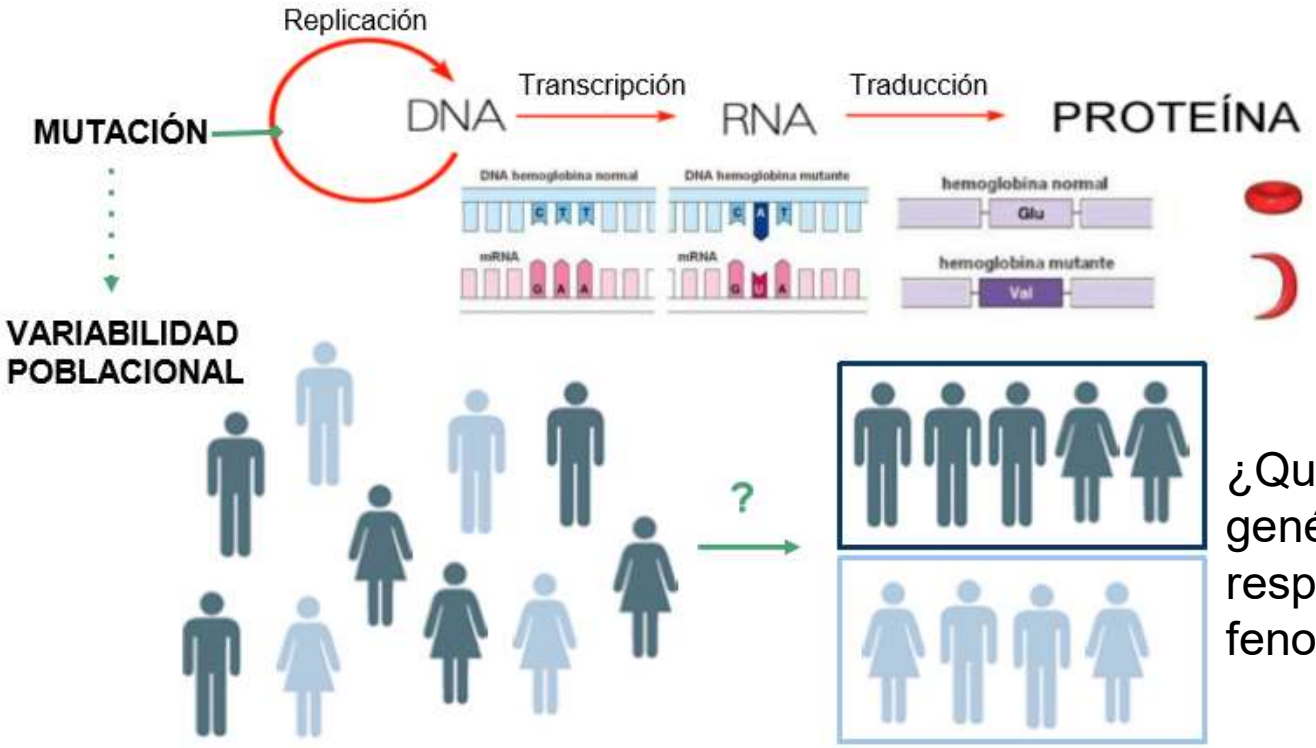
Instituto de Salud Carlos III

Variantes Genéticas



¿Qué diferencias genéticas son responsables de un fenotipo?

Tipos de variación genética

SNP > 1% en la población
INDELS

SVs: Variantes estructurales como duplicaciones, grandes deleciones ó inseciones (CNVs), inversiones y translocaciones.

NCBI **dbSNP** y **dbVAR**

Análisis de la variabilidad genética

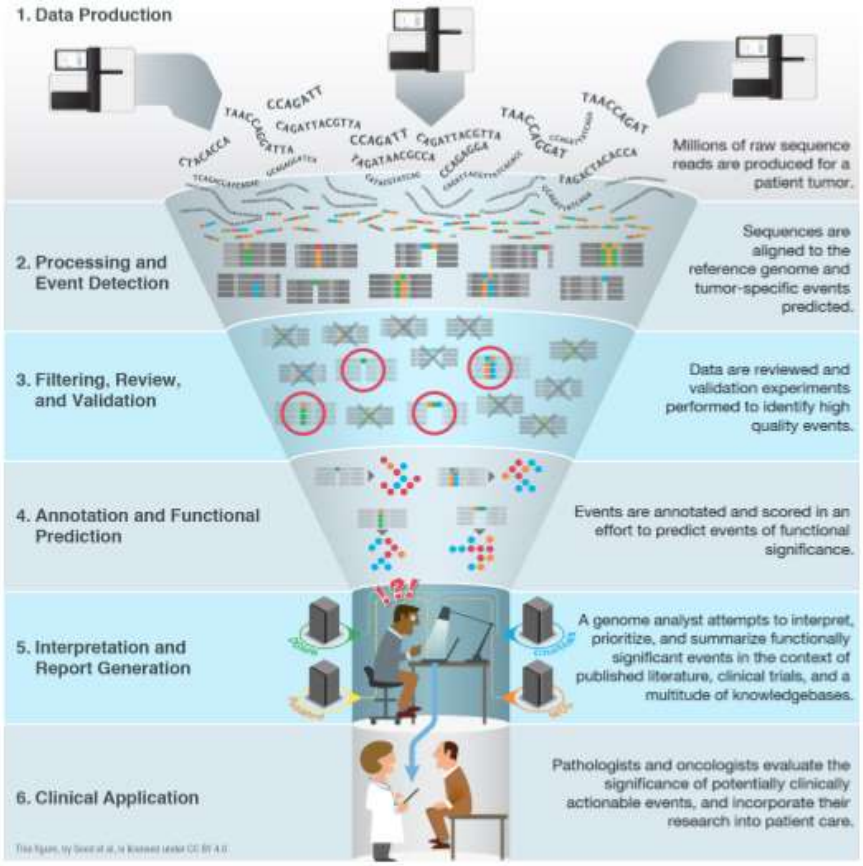
SNPs: 90% variabilidad genética
Aparece 1 SNP cada **1300 pb**
1 variante → 1 fenotipo ? . Enfermedades mendelianas
Sin impacto
Individualidad: pigmentación, receptores olfativos, …
Susceptibilidad a enfermedades complejas.
Metabolismo y inmunidad
Respuesta farmacológica

Gran número de variantes por individuo:

Exoma: 250.000 variantes.
Genoma: > 6 M de variantes.
Exoma clínico (6700 genes: 10.000-15.000 variantes)

Instituto de Salud Carlos III

NGS: Análisis de Variantes



Good BM, Ainscough BJ, McMichael JF, Su AI, Griffith OL. 2014. Genome Biology. 15(8):438.

Instituto de Salud Carlos III

Galaxy

```
tophat2 -p 4 \
--library-type fr-unstranded \
--no-novel-juncs \
-G Saccharomyces_cerevisiae.EF4.69.gtf \
-o control \
Saccharomyces_cerevisiae.EF4.69.dna.toplevel.fa \
control.fastq
```

Galaxy es una plataforma web de software libre para el análisis de datos en investigación biomédica.

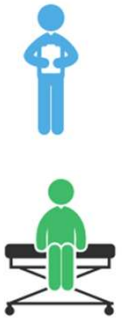
Accesible
Reproducible
Transparente

- https://wiki.galaxyproject.org/PublicGalaxyServers

NGS: Análisis de Variantes

Instituto de Salud Carlos III

1. Control de Calidad (CC)	2. Mapeo y CC	3. Llamada de variantes y CC	4. Priorización de variantes
"Limpieza" de las secuencias con baja calidad	Mapeo Recalibración por calidad de bases	Creación de ficheros con SNVs, indels y missing values	Anotación Análisis de Herencia Filtrado por frecuencia Priorización de una via Funcional…
FASTQ file	BAM file	VCF file	Informe



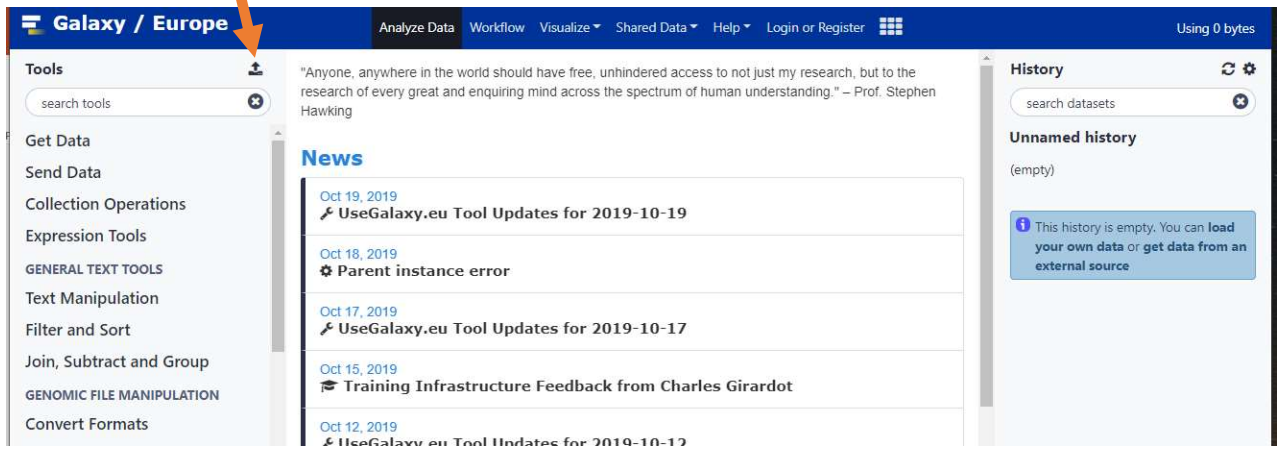
Análisis Primario

Instituto de Salud Carlos III

Diagnóstico del gen causante de la enfermedad de un niño afectado de osteopetrosis (proband), padre y madre no afectados a partir de los exomas del trio

