: Study Browser



Tool interface: Variant Browser



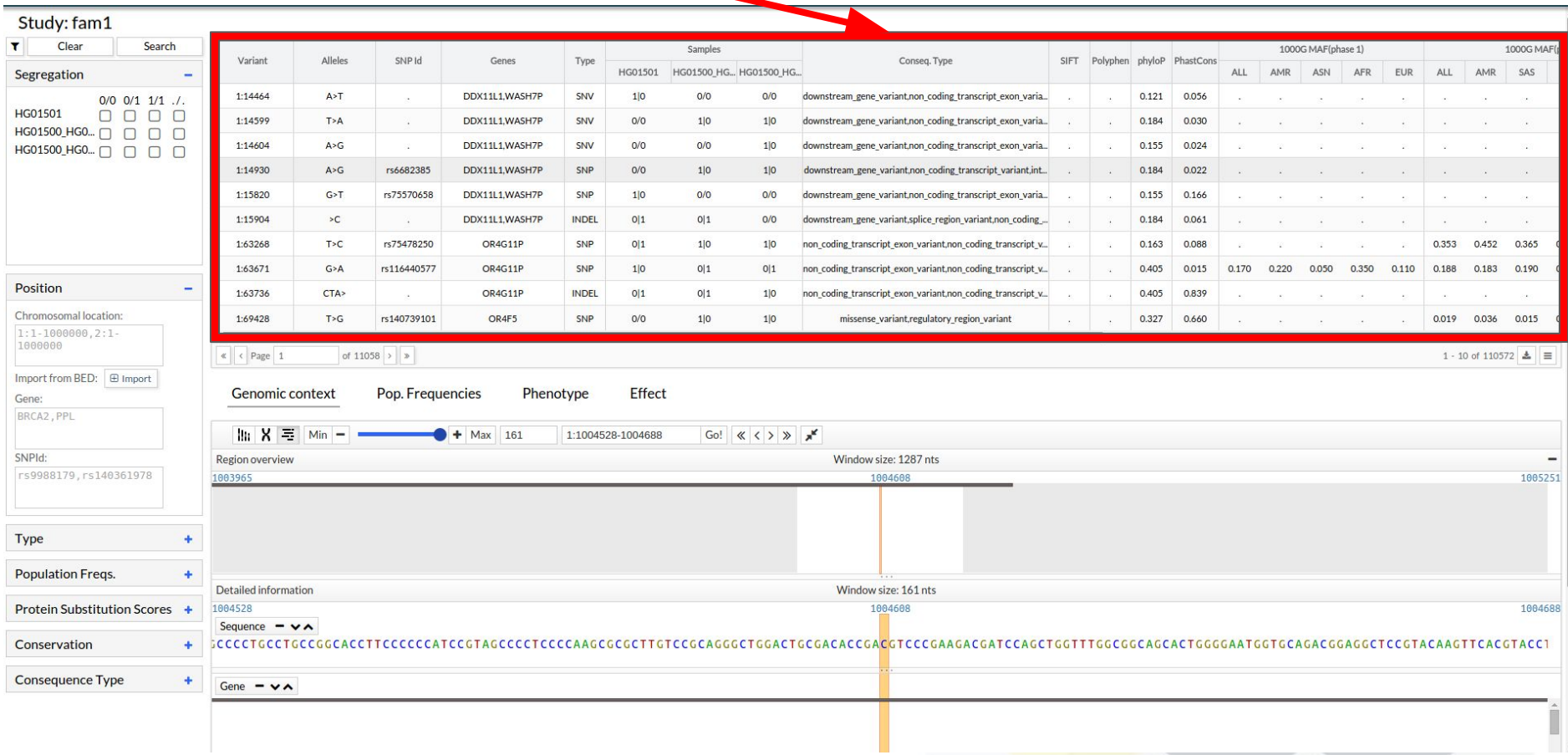
Tool interface: Variant Browser



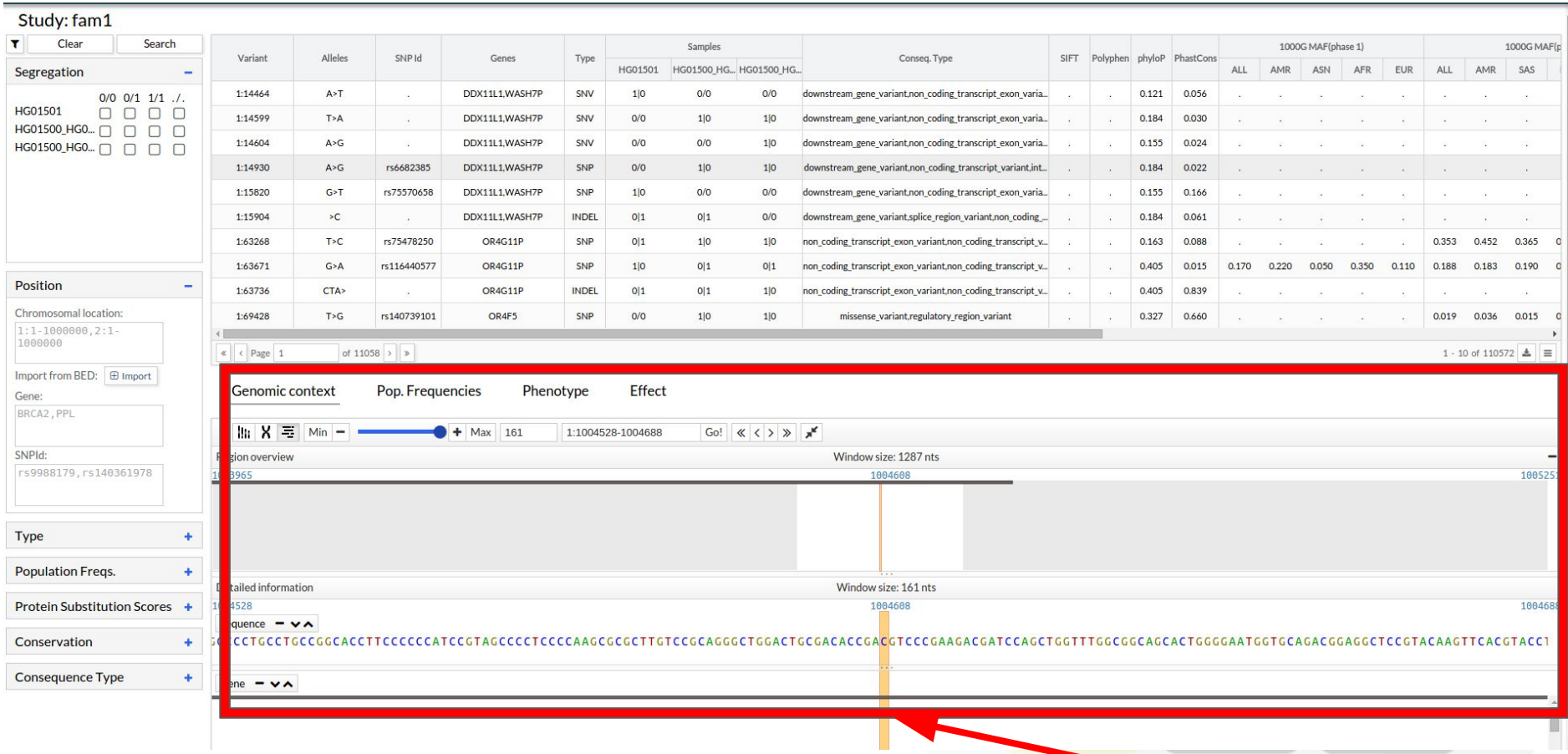
Filters

Tool interface: Variant Browser

Variants

