CURSO PRÁTICO DE BIOINFORMÁTICA:

Aula 5 – VCF

https://github.com/rpmartin85/BIOINFO

Renan Paulo Martin

19/07/19

Arquivo VCF

- Arquivo de texto contendo dados de variantes genéticas;
- Dividido principalmente em 3 partes:
  - Meta-information:
    - Linhas que iniciam com ##;
  - Header Line:
    - Uma linha iniciando com #;
  - Data lines:
    - Linhas contendo os dados das variantes genéticas.

Exemplo

```
##fileformat=VCFv4.2
##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
```

Meta-information lines

- Contém informações referentes aos valores apresentados da seção de dados;
- Formato chave = valor;
- Podem conter descrição dos seguintes campos:
  - INFO, FILTER e FORMAT;
- Chaves são separadas por vírgulas;
  - ID, Number, Type, Description.

Header line

- Contém 8 campos obrigatórios separados por tabulação:
  - #CHROM;
  - POS;
  - ID;
  - REF;
  - ALT;
  - QUAL;
  - FILTER;
  - INFO.
- Campos seguintes representam as amostras ao qual o sequenciamento se refere.

Data lines

- Contém 8 campos obrigatórios separados por tabulação, mesmos campos do header;
- Lista de variantes encontradas no sequenciamento;
- Múltiplos alelos são permitidos por linha;
- Se os dados do genótipo estiverem presentes, todas as amostras devem usar o mesmo formato;
- Cada linha suporta apenas uma posição no genoma;
- A posição é orientada de acordo com o Assembly ao qual o sequenciamento foi alinhado, seguindo as coordenadas desse assembly.

Limitações

- Normalmente apresenta apenas as informações referentes ao sequencimento;
- Dependendo da estratégia de sequenciamento, pode ser um arquivo muito longo;
- Sua vizualização em editores de planilha é dificultada devido ao tamanho do arquivo;
- Sua manipulação tem que ser feita com cautela para não perder formatação.

Vantagens

- Uso de programas específicos permitem sua manipulação mesmo estando compactado;
- Apresenta um arquivo index conjugado que aumenta a performance das manipulações necessárias;
- Formato universal e aberto, aceito pela grande maioria de ferramentas de análises;
- Permite anotações adicionais;
- Possibilidade de armazenamento de múltiplas amostras em um mesmo arquivo.

Geração do VCF

- Diversas ferramentas podem ser usadas para diferentes propostas:
  - GATK;
    - haplotype caller;
    - Mutect2.
- Arquivo BAM alinhado + genoma de referência (input).

HaplotypeCaller

COMANDO ENCONTRA-SE NO ARQUIVO COMANDO_AULA_5.txt

java -jar /data/bioinfo2019/GenomeAnalysisTK.jar -T HaplotypeCaller -R /data/hg38_baseFiles/hg38/hg38.fa -I /data/renan/HG00096_CHR20_sorted_MKD_GATKrecal.bam --emitRefConfidence GVCF --dbsnp /data/hg38_baseFiles/dbsnp151/Homo_sapiens_assembly38.dbsnp151.vcf.gz -L chr20 -o HG00096_CHR20.vcf -variant_index_type LINEAR -variant_index_parameter 128000

Manipulação BCFTOOLS

- Série de ferramentas:
  - annotate;
  - concat;
  - filter;
  - index;
  - merge;
  - norm;
  - query;
  - sort;
  - view.

```
[Renans-MacBook-Pro:~ renanmartin$ bcftools

Program: bcftools (Tools for variant calling and manipulating VCFs and BCFs)
Version: 1.9 (using htslib 1.9)

Usage:   bcftools [--version|--version-only] [--help] <command> <argument>

Commands:

 -- Indexing
    index        index VCF/BCF files

 -- VCF/BCF manipulation
    annotate     annotate and edit VCF/BCF files
    concat       concatenate VCF/BCF files from the same set of samples
    convert      convert VCF/BCF files to different formats and back
    isec         intersections of VCF/BCF files
    merge        merge VCF/BCF files files from non-overlapping sample sets
    norm         left-align and normalize indels
    plugin       user-defined plugins
    query        transform VCF/BCF into user-defined formats
    reheader     modify VCF/BCF header, change sample names
    sort         sort VCF/BCF file
    view         VCF/BCF conversion, view, subset and filter VCF/BCF files

 -- VCF/BCF analysis
    call         SNP/indel calling
    consensus    create consensus sequence by applying VCF variants
    cnv          HMM CNV calling
    csq          call variation consequences
    filter       filter VCF/BCF files using fixed thresholds
    gtcheck      check sample concordance, detect sample swaps and contamination
    mpileup      multi-way pileup producing genotype likelihoods
    roh          identify runs of autozygosity (HMM)
    stats        produce VCF/BCF stats

 Most commands accept VCF, bgzipped VCF, and BCF with the file type detected
 automatically even when streaming from a pipe. Indexed VCF and BCF will work
 in all situations. Un-indexed VCF and BCF and streams will work in most but
 not all situations.
```