1. En carpeta de Datos (forward/reverse orientation of all the fragments)
2. Subir a Galaxy archivos proband_R1.fq, proband_R2.fq como fastqsangergz

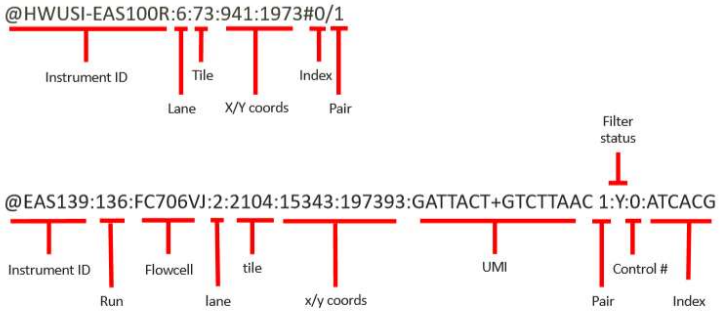
https://usegalaxy.eu/



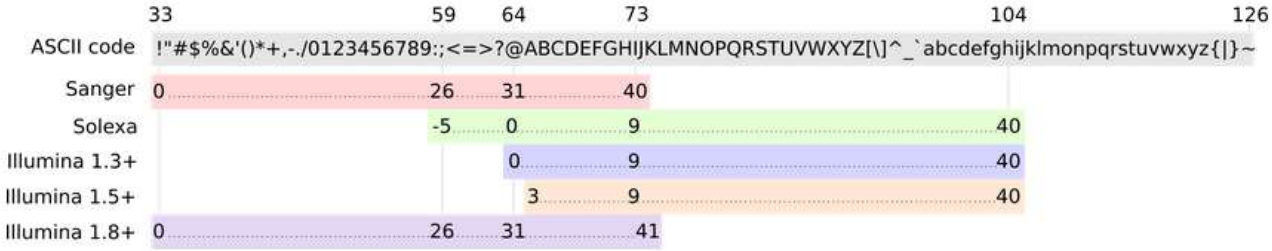
Visualizar el archivo que has subido Fastq

```
@HWUSI-EAS100R:6:73:941:1973#0/1
CTTTTTTATTTTTGTCTGACTGGGTTGATTCAAAA
+
CCCFFFFFHHHHGJHIIJHIHIIIFHIJJJJIJJG
```

- encabezado "@"
- Línea de secuencia
- Línea espaciadora
- Valores de calidad



Phred quality score



Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Instituto de Salud Carlos III

Tools

Fastqc

FASTA/FASTQ

Combine FASTA and QUAL into FASTQ

Manipulate FASTQ reads on various attributes

fastp - fast all-in-one preprocessing for FASTQ files

FASTQ Quality Control

FastQC Read Quality reports

Mapping

Executed **FastQC** and successfully added 1 job to the queue.

The tool uses this input:

- **4: proband_R1.fq.gz**

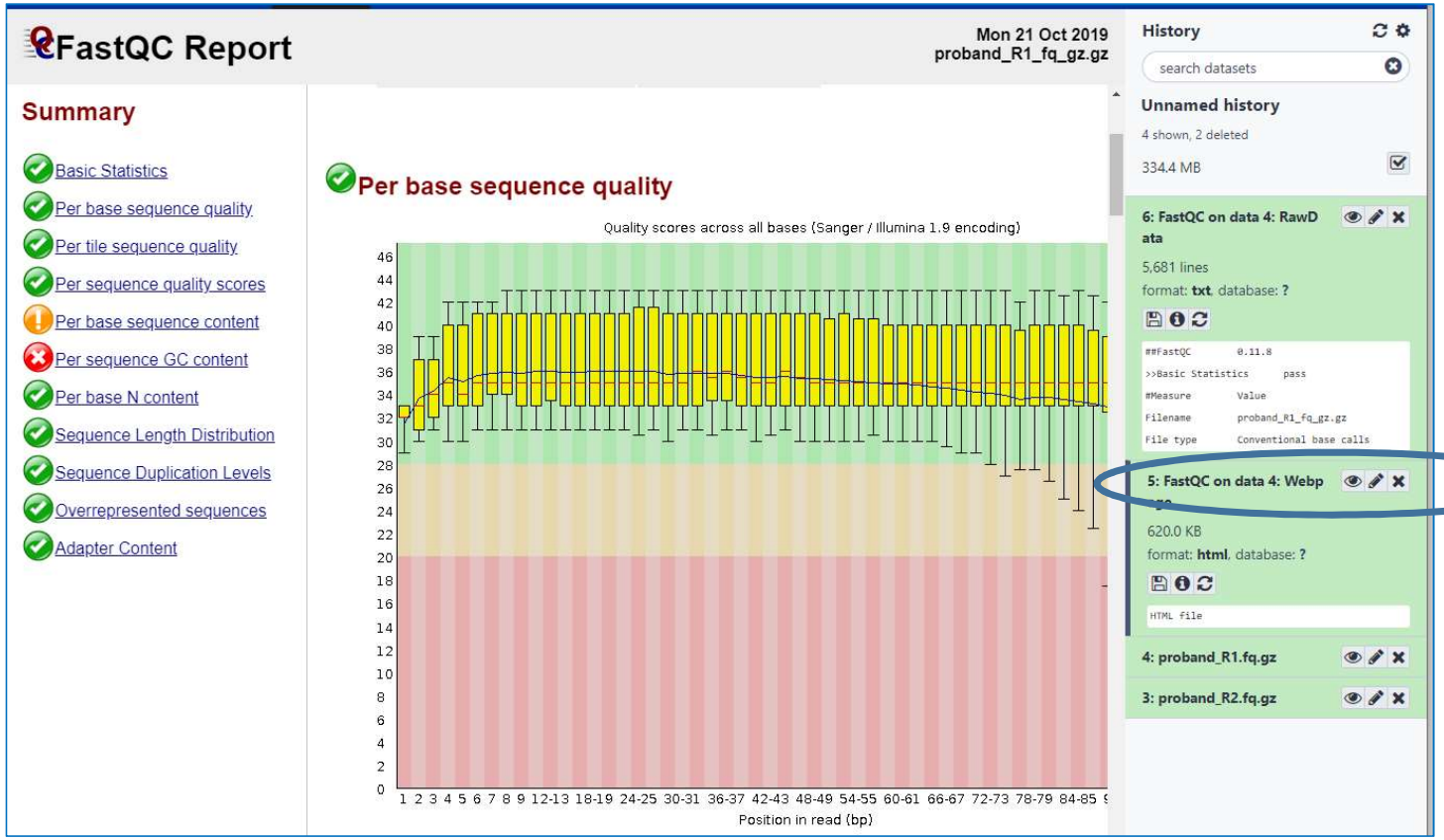
It produces 2 outputs:

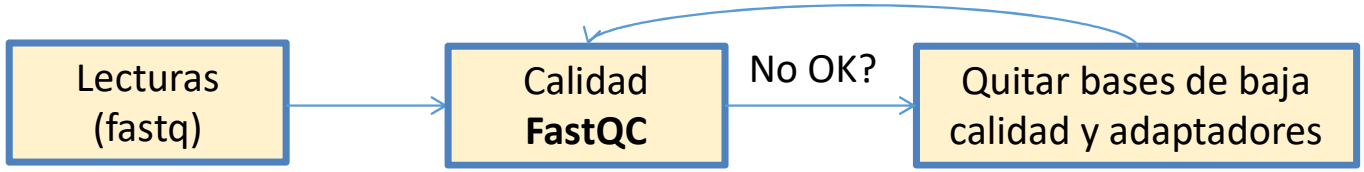
- **5: FastQC on data 4: Webpage**
- **6: FastQC on data 4: RawData**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Instituto de Salud Carlos III