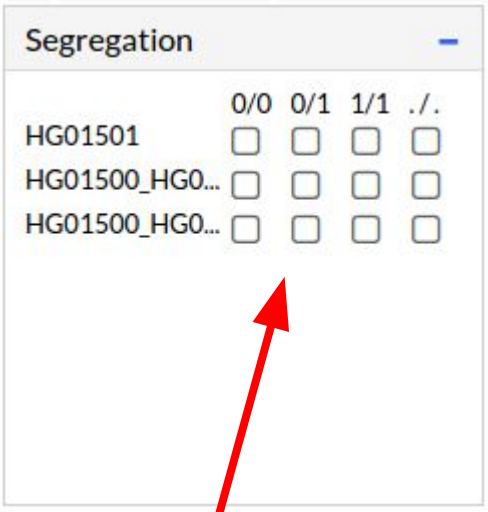


Tool interface: Variant Browser

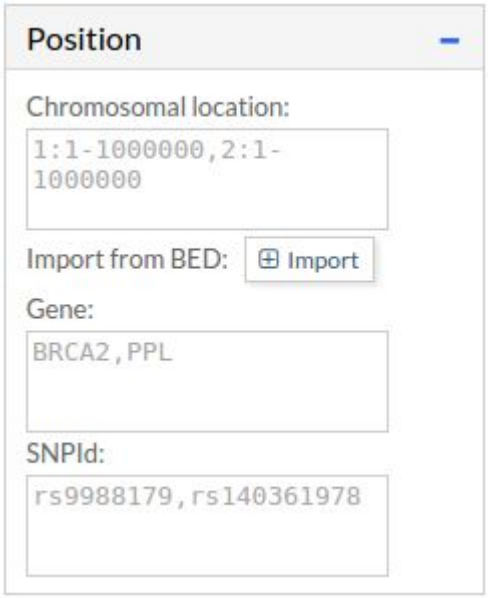


More info

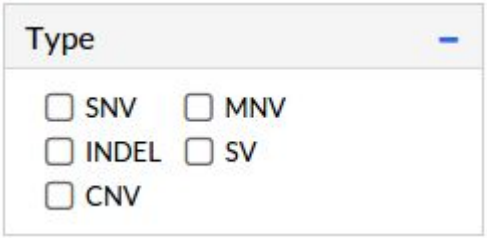
Tool interface: Filters



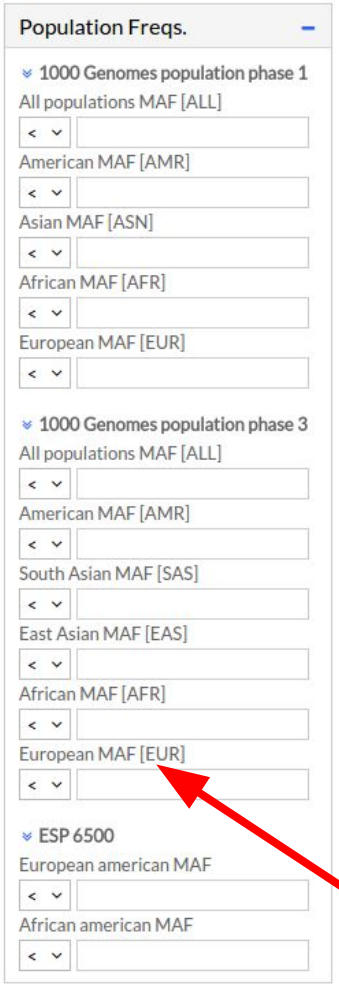
Choose your genotypes.

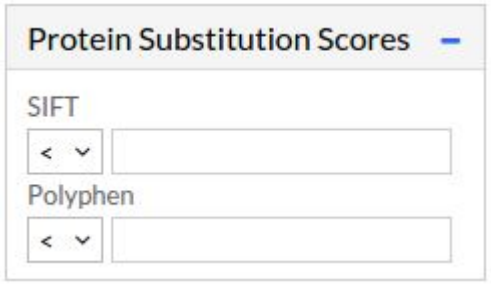


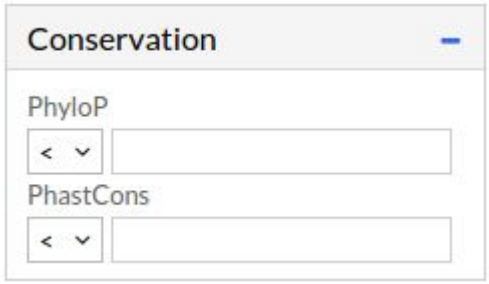
You can filter by region, Chromosome, Gene and SNPid.
You can also import regions from a BED file