Exemplos de uso

- Criação de index:
  - bcftools index –t Arquivo.vcf.gz
- Exibição de arquivo VCF na tela do terminal:
  - bcftools view Arquivo.vcf.gz
- Ordenação das variantes:
  - bcftools sort Arquivo.vcf.gz –Oz –o Arquivo.vcf.gz
- Obter variantes em uma determinada região do genoma:
  - bcftools norm –T arquivoComRegioes.txt –Arquivo.vcf.gz –Oz –o ArquivoRegiao.vcf.gz

Atividade prática de manipulação

- Compactar o Arquivo.vcf;
- Gerar o index do Arquivo.vcf.gz criado no passo anterior;
- Exibir na tela todas as variantes do Arquivo.vcf.gz;
- Remover os dados dos genótipo do Arquivo.vcf.gz criando o ArquivoNoGT.vcf.gz e seu referido index;
- Criar um arquivo contendo somente as variantes no cromossomo X compactado e seu respectivo index;

Anotação

- O arquivo VCF precisa ser anotado para que seja feita as análises e interpretações;
  - A anotação feita por programas externos ao BCFTOOLS deve ser executada somente após o término das manipulações desejadas.
- Principais anotações:
  - Função Gênica;
  - Genética de População;
  - Bancos de Dados adicionais;
  - Preditores de Patogenicidade.

Função Gênica

- Arquivo VCF não contem informação referente ao gene em que a variante está associada;
- Coordenada no Genoma de referência (HG19-GRCh37 ou HG38-GRCh38);
- Anotação de algum assembly é necessário para determinar a função gênica da variante;
- RefSeq;
- KnownGene;
- Ensembl.

RefSeq

- Func.refGene;
  - Localização da variante (intergênica, intrônica, exônica e etc).
- Gene.refGene;
  - Gene ID no RefSeq.
- ExonicFunc.refGene;
  - Função exônica associada a variante (sinônima, não sinônima, nonsense e etc).
- AAChange.refGene;
  - Informação referente à troca aminoácido / base.

Genética de População

- Uma das principais anotações;
- Diversas bases de dados disponíveis:
  - ESP6500 (6.500 exomas);
  - ExAC (65.000 exomas);
  - gnomAD (123.136 exomas + 15.496 genomas);
  - ABraOM (609 exomas brasileiros);
  - 1000g (2.504 genomas);
  - KAVIAR (64.000 exomas + 13.000 genomas);
- Permitem consultas através de website.

Interested in working on the development of this resource? Apply here.

ExAC Browser (Beta) | Exome Aggregation Consortium

Search for a gene or variant or region

Examples - Gene: PCSK9, Transcript: ENST00000407236, Variant: 22-46615880-T-C, Multi-allelic variant: rs1800234, Region: 22:46615715-46615880

About ExAC

The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

The data set provided on this website spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. The ExAC Principal Investigators and groups that have contributed data to the current release are listed here.

All data here are released under a Fort Lauderdale Agreement for the benefit of the wider biomedical community - see the terms of use here.

Sign up for our mailing list for future release announcements here.

Recent News

August 8, 2016

- CNV calls are now available on the ExAC browser

March 14, 2016

- Version 0.3.1 ExAC data and browser (beta) is released! (Release notes)

January 13, 2015

- Version 0.3 ExAC data and browser (beta) is released! (Release notes)

October 29, 2014

- Version 0.2 ExAC data and browser (beta) is released! Sign up for our mailing list for future release announcements here.