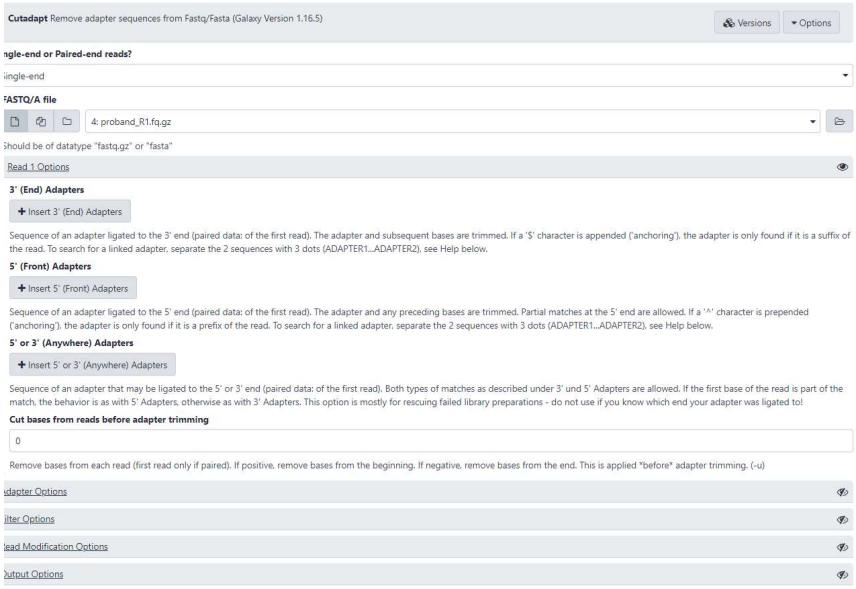
Análisis de calidad con **FastQC**





Phred score ~20, se deberían filtrar estas lecturas, usando **Cutadapt**

- “Minimum length”: 20
- “Quality cutoff”: 20
- Adapter removal (if any), we choose to remove ends (5’ and/or 3’) with low-quality, here below 20 in quality).



Instituto de Salud Carlos III

Mapeo con **BWA-MEM** utilizando el genoma Humano hg19

Tools

BWA

FASTA/FASTQ

Trim Galore! Quality and adapter trimmer of reads

AB-SOLID DATA

Convert SOLiD output to fastq

Annotation

TB-Profiler Profile Infer strain types and drug resistance markers from sequences

Mapping

Map with minimap2 A fast pairwise aligner for genomic and spliced nucleotide sequences

Map with BWA - map short reads (< 100 bp) against reference genome

Map with BWA-MEM - map medium and long reads (> 100 bp) against reference genome

Map with BWA-MEM

Variant Calling

SnpEff build: database from Genbank or GFF record

snippy Snippy finds SNPs between a haploid reference genome and your NGS sequence reads.

Epigenetics

bwameth Fast and accurate aligner of

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Using reference genome

Human (Homo sapiens): hg19 Full

Select genome from the list

Single or Paired-end reads

Paired

Select between paired and single end data

Select first set of reads

4: proband_R1.fq.gz

Specify dataset with forward reads

Select second set of reads

3: proband_R2.fq.gz

Specify dataset with reverse reads

Enter mean, standard deviation, max, and min for insert lengths.

-I; This parameter is only used for paired reads. Only mean is required while sd, max, and min will be inferred. Examples: both "250" and "250,25" will work while "250,,10" will not. See below for details.

Set read groups information?

Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode

1.Simple Illumina mode

✔ Execute

History

search datasets

Unnamed history

4 shown, 2 deleted

334.4 MB

6: FastQC on data 4: RawData

5,681 lines

format: **txt**, database: **?**

```
##FastQC        0.11.8
>>Basic Statistics      pass
#Measure        Value
Filename        proband_R1_fq_gz.gz
File type       Conventional base calls
```

5: FastQC on data 4: Webpage

620.0 KB

format: **html**, database: **?**

HTML file

4: proband_R1.fq.gz

3: proband_R2.fq.gz

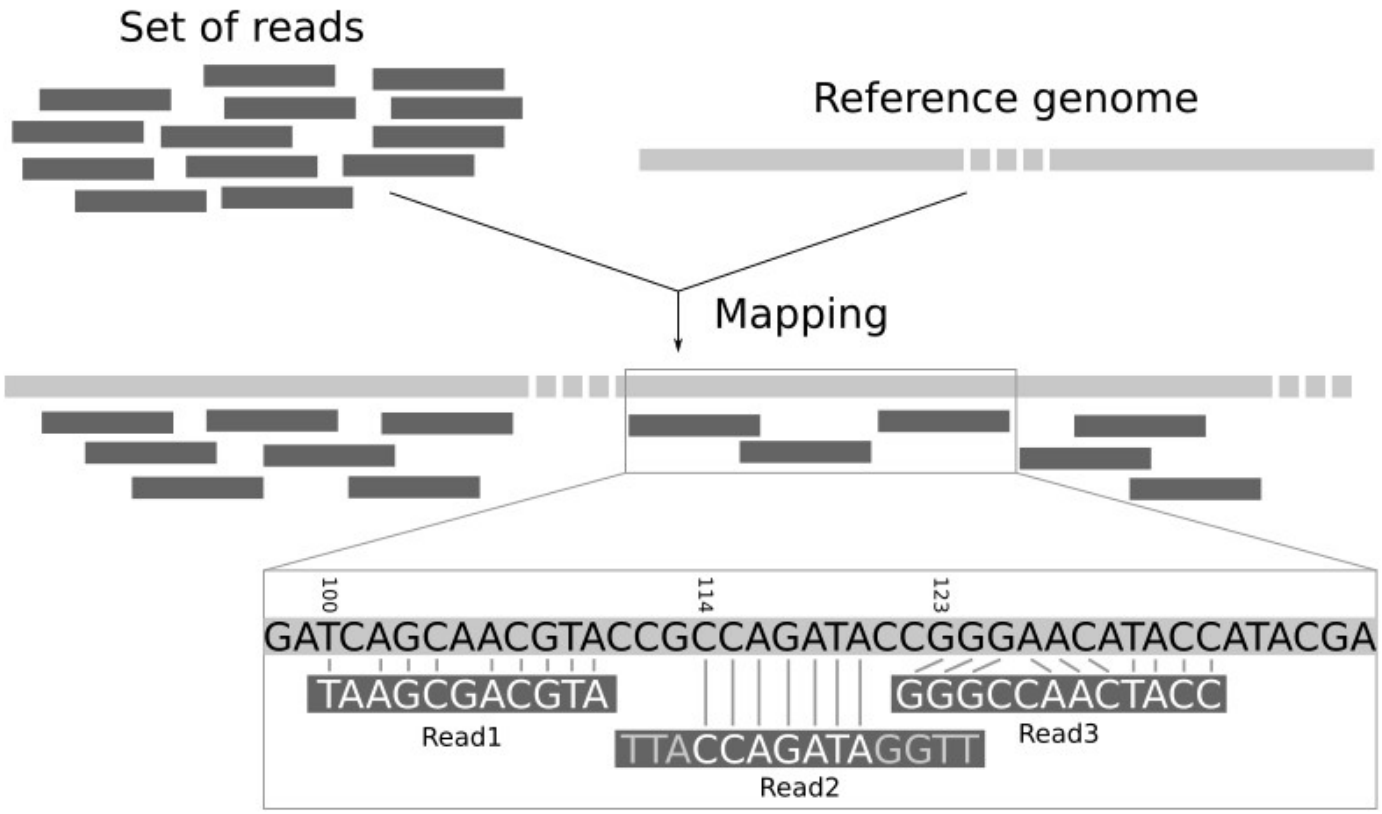
Instituto de Salud Carlos III

FastQC

OK?

Mapeo de las lecturas al genoma de referencia utilizando BWA-MEM

Archivo BAM (binario SAM)



Source: Galaxy

Instituto de Salud Carlos III

SAM = Sequence Alignment Map
BAM = Binary SAM = compressed SAM

Encabezado

@HD The header line
- VN: format version
- SO: Sorting order of alignments

@SQ Reference sequence dictionary
- SN: reference sequence name
- LN: reference sequence length
- SP: species