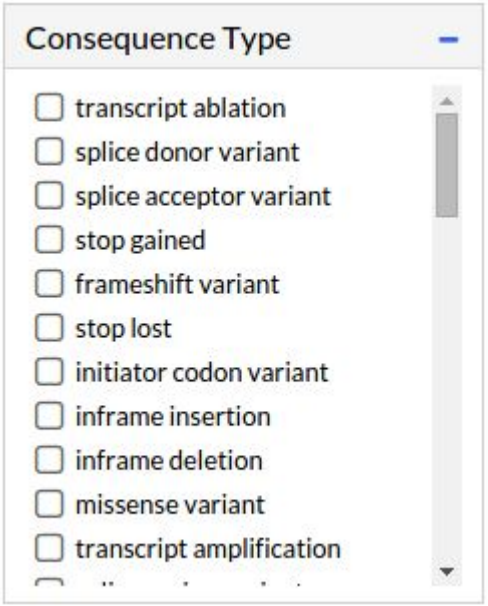


Tool interface: Filters









Filter by MAF

- 1000G phase1
- 1000G phase3
- ESP 6500

Tool interface: Variant grid

Resizable columns

Variant	Alleles	SNP Id	Genes	Type	Samples			Conseq. Type	SIFT	Polyphen	phyloP	PhastCons	1000G MAF(phase 1)					1000G MAF(p			
					HG01501	HG01500_HG...	HG01500_HG...						ALL	AMR	ASN	AFR	EUR	ALL	AMR	SAS	
1:14464	A>T	.	DDX11L1,WASH7P	SNV	1|0	0/0	0/0	downstream_gene_variant,non_coding_transcript_exon_varia...	.	.	0.121	0.056	.	.	.	.	.	.	.	.	
1:14599	T>A	.	DDX11L1,WASH7P	SNV	0/0	1|0	1|0	downstream_gene_variant,non_coding_transcript_exon_varia...	.	.	0.184	0.030	.	.	.	.	.	.	.	.	
1:14604	A>G	.	DDX11L1,WASH7P	SNV	0/0	0/0	1|0	downstream_gene_variant,non_coding_transcript_exon_varia...	.	.	0.155	0.024	.	.	.	.	.	.	.	.	
1:14930	A>G	rs6682385	DDX11L1,WASH7P	SNP	0/0	1|0	1|0	downstream_gene_variant,non_coding_transcript_variant,int...	.	.	0.184	0.022	.	.	.	.	.	.	.	.	
1:15820	G>T	rs75570658	DDX11L1,WASH7P	SNP	1|0	0/0	0/0	downstream_gene_variant,non_coding_transcript_exon_varia...	.	.	0.155	0.166	.	.	.	.	.	.	.	.	
1:15904	>C	.	DDX11L1,WASH7P	INDEL	0|1	0|1	0/0	downstream_gene_variant,splice_region_variant,non_coding_...	.	.	0.184	0.061	.	.	.	.	.	.	.	.	
1:63268	T>C	rs75478250	OR4G11P	SNP	0|1	1|0	1|0	non_coding_transcript_exon_variant,non_coding_transcript_v...	.	.	0.163	0.088	.	.	.	.	.	0.353	0.452	0.365	0
1:63671	G>A	rs116440577	OR4G11P	SNP	1|0	0|1	0|1	non_coding_transcript_exon_variant,non_coding_transcript_v...	.	.	0.405	0.015	0.170	0.220	0.050	0.350	0.110	0.188	0.183	0.190	0
1:63736	CTA>	.	OR4G11P	INDEL	0|1	0|1	1|0	non_coding_transcript_exon_variant,non_coding_transcript_v...	.	.	0.405	0.839	.	.	.	.	.	.	.	.	
1:69428	T>G	rs140739101	OR4F5	SNP	0/0	1|0	1|0	missense_variant,regulatory_region_variant	.	.	0.327	0.660	.	.	.	.	.	0.019	0.036	0.015	0

« ‹ Page 1 of 11058 › »