October 20, 2014

- Public release of ExAC Browser (beta) at ASHG!

October 15, 2014

- Internal release to consortium now available!

Reference: hg19

Search for Gene Name, Region, Position or Variant ID

Gene Name (ex: HBB), Region (ex: 11:5246696-5248301 or chr11:5246696-5248301), Position (ex: 11:5248232 or chr11:5248232) or Variant ID (ex: rs334)

ABraOM

Arquivo **Bra**sileiro **O**nline de **M**utações

Online Archive of Brazilian Mutations

This variant repository contains genomic variants of Brazilians. Our goal is to provide the community with genetic variability found in Brazil.

The initial deposited cohort 'SABE609' comprise exomic variants of 609 elderly individuals from a census-based sample from the city of São Paulo.

A total of **2,382,573** variants were called before filtering and are available at our browser. From that total, **1,264,224** are high confidence (GATK PASS flags **and** excluding CEGH-USP FDP/FAB flags).

Please refer to the about page for more information on the cohort, flags, counts and summary statistics

Terms of Use

For Academic use only.
By using this resource you agree to cite the following paper in your work:

"Exomic variants of an elderly cohort of Brazilians in the ABraOM database"
by Naslavsky, Yamamoto et al. is published in Human Mutation:
http://onlinelibrary.wiley.com/doi/10.1002/humu.23220/full

Signed consent forms from all subjects were obtained prior to this publication. No data from any given individual was shared. Collaborations are welcome. Contact us for any further information.

Centro de Pesquisa sobre o Genoma Humano e Células-Tronco

Support

CEPID FAPESP

inct institutos nacionais de ciência e tecnologia CNPq

Search Human (*Homo sapiens*)

Search all categories ▾ | Search Human... | Go

e.g. **BRCA2** or **17:64155265-64255266** or **rs1333049** or **osteoarthritis**

Genome assembly: GRCh37.p13 (GCA_000001405.14)

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh37.p13 coordinates
- Display your data in Ensembl

Other assemblies

GRCh38 (Ensembl release 93) | Go


View karyotype


Example region

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild, including RNASeq gene expression models
- Download genes, cDNAs, ncRNA, proteins (FASTA)
- Update your old Ensembl IDs

Example gene


Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

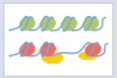
- More about comparative analysis
- Download alignments (EMF)


Example gene tree

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.

- More about the Ensembl regulatory build and microarray annotation
- Experimental data sources
- Download all regulatory features (GFF)
- Download regulatory evidence data features (BED)


Example regulatory feature

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

- More about variation in Ensembl
- Download all variants (GVF)
- Variant Effect Predictor

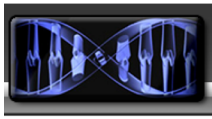
Example variant


Example phenotype


Example structural variant

Preditores de Patogenicidade

- SIFT;
- PolyPhen2;
- MutationTaster;
- GERP++;
- CADD;
- Trap.



PolyPhen-2 prediction of functional effects of human nsSNPs

Home | About | Help | Downloads | Batch query | WHESS.db

PolyPhen-2 (**Poly**morphism **Phen**otyping v**2**) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. Please, use the form below to submit your query.

Query Data

Protein or SNP identifier	
Protein sequence in FASTA format	
Position	
Substitution	AA_1 A R N D C E Q G H I L K M F P S T W Y V AA_2 A R N D C E Q G H I L K M F P S T W Y V
Query description	

Submit Query | Clear | Check Status

Display advanced query options

Software & web support: Ivan Adzhubey

Web design & development: biobyte solutions

Single nucleotide variant (SNV) lookup

This form allows you to quickly access the score (and annotation) of a single nucleotide variant (SNV) or all scores at a specific genomic position. If you are investigating multiple or even ranges of CADD SNV scores, please have a look at our Multi-SNV scoring form.

CADD scores are freely available for all non-commercial applications. If you are planning on using them in a commercial application, please contact us.