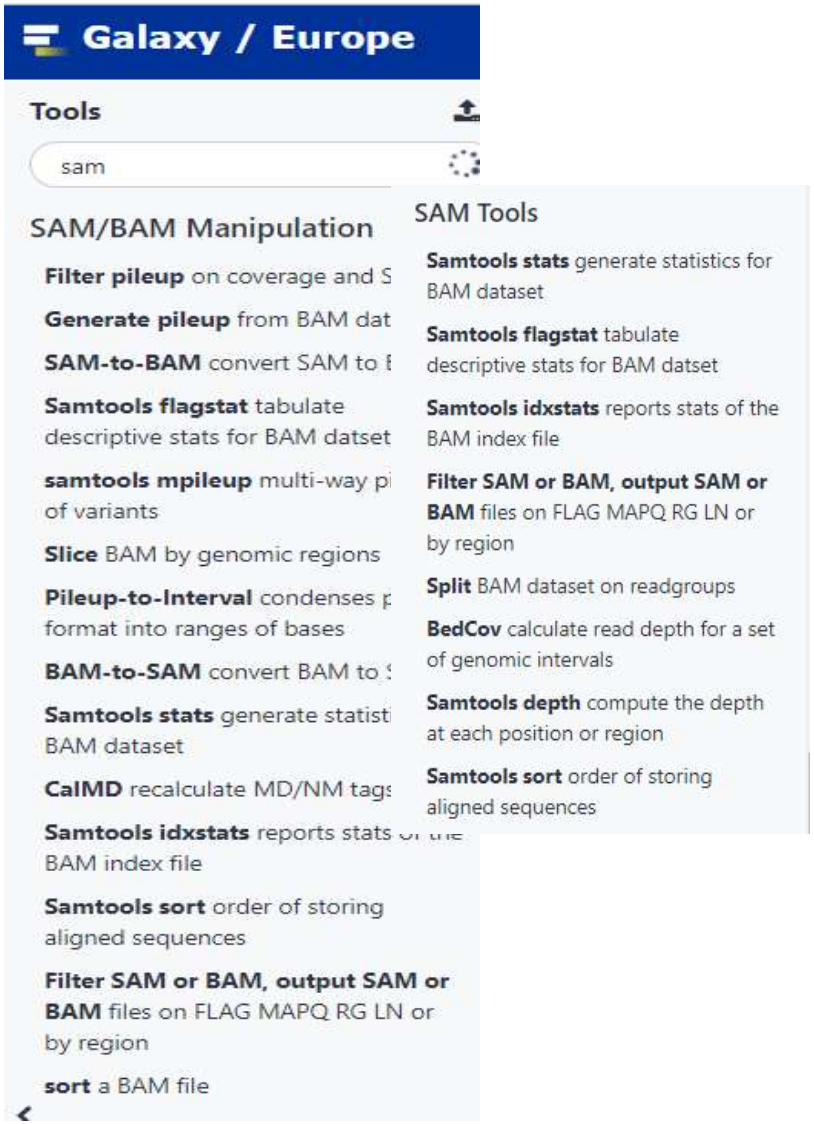
@RG Read group
- ID: read group identifier
- CN: name of sequencing center
- SM: sample name

@PG Program
- PN: program name
- VN: program version

Alineamiento

Nº	Campo	Tipo	Exp Regular / Rango	Descripcion
1	QNAME	Cadena	$[!-?A-\sim]\{1,255\}$	Nombre de la consulta
2	FLAG	Entero	$[0, 2^{16} - 1]$	Bandera de opciones
3	RNAME	Cadena	$\backslash * \|[! - () + - <> - \sim][!- \sim]*$	Referencia del nombre de la secuencia
4	POS	Entero	$[0, 2^{29} - 1]$	Posición de la primera base más a la izquierda
5	MAPQ	Entero	$[0, 2^8 - 1]$	Calidad del Mapeo
6	CIGAR	Cadena	$\backslash * \|([0 - 9] + [MIDNSHPX =])+$	Cadena CIGAR
7	RNEXT	Cadena	$\backslash * \| = \|[! - () + - <> - \sim][!- \sim]*$	Nombre de referencia del siguiente fragmento
8	PNEXT	Entero	$[0, 2^{29} - 1]$	Posición el siguiente fragmento
9	TLEN	Entero	$[-2^{29} + 1, 2^{29} - 1]$	Longitud de la plantilla
10	SEQ	Cadena	$\backslash * \|[A - Za - z = .]+$	Fragmento de secuencia
11	QUAL	Cadena	$[!- \sim]+$	Calidad de la secuencia

https://samtools.github.io/hts-specs/SAMv1.pdf



Estadística del alineamiento BAM

Samtools flagstat

Control de calidad de los archivos BAM

RmDup Quita duplicados de PCR

"Filter on bitwise flag": yes

"Only output alignments with all of these flag bits set":

param-check *"Read is mapped in a proper pair"*

"Is this paired-end or single end data": BAM is paired-end

"Treat as single-end": No

Visualizar los archivos BAM

IGV

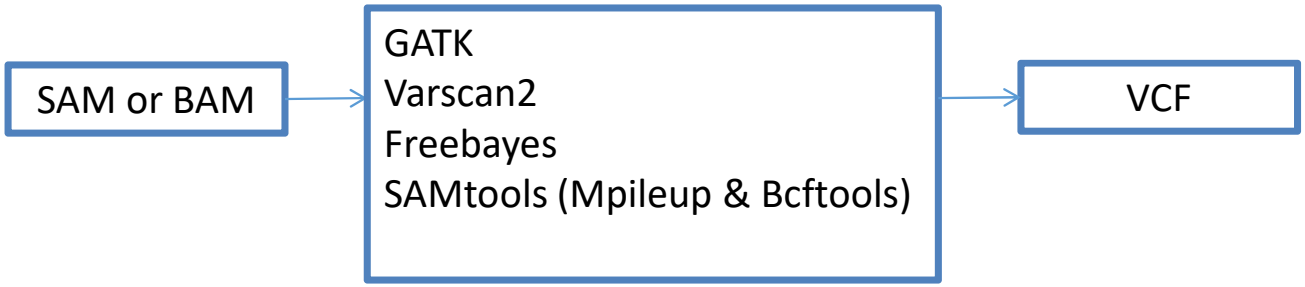
Instituto de Salud Carlos III

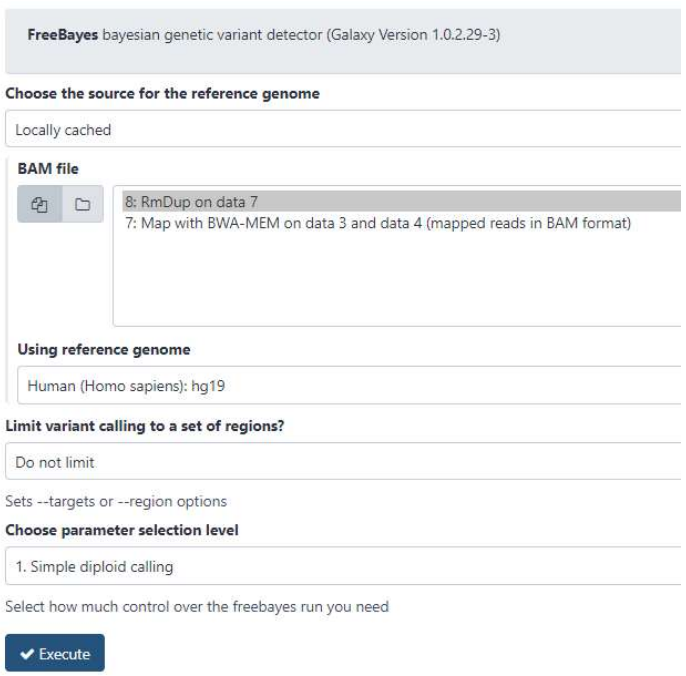
Llamada de variantes

```
REFERENCE: atcatgacggcaGtagcatat
---------------------------------
READ1:     atcatgacggcaGtagcatat
READ2:         tgacggcaGtagcatat
READ3:     atcatgacggcaAtagca
READ4:             cggcaGtagcatat
READ5:     atcatgacggcaGtagc
```

Podría ser:

- una verdadera variante
- un artefacto experimental, (un error de preparación de la muestra)
- un error de llamada de la base
- un error de análisis, una desalineación (aunque poco probable en el ejemplo anterior)





bcftools norm.

https://genome.sph.umich.edu/wiki/Variant_Normalization

- *"When any REF allele does not match the reference genome base"*: ignore the problem (-w)
- *"Left-align and normalize indels?"*: Yes
- *"Perform deduplication for the folowing types of variant records"*: do not deduplicate any records
- Freebayes is not producing any duplicate calls.
- *"~multiallelics"*: split multiallelic sites into biallelic records (-)
- *split the following variant types"*: both