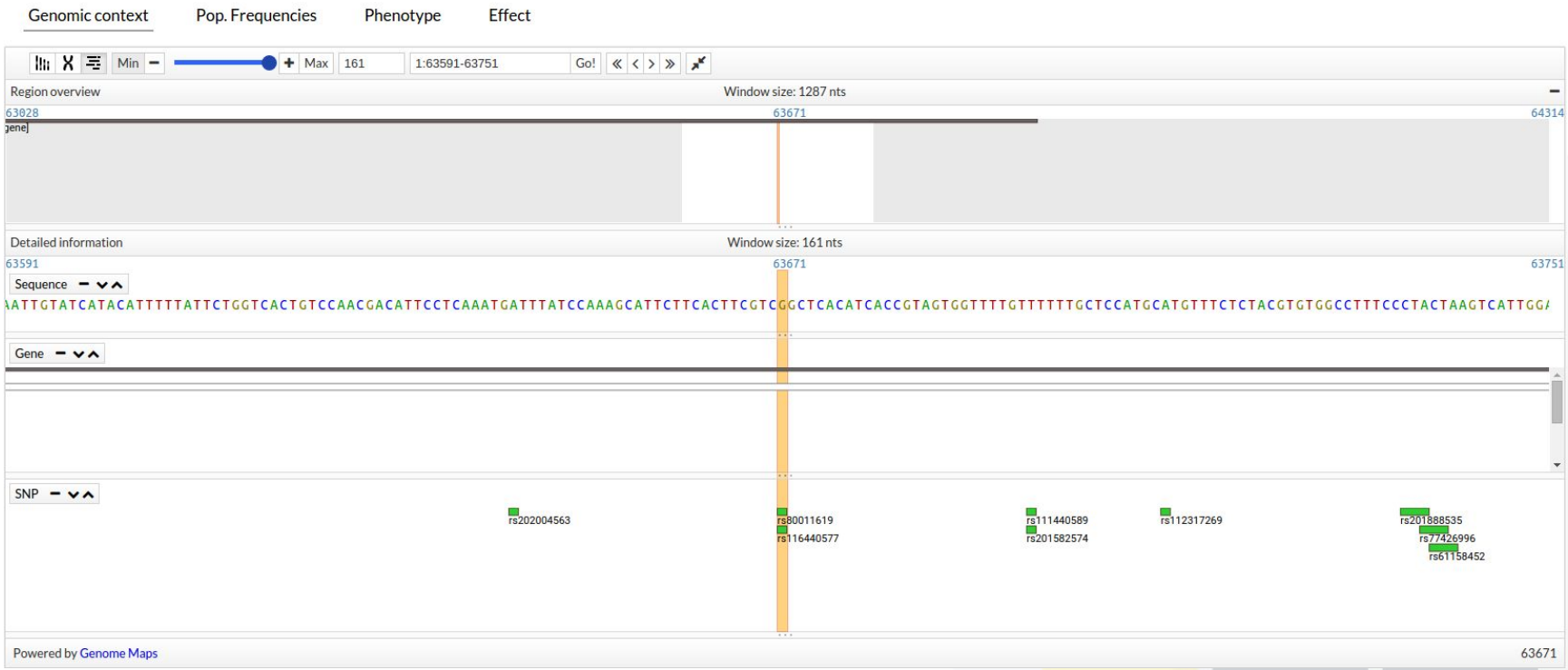
1 - 10 of 110572

Pagination system

Export to CSV

Column Selector

Tool interface: Genomic Context

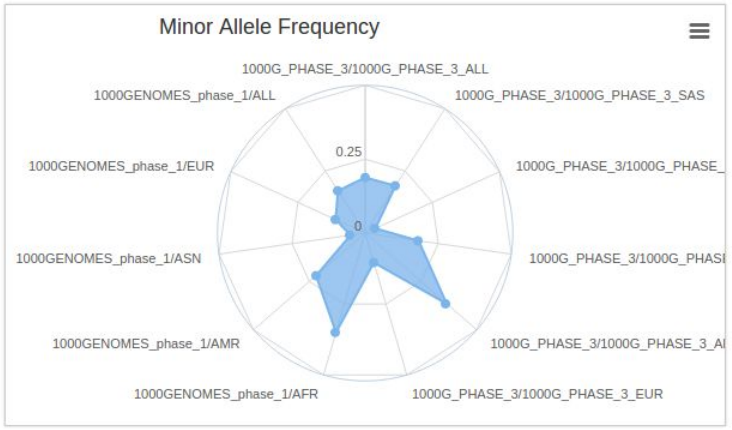


Tool interface: Pop. Frequencies

Genomic context | Pop. Frequencies | Phenotype | Effect

Study	Population	SuperPopulation	Ref. Allele	Alt. Allele	Ref. Allele Fr...	Alt. Allele Fr...	MAF	0/0	0/1	
1000G_PHASE_3	1000G_PHASE_3_ALL	1000G_PHASE_3_ALL	G	A	0.813	0.188	0.188	0	0	0
1000G_PHASE_3	1000G_PHASE_3_SAS	1000G_PHASE_3_SAS	G	A	0.810	0.190	0.190	0	0	0
1000G_PHASE_3	1000G_PHASE_3_EAS	1000G_PHASE_3_EAS	G	A	0.962	0.038	0.038	0	0	0
1000G_PHASE_3	1000G_PHASE_3_AMR	1000G_PHASE_3_AMR	G	A	0.817	0.183	0.183	0	0	0
1000G_PHASE_3	1000G_PHASE_3_AFR	1000G_PHASE_3_AFR	G	A	0.635	0.365	0.365	0	0	0
1000G_PHASE_3	1000G_PHASE_3_EUR	1000G_PHASE_3_EUR	G	A	0.896	0.104	0.104	0	0	0
1000GENOMES_phase_1	AFR	AFR	G	A	0.650	0.350	0.350	0	0	0
1000GENOMES_phase_1	AMR	AMR	G	A	0.780	0.220	0.220	0	0	0
1000GENOMES_phase_1	ASN	ASN	G	A	0.950	0.050	0.050	0	0	0
1000GENOMES_phase_1	EUR	EUR	G	A	0.890	0.110	0.110	0	0	0
1000GENOMES_phase_1	ALL	ALL	G	A	0.830	0.170	0.170	0	0	0

Total: 11



Tool interface: Phenotype & Effect

Genomic context | Pop. Frequencies | Phenotype | Effect

Cosmic:

Gene name	Histology subtype	Mutation ID	Mutation somatic status	Primary histology	Primary site	Sample source	Site subtype	Tumour orig
AGRN	adenocarcinoma	1126908	Confirmed somatic variant	carcinoma	prostate	fresh/frozen - NOS	NS	primary
AGRN	adenocarcinoma	1126908	Confirmed somatic variant	carcinoma	large_intestine	NS	colon	NS
AGRN	neoplasm	1126908	Confirmed somatic variant	other	thyroid	NS	NS	NS
AGRN	neoplasm	1126908	Confirmed somatic variant	other	thyroid	NS	NS	NS
AGRN	adenocarcinoma	1126908	Confirmed somatic variant	carcinoma	large_intestine	NS	colon	NS

Total: 5

GWAS:

No results found.