Chromosome: 22 **Position:** 43451447

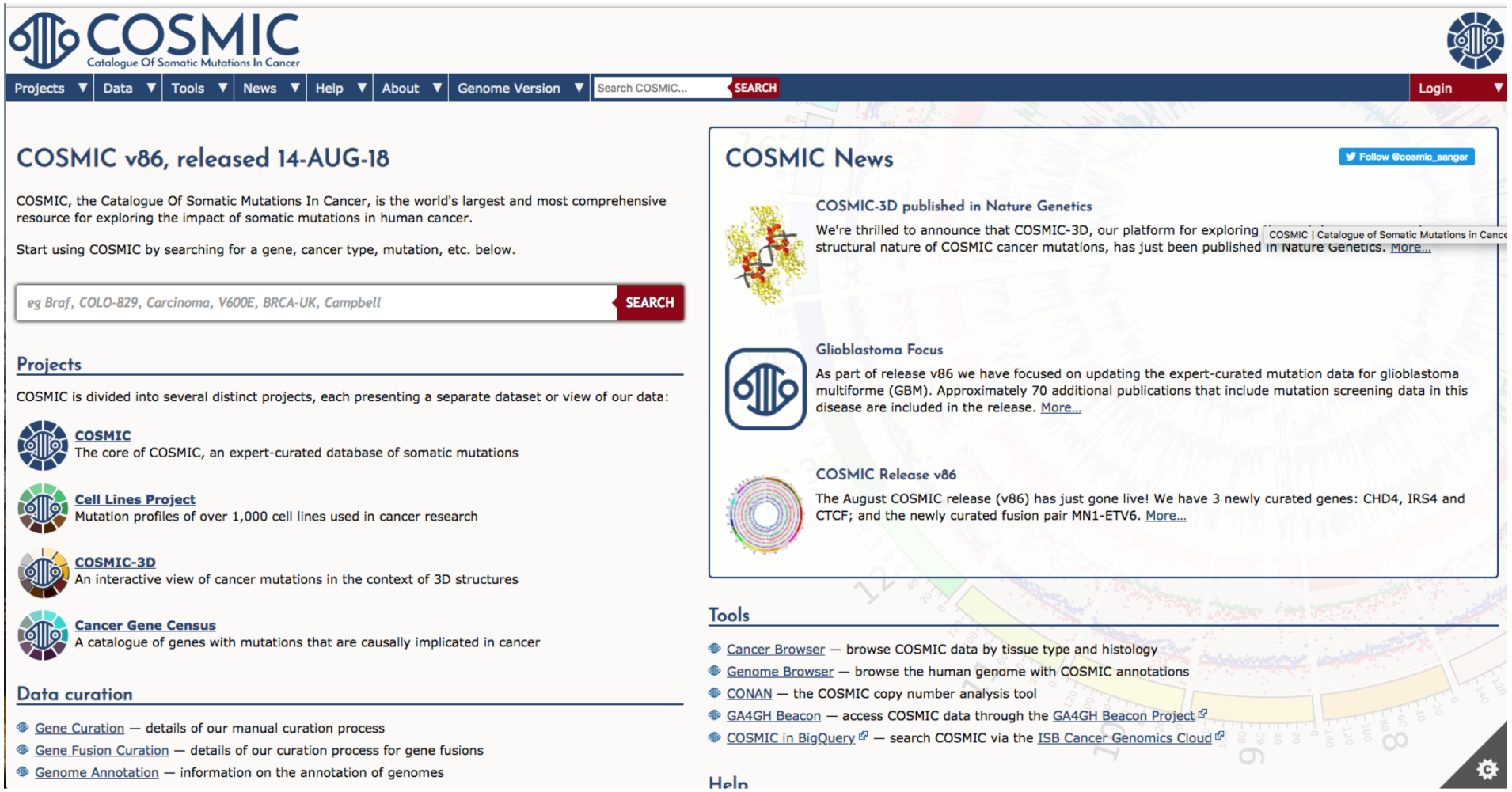
Ref (optional): T **Alt (optional):** A

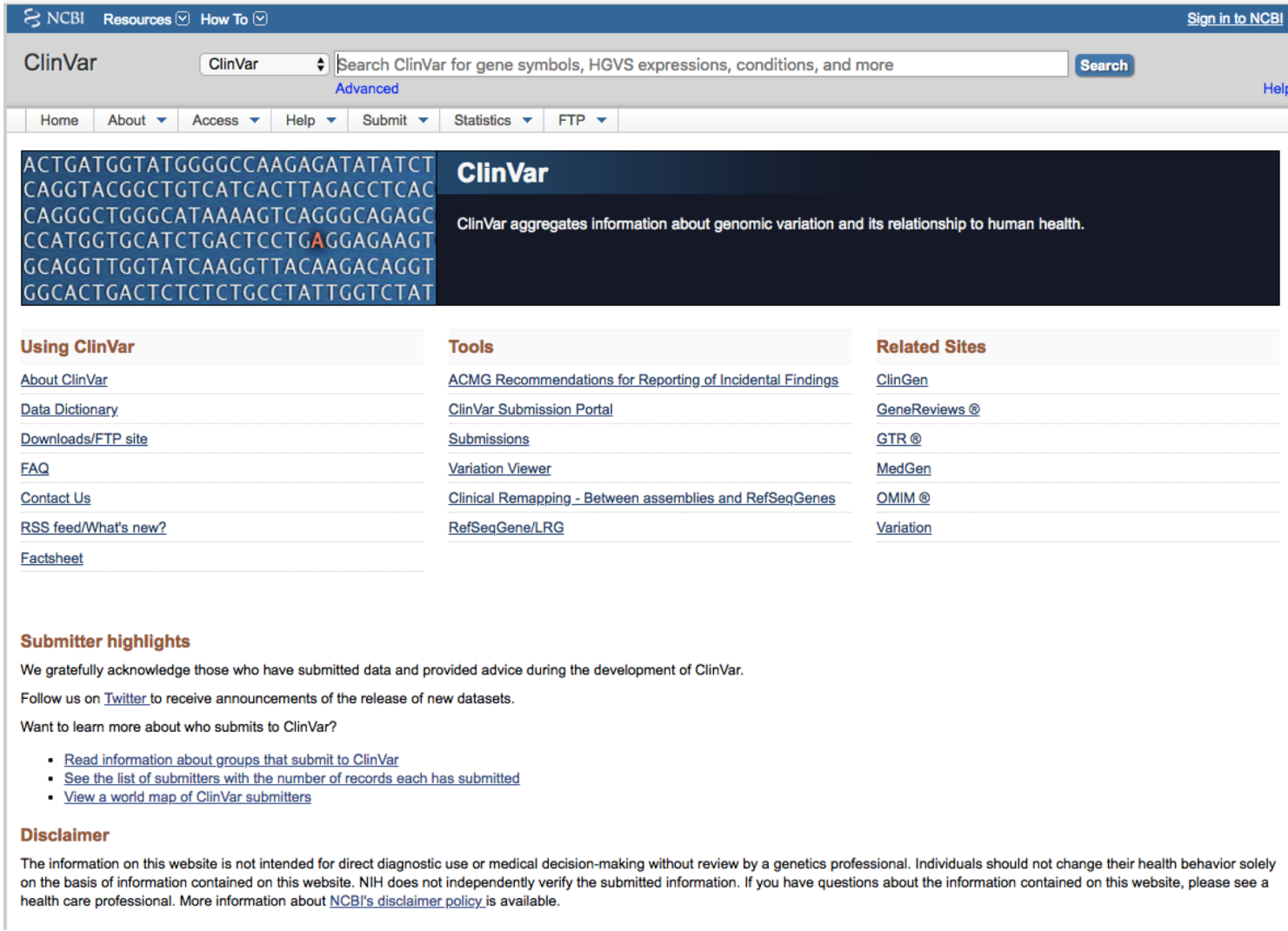
CADD model: GRCh37-v1.4 INCLUDE ANNOTATIONS TRANSPOSE TABLE

LOOKUP VARIANT(S)

Banco de Dados Fenotípicos

- COSMIC (Catalog Of Somatic Mutation in Cancer):
  - https://cancer.sanger.ac.uk/cosmic
- ClinVar:
  - https://www.ncbi.nlm.nih.gov/clinvar/
- OMIM (Online Mendelian Inheritance in Men):
  - https://omim.org/
- HGMD (Human Gene Mutation Database):
  - http://www.hgmd.cf.ac.uk/ac/index.php