VCF = Variant Call Format

http://samtools.github.io/hts-specs/VCFv4.3.pdf

Instituto de Salud Carlos III

Priorización de Variantes

- VEP
- GEMINI

VeP

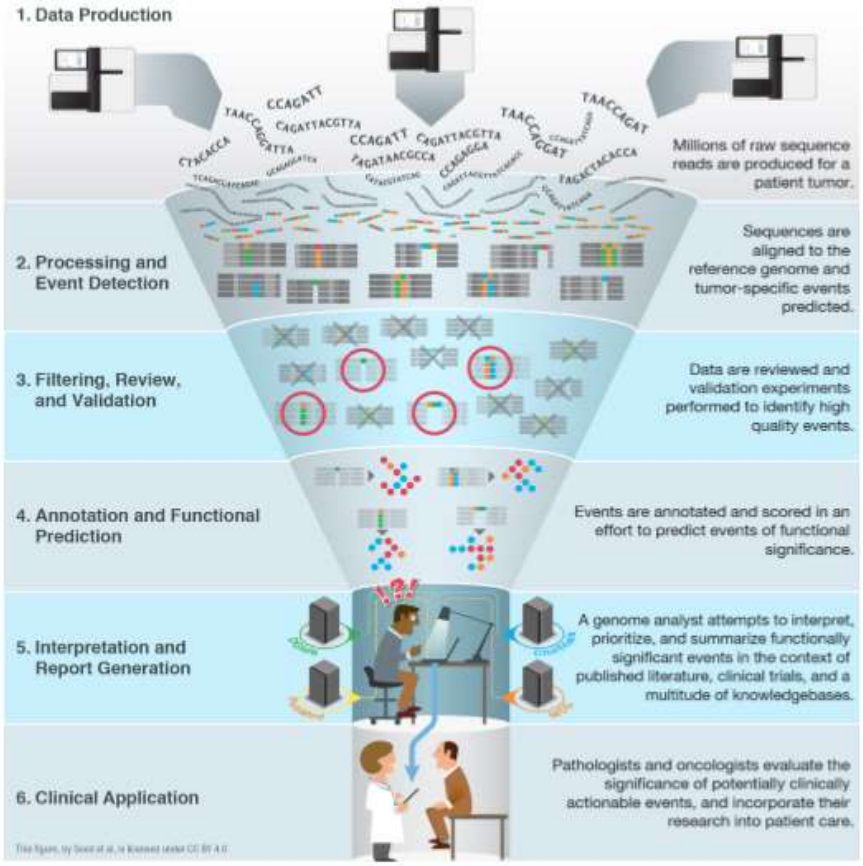
Galaxy

igv Integrative Genomics Viewer

Instituto de Salud Carlos III

Instituto de Salud Carlos III

NGS: Análisis de Variantes



Good BM, Ainscough BJ, McMichael JF, Su AI, Griffith OL. 2014. Genome Biology. 15(8):438.

Anotación en Bases de datos clínicas

Guia American *College of Medical Genetics and Genomics* /*Association of Molecular Pathology*
ACMG/AMP 2015

	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data						Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3					
Segregation data	Nonsegregation with disease BS4					
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		

Significado incierto
Benigna
Probablemente benigna
Probablemente patogénica
Patogénica

Asociada con la condición estudiada y modo de herencia

Frecuencia poblacional

Instituto de Salud Carlos III

Catálogos de variabilidad genética.
El proyecto 1000 genomas:
http://www.1000genomes.org

Exome Sequencing Project (ESP):
http://evs.gs.washington.edu/EVS

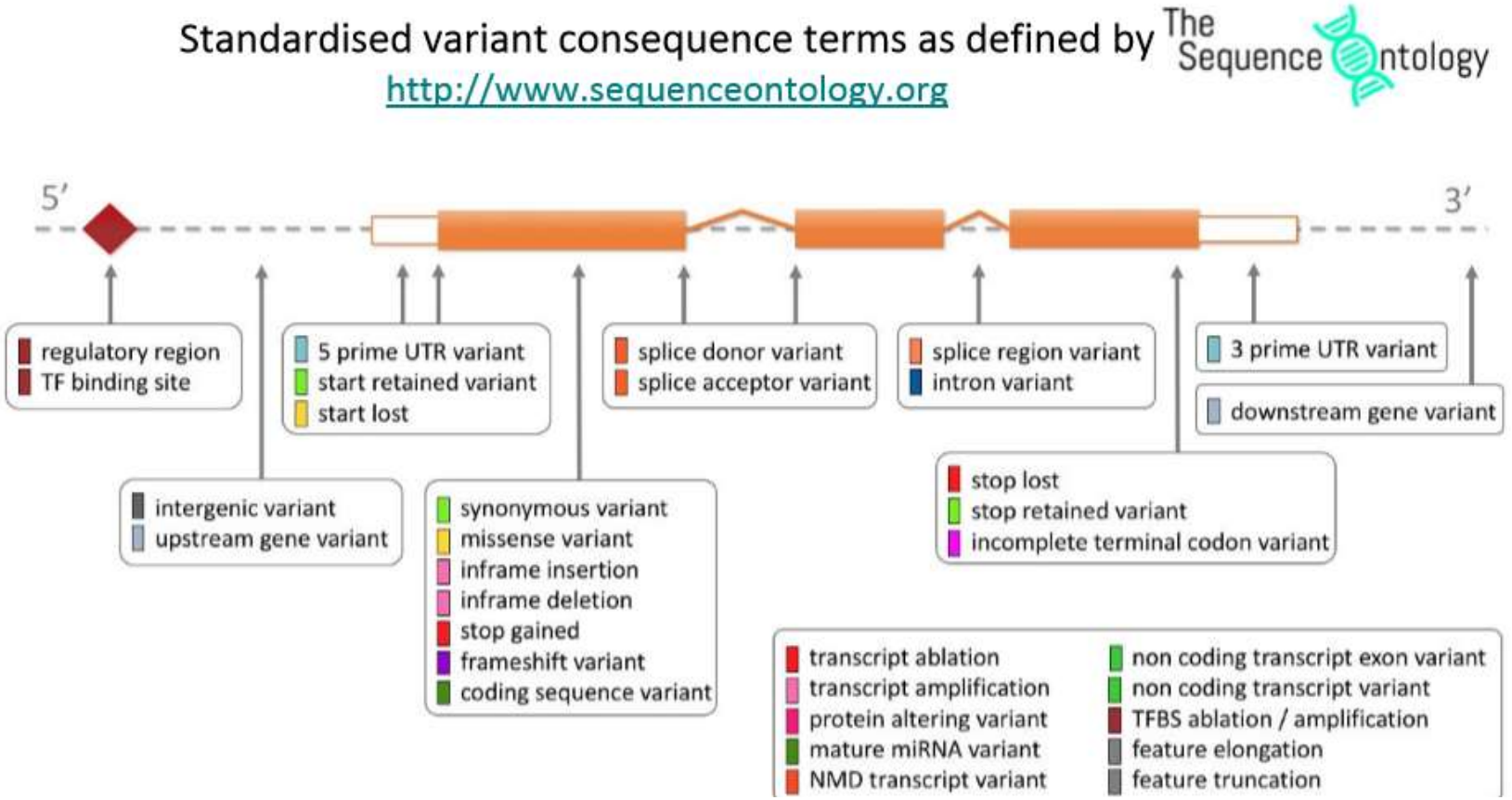
Exome Aggregation Consortium (ExAC/ gnomAD):
http://exac.broadinstitute.org
http://gnomad.broadinstitute.org/about

The CIBERER Spanish Variant Server (CSVS):
http://csvs.clinbioinfosspa.es

MAF: Frecuencia del alelo minoritario → Incidencia enfermedad

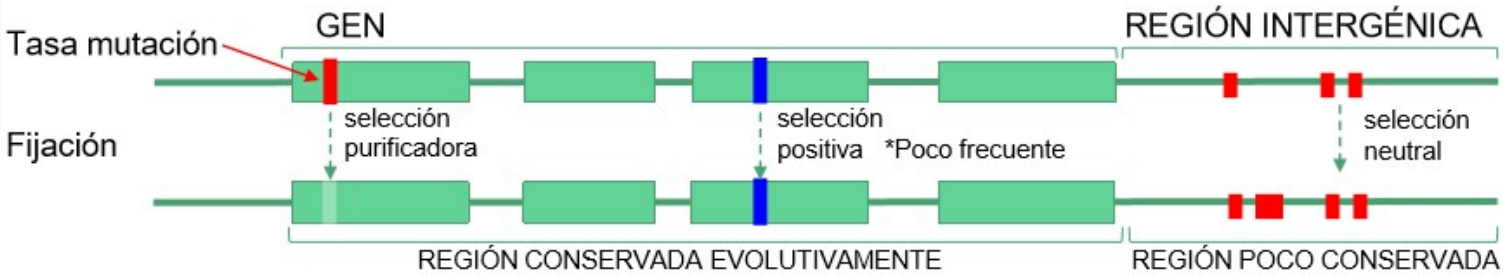
- Tipo de herencia
- Penetrancia
- Edad de inicio

Consecuencia: Sequence Ontology (SO)



LoF (Loss of Function): Provocan pérdida de función: splicing, STOP, ...

Conservación



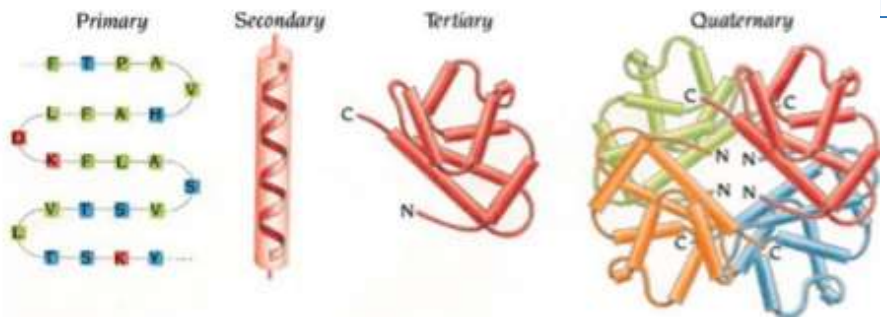
Regiones con funciones importantes para la célula acumulan poca variabilidad: más conservadas

GERP: Genomic Evolutionary Rate Profiling. Método para estimar la presión selectiva de una posición específica mediante máxima verosimilitud. A mayor valor, mayor presión selectiva.