Clinvar:

Accession	Clinical significance	Gene name	Review status	Traits
RCV000116259	Benign	AGRN	CLASSIFIED_BY_SINGLE_SUBMITTER	not specified,AllHighlyPenetrant,Not Specified

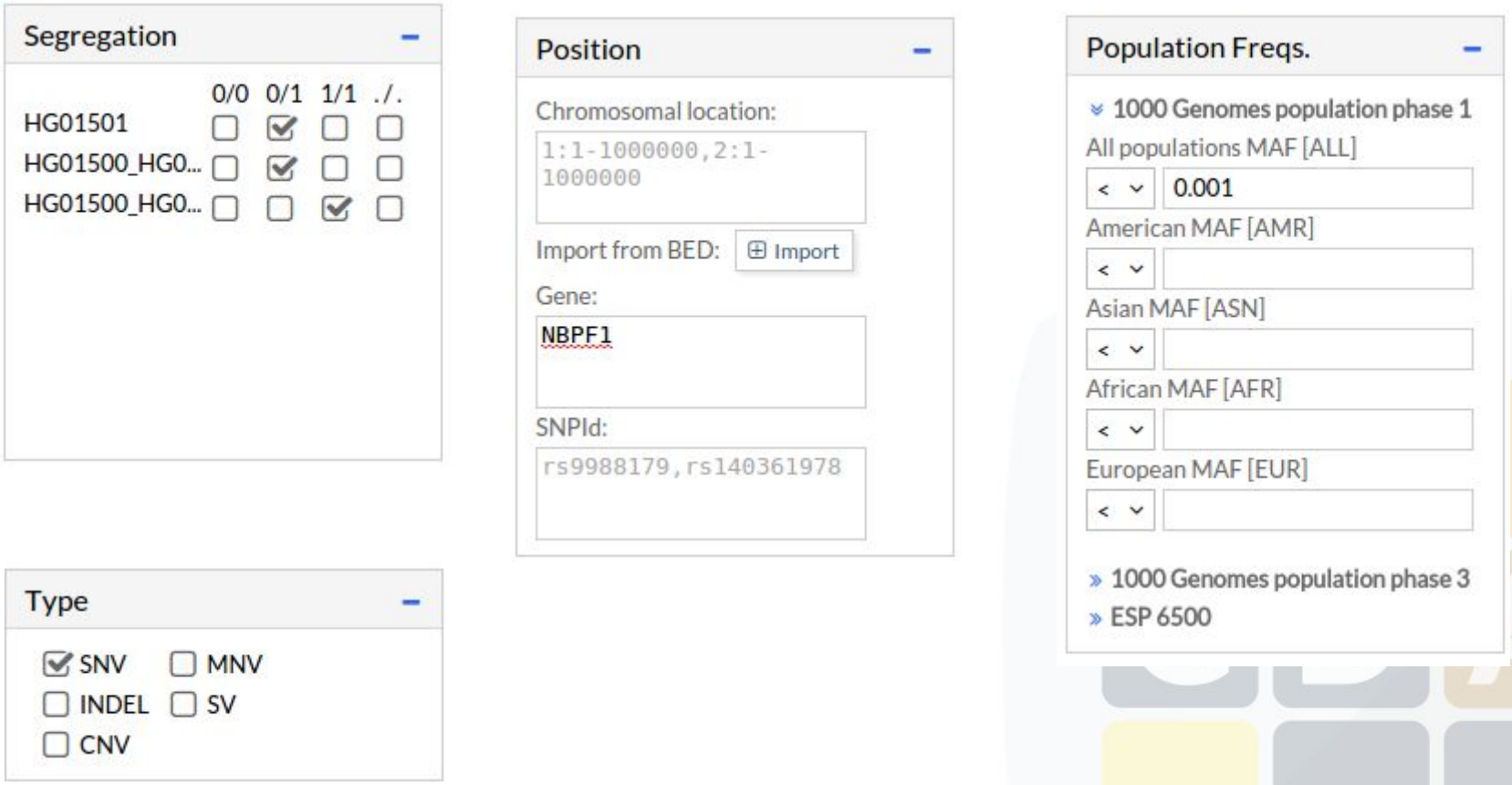
Total: 1

Genomic context | Pop. Frequencies | Phenotype | Effect

Gene Name	Ensembl Gene Id	Ensembl Transcript Id	Conseq. type	Relative Position	Codon	Strand	Biotype	cDna Position	cds Position	AA Position	AA Change	Sift	Polyphen
AGRN	ENSG00000188157	ENST00000379370	synonymous_variant		tcA/tcG	+	protein_coding	3116	3066	1022	SER/SER		
AGRN	ENSG00000188157	ENST00000479707	2KB_downstream_gene_variant			+	retained_intron						
AGRN	ENSG00000188157	ENST00000466223	2KB_upstream_gene_variant			+	retained_intron						
AGRN	ENSG00000188157	ENST00000478677	2KB_upstream_gene_variant			+	retained_intron						
AGRN	ENSG00000188157	ENST00000492947	2KB_upstream_gene_variant			+	retained_intron						
AGRN	ENSG00000188157	ENST00000419249	upstream_gene_variant			+	protein_coding						
			regulatory_region_variant										