COSMIC
Catalogue Of Somatic Mutations In Cancer
Projects
Data
Tools
News
Help
About
Genome Version
Search COSMIC...
SEARCH
Login
COSMIC v86, released 14-AUG-18
COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.
Start using COSMIC by searching for a gene, cancer type, mutation, etc. below.
eg Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell
SEARCH
Projects
COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:
COSMIC
The core of COSMIC, an expert-curated database of somatic mutations
Cell Lines Project
Mutation profiles of over 1,000 cell lines used in cancer research
COSMIC-3D
An interactive view of cancer mutations in the context of 3D structures
Cancer Gene Census
A catalogue of genes with mutations that are causally implicated in cancer
Data curation
Gene Curation — details of our manual curation process
Gene Fusion Curation — details of our curation process for gene fusions
Genome Annotation — information on the annotation of genomes
COSMIC News
Follow @cosmic_sanger
COSMIC-3D published in Nature Genetics
We're thrilled to announce that COSMIC-3D, our platform for exploring
COSMIC | Catalogue of Somatic Mutations in Cance
structural nature of COSMIC cancer mutations, has just been published in Nature Genetics. More...
Glioblastoma Focus
As part of release v86 we have focused on updating the expert-curated mutation data for glioblastoma multiforme (GBM). Approximately 70 additional publications that include mutation screening data in this disease are included in the release. More...
COSMIC Release v86
The August COSMIC release (v86) has just gone live! We have 3 newly curated genes: CHD4, IRS4 and CTCF; and the newly curated fusion pair MN1-ETV6. More...
Tools
Cancer Browser — browse COSMIC data by tissue type and histology
Genome Browser — browse the human genome with COSMIC annotations
CONAN — the COSMIC copy number analysis tool
GA4GH Beacon — access COSMIC data through the GA4GH Beacon Project
COSMIC in BigQuery — search COSMIC via the ISB Cancer Genomics Cloud
Help


NCBI
Resources
How To
Sign in to NCBI
ClinVar
ClinVar
Search ClinVar for gene symbols, HGVS expressions, conditions, and more
Search
Advanced
Help
Home
About
Access
Help
Submit
Statistics
FTP
ACTGATGGTATGGGGCCAAGAGATATATCT
CAGGTACGGCTGTCATCACTTAGACCTCAC
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC
CCATGGTGCATCTGACTCCTGAGGAGAAGT
GCAGGTTGGTATCAAGGTTACAAGACAGGT
GGCACTGACTCTCTCTGCCTATTGGTCTAT
ClinVar
ClinVar aggregates information about genomic variation and its relationship to human health.
Using ClinVar
About ClinVar
Data Dictionary
Downloads/FTP site
FAQ
Contact Us
RSS feed/What's new?
Factsheet
Tools
ACMG Recommendations for Reporting of Incidental Findings
ClinVar Submission Portal
Submissions
Variation Viewer
Clinical Remapping - Between assemblies and RefSeqGenes
RefSeqGene/LRG
Related Sites
ClinGen
GeneReviews ®
GTR ®
MedGen
OMIM ®
Variation
Submitter highlights
We gratefully acknowledge those who have submitted data and provided advice during the development of ClinVar.
Follow us on Twitter to receive announcements of the release of new datasets.
Want to learn more about who submits to ClinVar?
Read information about groups that submit to ClinVar
See the list of submitters with the number of records each has submitted
View a world map of ClinVar submitters
Disclaimer
The information on this website is not intended for direct diagnostic use or medical decision-making without review by a genetics professional. Individuals should not change their health behavior solely on the basis of information contained on this website. NIH does not independently verify the submitted information. If you have questions about the information contained on this website, please see a health care professional. More information about NCBI's disclaimer policy is available.

50 YEARS OMIM
Human Genetics Knowledge for the World

OMIM®

OMIM - Online Mendelian Inheritance in Man

Online Mendelian Inheritance in Man®

An Online Catalog of Human Genes and Genetic Disorders

Updated August 30, 2018

Search OMIM for clinical features, phenotypes, genes, and more...

Advanced Search : OMIM, Clinical Synopses, Gene Map

Need help? : Example Searches, OMIM Search Help

Mirror site : mirror.omim.org

OMIM is supported by a grant from NHGRI, licensing fees, and generous contributions from people like you.