PhastCons: Calcula la conservación por ventanas usando un modelo filogenético HMM (hidden Markov model). - [0 - 1]: Representa la probabilidad de que una región esté sometida a su mayor conservación posible.

PhyloP: Calcula el -log p-valor bajo una hipótesis nula de evolución neutral. - En humanos [-14, 3]: scores positivos indican conservación, negativos indican una aceleración.

Patogenicidad



La **función** de una proteína depende de su **estructura** cuaternaria final.
Su estructura depende de las **propiedades fisicoquímicas** de los **aminoácidos** que la componen.

SIFT [0 - 1] sorts intolerant from tolerant Probabilidad de tolerancia basada en la homología de proteínas entre alineamientos.

Polyphen [0 - 1] Polymorphism Phenotyping v2 Usa **parámetros fisicoquímicos** y anotación **estructural** para calcular, mediante métodos bayesianos, la patogenicidad de un cambio de aminoácido.

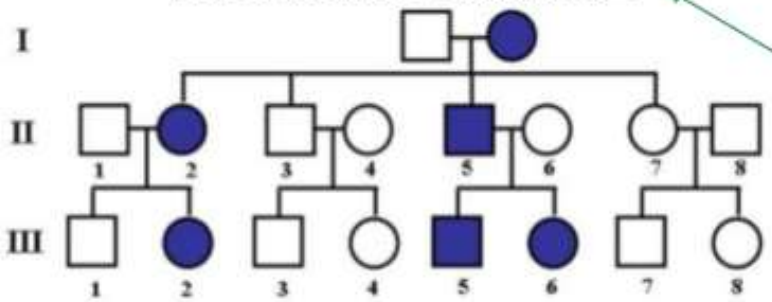
CADD: Combined Annotation Dependent Depletion. Usando **información anotada disponible**, evalúa en un ranking el impacto de la variante.
ScaledCADD 10 = 10%
ScaledCADD 20 = 1%

Valor de SIFT	Predicción
< 0.05	DAMAGING
>= 0.05	TOLERATED
Valor de Polyphen	**Predicción**
> 0.908	Probably damaging
0.446 - 0.908	Possibly damaging
< 0.446	Benign

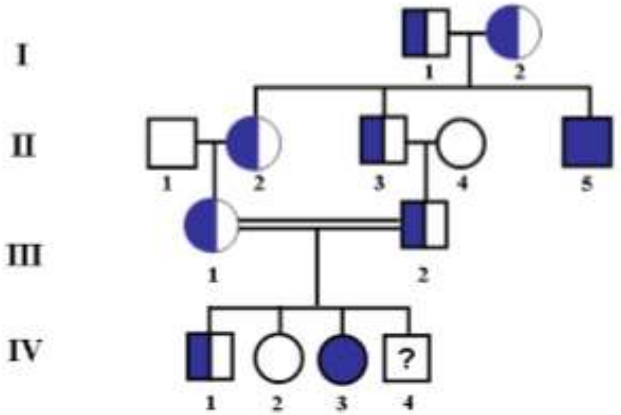
dbNSFP Prediccion functional de variantes no sinónimas del genoma

Modos de herencia: Genotipo

HERENCIA DOMINANTE



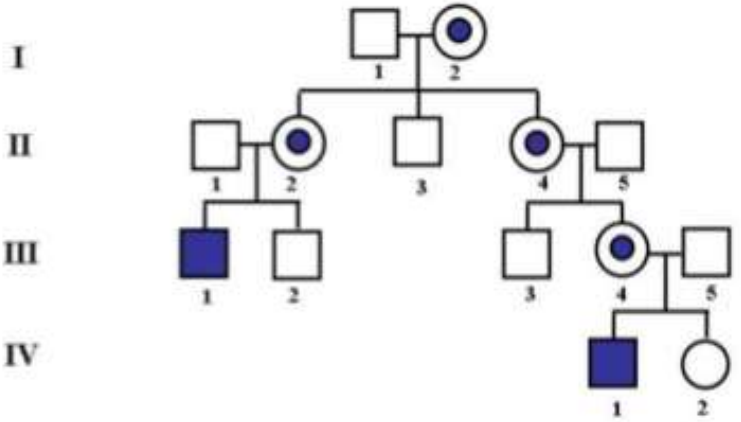
HERENCIA RECESIVA



HETEROCIGOSIS COMPUESTA

HERENCIA LIGADA AL X (DOMINANTE)

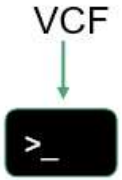
HERENCIA LIGADA AL X (RECESIVA)



HERENCIA LIGADA AL Y

Sólo varones afectados, transmisión de padre a hijos.

Software



ANNOVAR
snpEff
GEMINI
…

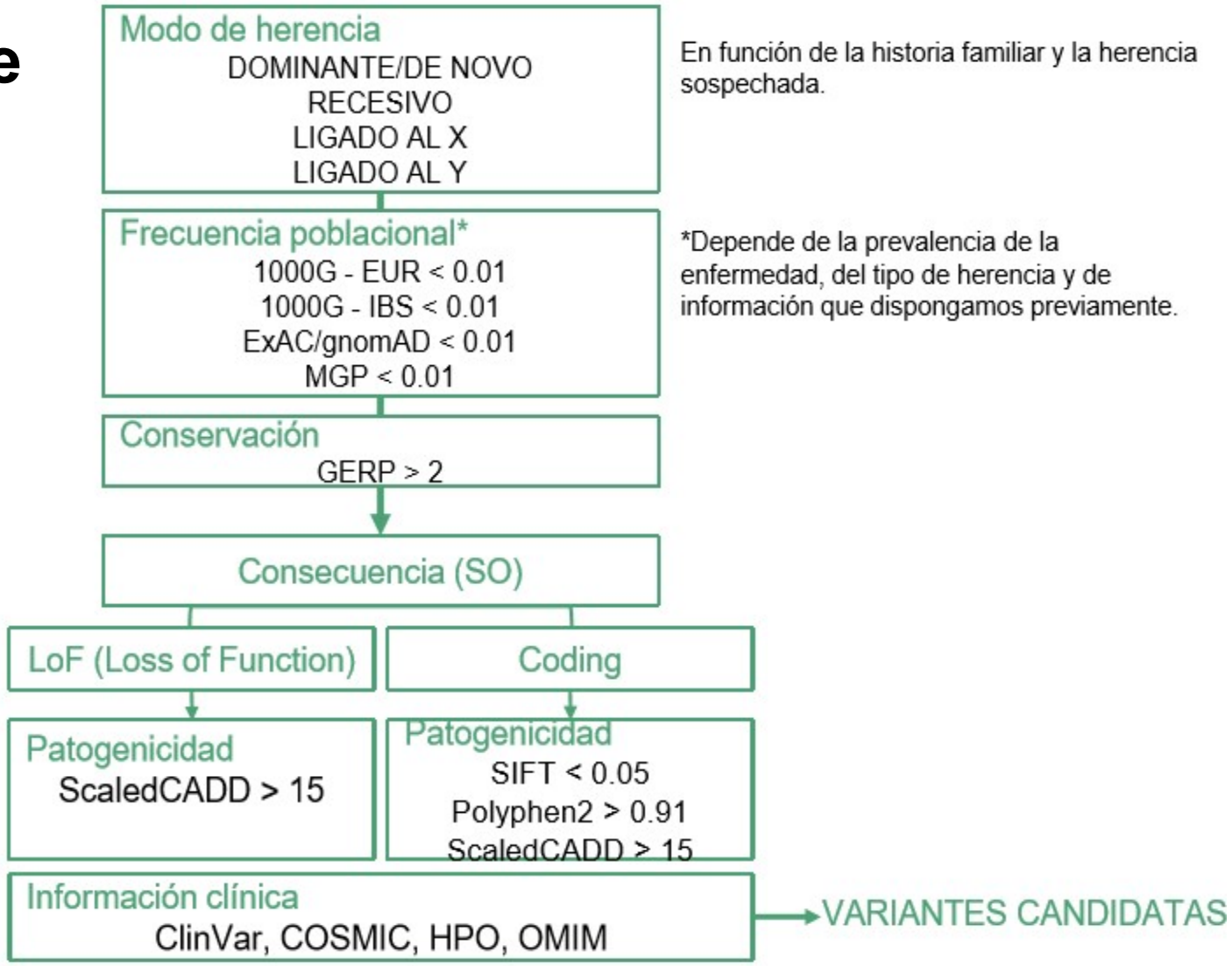
Variant Effect Prediction VEP

Bystro
…

Tener presente:

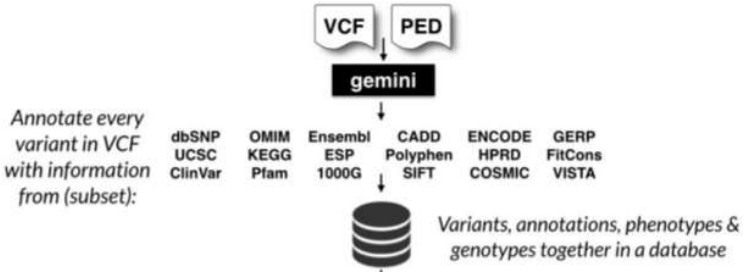
Los resultados pueden variar de una herramienta a otra
Falsos positivos
Errores en el efecto que describen de la variante…
(frameshift en lugar de stop loss…)

Priorización de Variantes



!! Algunas variantes LoF tampoco tienen CADD asociado

GEMINI (GEnome MINIng) in Galaxy



3 archivos VCF (father, mother, proband) anotados con SnpEff (Merge Galaxy)
archivo PED "pedigrí"

#family_id	name	paternal_id	maternal_id	sex	phenotype
FAM	father	0	0	1	1
FAM	mother	0	0	2	1
FAM	proband	father	mother	1	2

family_trio.sqlite
(en datos del curso)
Subirlo con extensión sqlite y anotado en hg19

GEMINI inheritance pattern

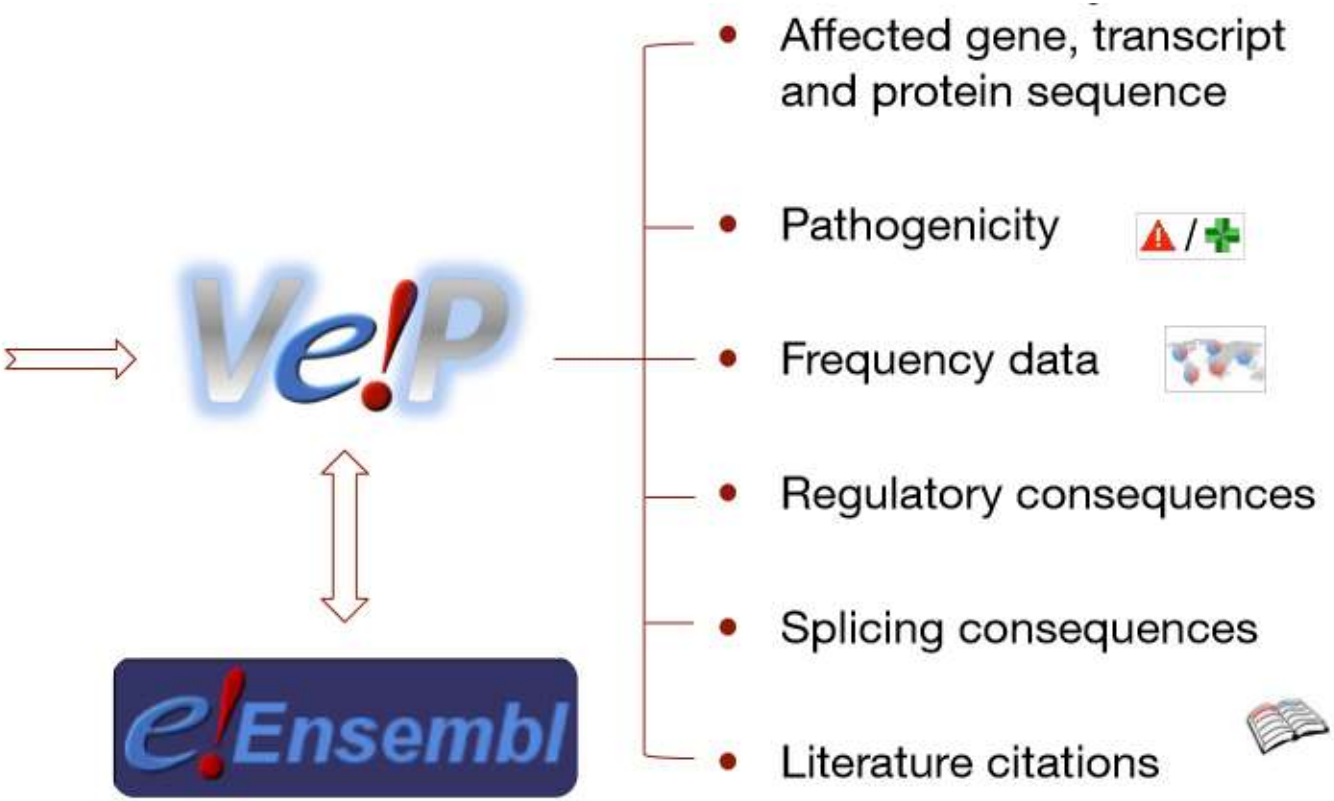
"Additional constraints expressed in SQL syntax"

impact_severity != 'LOW'

Ouput Custom. Set of columns to include in the variant report table:
clinvar_sig, clinvar_disease_name, clinvar_gene_phenotype, rs_ids

Anotación de Variantes

Datos de Variantes



Instituto de Salud Carlos III

VeP

IMPORTACIÓN DE VARIANTES DE DIVERSAS BASES DE DATOS

CONTROL DE CALIDAD

Instituto de Salud Carlos III

VeP

IMPORTACIÓN DE VARIANTES DE DIVERSAS BASES DE DATOS

CONTROL DE CALIDAD

ANOTACION CON DATOS DE POBLACION

HGMD®
The Human Gene Mutation Database
Cardiff

1000 Genomes
A Deep Catalog of Human Genetic Variation

ClinVar

Europe PMC

OMIM®

gnomAD
genome aggregation database

UK 10K
UK10K
Rare Genetic Variants in Health and Disease

NHLBI Trans-Omics for Precision Medicine

Instituto de Salud Carlos III

VeP

IMPORTACIÓN DE VARIANTES DE DIVERSAS BASES DE DATOS

CONTROL DE CALIDAD

ANOTACION CON DATOS DE POBLACION

PREDICCION

Predicción:

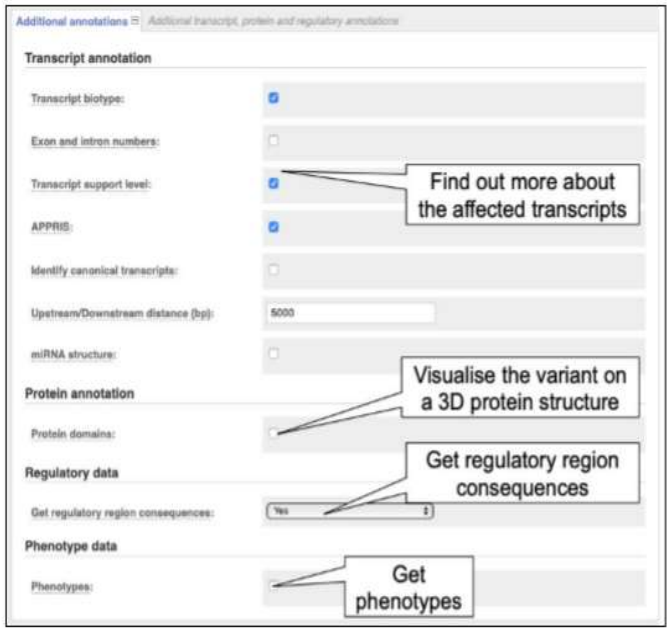
- Consecuencia de las variantes
- Predicción de la Función de proteínas
- Datos de desequilibrio de ligamiento
- Conservación de la variante entre especies

Instituto de Salud Carlos III

http://grch37.ensembl.org/Homo_sapiens/Tools/VEP

Subir el fichero probandch8.vcf_bgzip

Añadir que incluya el dato de fenotipo
Filtrar por IMPACT is HIGH
Consequence is stop_gained



Navigation (per variant)
Page: 1 of 2 | Show: 1 5 10 50 All variants

Filters
Impact is HIGH
Consequence is stop_gained
Clear filters
Match all of the above rules
Update
Uploaded variant is defined
Add

Download
All: VCF VEP TXT
Filtered: VCF VEP TXT
BioMart: Variants Genes

New job

Show/hide columns (13 hidden)

	Feature type	Feature	Biotype	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand	SIFT	PolyPhen	AF	Clinical significance	Phenotype or disease	Pubmed	Associated phenotypes
000104267	Transcript	ENST00000285379.5	protein_coding	3/7	521	291	97	W/*	TGG/TGA	rs1173332839	1	-	-	-	-	-	-	OSTEOPETROSIS AUTOSOMAL RECESSIVE 3 (ENSG00000104267,MIM morbid) OSTEOPETROSIS AUTOSOMAL RECESSIVE TYPE 3 (ENSG00000104267,DDG2P) Osteopetrosis with renal tubular acidosis (ENSG00000104267,Orphanet)
000186523	Transcript	ENST00000321602.8	protein_coding	2/4	285	50	17	S/*	TCA/TGA	rs201723492	-1	-	-	-	-	-	-	-