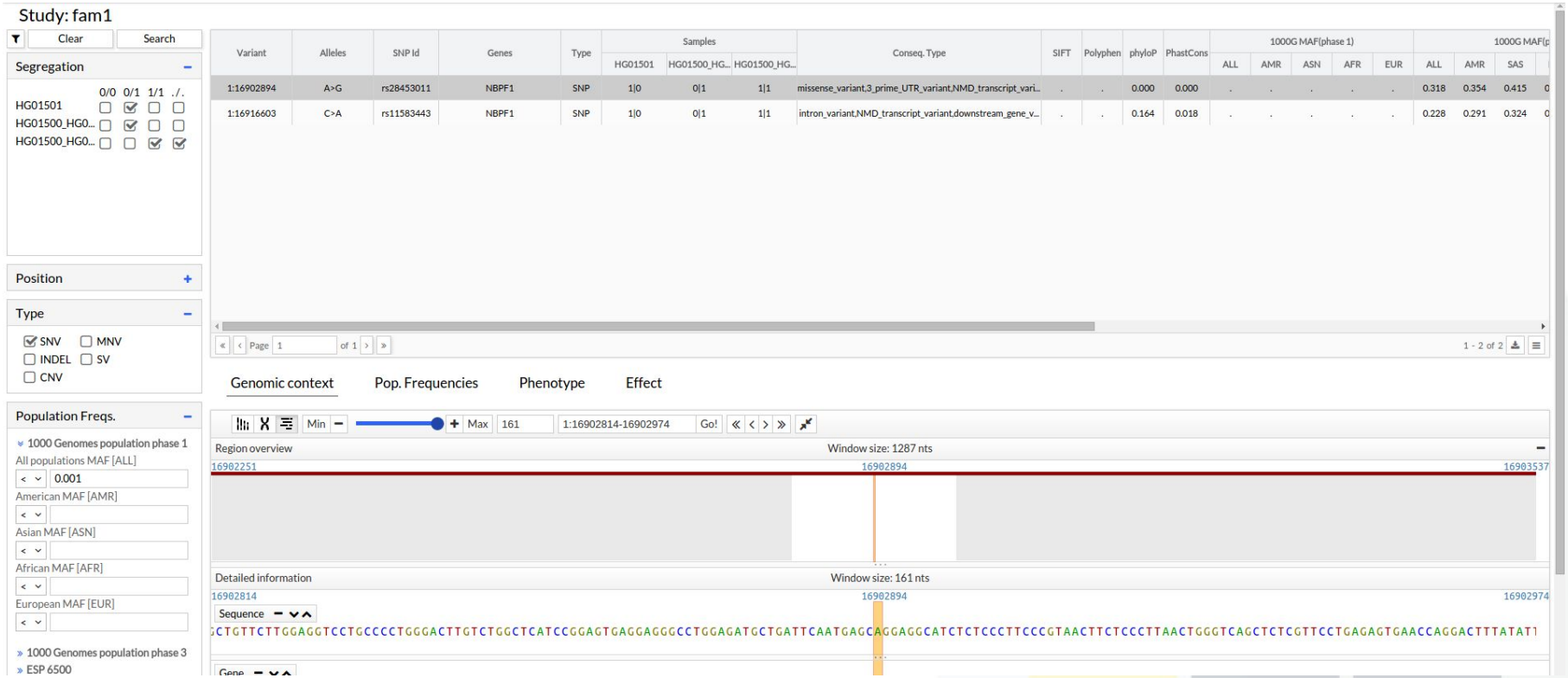
Total: 7

Results



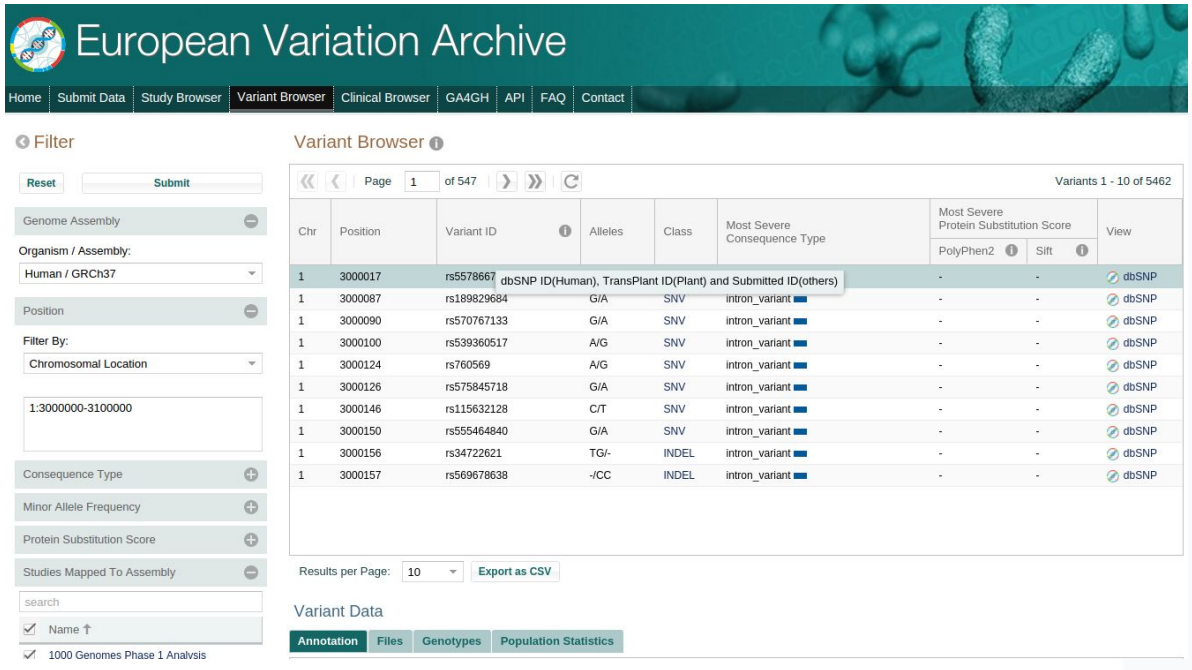
Click on “Search” and view the results

Results


Study: fam1
Clear
Search
Segregation
0/0 0/1 1/1 ./.
HG01501
HG01500_HG0...
HG01500_HG0...
Position
Type
SNV
MNV
INDEL
SV
CNV
Population Freqs.
1000 Genomes population phase 1
All populations MAF [ALL]
0.001
American MAF [AMR]
Asian MAF [ASN]
African MAF [AFR]
European MAF [EUR]
1000 Genomes population phase 3
ESP 6500
Variant
Alleles
SNP Id
Genes
Type
Samples
HG01501
HG01500_HG...
HG01500_HG...
Conseq. Type
SIFT
Polyphen
phyloP
PhastCons
1000G MAF(phase 1)
ALL
AMR
ASN
AFR
EUR
SAS
1:16902894 A>G rs28453011 NBPF1 SNP 1|0 0|1 1|1 missense_variant,3_prime_UTR_variant,NMD_transcript_vari... 0.000 0.000 0.318 0.354 0.415
1:16916603 C>A rs11583443 NBPF1 SNP 1|0 0|1 1|1 intron_variant,NMD_transcript_variant,downstream_gene_v... 0.164 0.018 0.228 0.291 0.324
Page 1 of 1
1 - 2 of 2
Genomic context
Pop. Frequencies
Phenotype
Effect
Min
Max
161
1:16902814-16902974
Go!
Region overview
Window size: 1287 nts
16902251
16902894
16903537
Detailed information
Window size: 161 nts
16902814
16902894
16902974
Sequence
Gene

Who is using BierApp?

ciberer isciii | Centro de Investigación Biomédica en Red Enfermedades Raras



IT4Innovations national supercomputing center

The EVA has its own customized version of BiERApp.

Conclusions

- 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.**

More info: publication

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

More info: new features

- ❏ Case-Control studies.
- ❏ Export results to VCF.
- ❏ Improve the database to allow larger datasets.
- ❏ Add new filters: HPO, GO, tissues, pathways, ...

More info: BiERApp behind the scenes