Make a donation!

McKusick-Nathans Institute of Genetic Medicine

JOHNS HOPKINS MEDICINE

Follow us on Twitter

The Human Gene Mutation Database

at the Institute of Medical Genetics in Cardiff

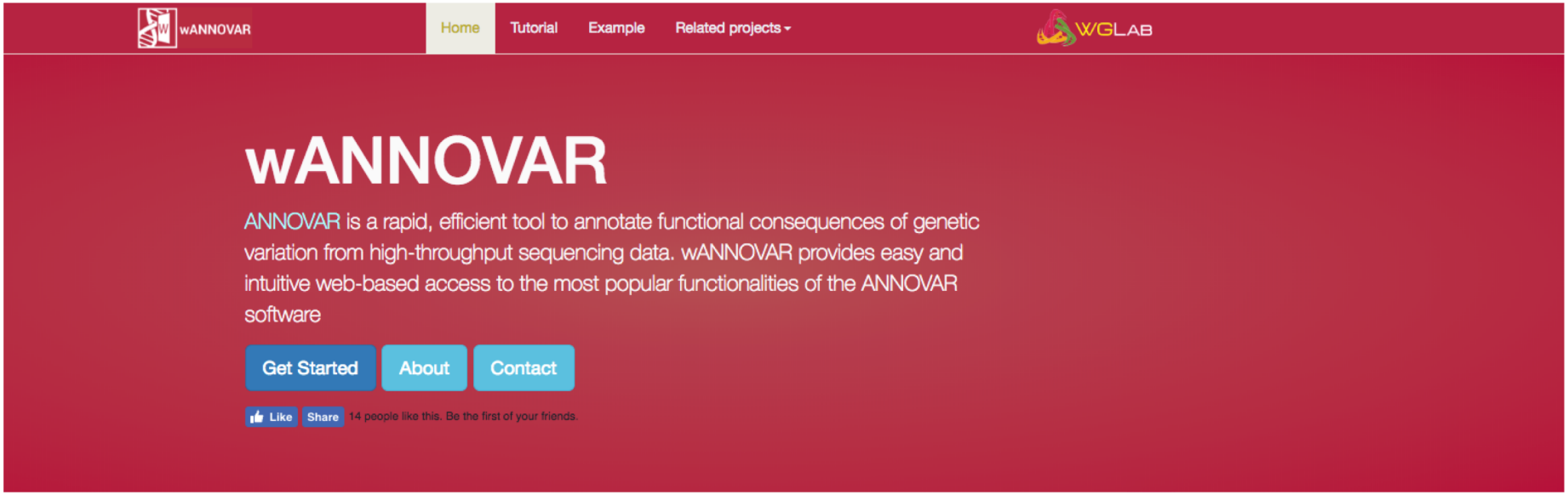
The Human Gene Mutation Database (**HGMD®**) represents an attempt to collate all known (published) gene lesions responsible for human inherited disease and is maintained in Cardiff by D.N. Cooper, E.V. Ball, P.D. Stenson, A.D. Phillips, K. Evans, S. Heywood, M.J. Hayden, M.M. Chapman, M.E Mort, L. Azevedo and M. Mort