Instituto de Salud Carlos III

Anotación de Variantes

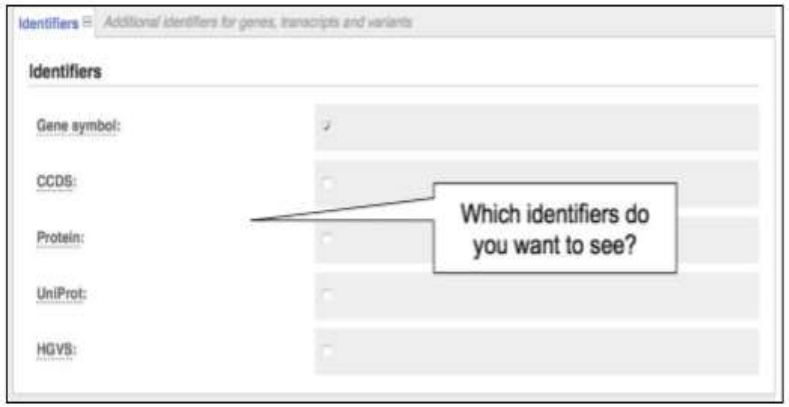
Variant coordinates (Ensembl default)	1 881907 881906 -/C + 5 140532 140532 T/C + 12 1017956 1017956 T/A + 2 946507 946507 G/C + 14 19584687 19584687 C/T -
HGVS notation	ENST00000285667.3:c.1047_1048insC 5:g.140532T>C NM_153681.2:c.7C>T ENSP00000439902.1:p.Ala2233Asp NP_000050.2:p.Ile2285Val
VCF	#CHROM POS ID REF ALT 20 14370 rs6054257 G A 20 17330 . T A 20 1110696 rs6040355 A G,T 20 1230237 . T .
Variant IDs	rs41293501 COSM327779 rs146120136 FANCD1:c.475G>A rs373400041

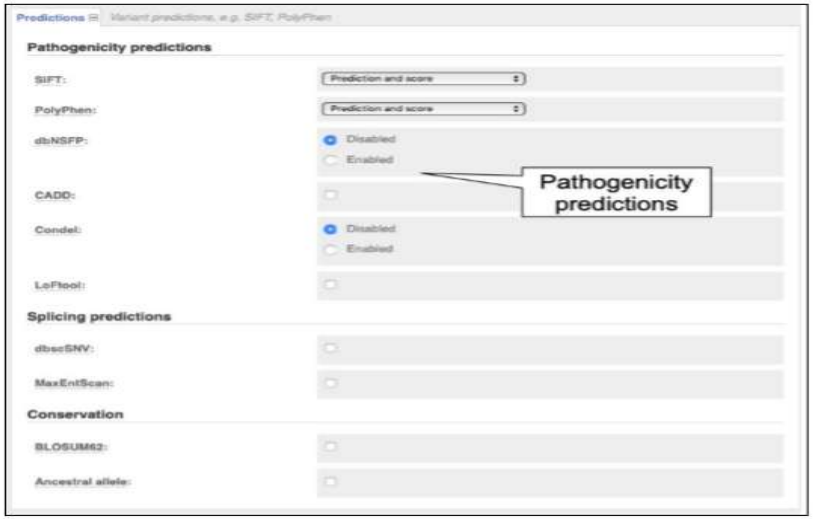
http://varnomen.hgvs.org/recommendations/general/

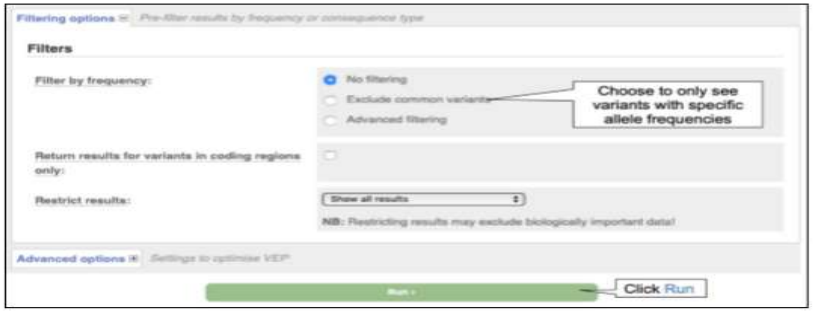
Instituto de Salud Carlos III

VeP

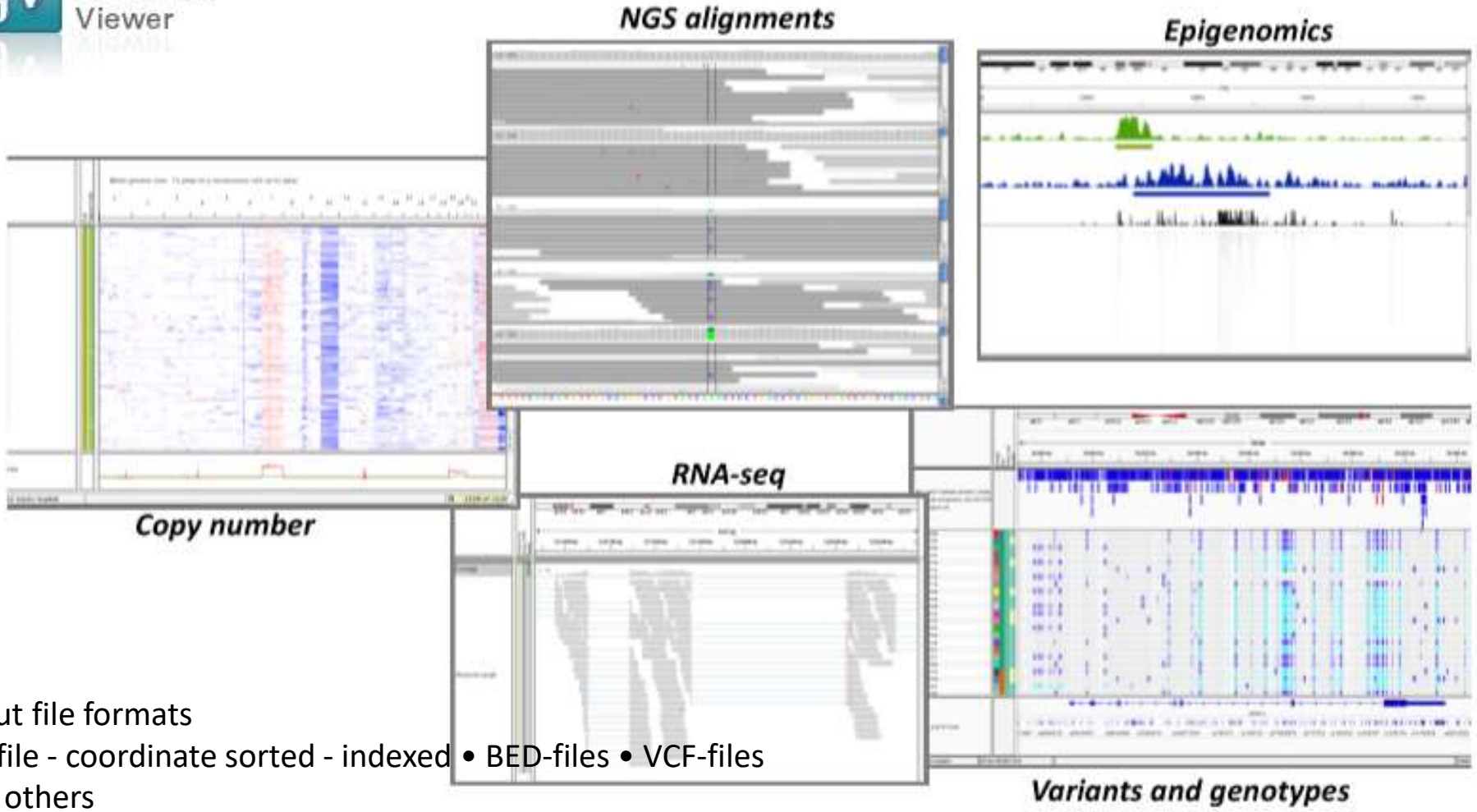
https://www.ensembl.org/Homo_sapiens/Tools/VEP







Integrative Genomics Viewer



IGV input file formats

- BAM-file - coordinate sorted - indexed • BED-files • VCF-files
- Many others

http://software.broadinstitute.org/software/igv/

Instituto de Salud Carlos III

1. Inspeccionar alineamientos y covertura de archivos BAM
2. SNVs
3. InDels

Por defecto, IGV tiene cargado Human (hg19).
Puedes cargar pistas adicionales (genes Ensembl, dbSNP…) desde Archivo -> Cargar desde el servidor
Navega por cromosomas ó genes. En este caso vamos al gen Ca2

- Por defecto, IGV tiene cargado Human (hg19).
- Puedes cargar pistas adicionales (genes Ensembl, dbSNP…) desde Archivo -> Cargar desde el servidor
- Navega por cromosomas ó genes. En este caso vamos al gen Ca2