Table:	Description:	Public entries: This site. Academic/non-profit users only	Total entries: HGMD Professional 2018.2
	Mutation totals (as of 2018-08-31)	**157131**	**224642**
Gene symbol	The gene description, gene symbol (as recommended by the HUGO Nomenclature Committee) and chromosomal location is recorded for each gene. In cases where a gene symbol has not yet been made official, a provisional symbol has been adopted which is denoted by lower-case letters.	6480	8784
cDNA sequence	cDNA reference sequences are provided, numbered by codon.	6471	8848
Genomic coordinates	Genomic (chromosomal) coordinates have been calculated for missense/nonsense, splicing, regulatory, small deletions, small insertions and small indels.	0	199696
HGVS nomenclature	Standard HGVS nomenclature has been obtained for missense/nonsense, splicing, regulatory, small deletions, small insertions and small indels.	0	200177
Missense/nonsense	Single base-pair substitutions in coding regions are presented in terms of a triplet change with an additional flanking base included if the mutated base lies in either the first or third position in the triplet.	87391	127200
Splicing	Mutations with consequences for mRNA splicing are presented in brief with information specifying the relative position of the lesion with respect to a numbered intron donor or acceptor splice site. Positions given as positive integers refer to a 3' (downstream) location, negative integers refer to a 5' (upstream) location.	14329	20132
Regulatory	Substitutions causing regulatory abnormalities are logged in with thirty nucleotides flanking the site of the mutation on both sides. The location of the mutation relative to the transcriptional initiation site, initiaton codon, polyadenylation site or termination codon is given.	3050	4029
Small deletions	Micro-deletions (20 bp or less) are presented in terms of the deleted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	23686	33090
Small insertions	Micro-insertions (20 bp or less) are presented in terms of the inserted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	9844	13864
Small indels	Micro-indels (20 bp or less) are presented in terms of the deleted/inserted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	2289	3088
Gross deletions	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	11705	16645
Gross insertions	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	2806	4117
Complex rearrangements	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	1576	1960
Repeat variations	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	455	517

7,437,538 queries successfully served since 2007.

ANNOVAR

- Software open source usado para anotação de VCF;
- http://annovar.openbioinformatics.org/en/latest/
- Versão online disponível wANNOVAR:
  - http://wannovar.wglab.org/
- Capacidade de annotar diferentes bases de dados em diferentes versões do genoma;
- Script escrito em Perl que permite a execução muiltplataforma;
- Permite a integração com softwares de terceiros.



Basic Information

Email Email

Sample Identifier Sample Identifier

Input File + Input File

or Paste Variant Calls paste your variant call here

Recent Updates

[10/19/2017] The detailed amino acid changes for indels are now included in the output (through -polish argument in table_annovar). The server also handles duplicated entries (multiple identical variants) in the input file correctly.

[08/25/2017] The variants reduction method (for disease gene finding from personal genomes)

Build	Table Name	Explanation	Date
hg18	refGene	FASTA sequences for all annotated transcripts in RefSeq Gene	20170601
hg19	refGene	same as above	20170601
hg38	refGene	save as above	20170601
hg18	knownGene	FASTA sequences for all annotated transcripts in UCSC Known Gene	20170601
hg19	knownGene	same as above	20170601
hg38	knownGene	same as above	20170601
hg18	ensGene	FASTA sequences for all annotated transcripts in ENSEMBL Gene	20170601
hg19	ensGene	same as above	20170601
hg38	ensGene	based on Gencode v26 Basic collection	20170912

hg18	dbnsfp33a	whole-exome SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, MetaSVM, MetaLR, VEST, M-CAP, CADD, GERP++, DANN, fathmm-MKL, Eigen, GenoCanyon, fitCons, PhyloP and SiPhy scores from dbNSFP version 3.3a	20170221
hg19	dbnsfp33a	same as above	20170221
hg38	dbnsfp33a	same as above	20170221

hg19	cosmic64	COSMIC database version 64	20130520
hg19	cosmic65	COSMIC database version 65	20130706
hg19	cosmic67	COSMIC database version 67	20131117
hg19	cosmic67wgs	COSMIC database version 67 on WGS data	20131117
hg19	cosmic68	COSMIC database version 68	20140224
hg19	cosmic68wgs	COSMIC database version 68 on WGS data	20140224
hg19	cosmic70	same as above	20140911

hg19	clinvar_20180603	Clinvar version 20180603 with separate columns (CLNALLELEID CLNDN CLNDISDB CLNREVSTAT CLNSIG)	20180708
hg38	clinvar_20180603	same as above	20180708

hg19	exac03	ExAC 65000 exome allele frequency data for ALL, AFR (African), AMR (Admixed American), EAS (East Asian), FIN (Finnish), NFE (Non-finnish European), OTH (other), SAS (South Asian)). version 0.3. Left normalization done.	20151129
hg18	exac03	same as above	20151129
hg38	exac03	same as above	20151129

hg19	gnomad_exome	gnomAD exome collection	20170311
hg38	gnomad_exome	gnomAD exome collection	20170311
hg19	gnomad_genome	gnomAD genome collection	20170311
hg38	gnomad_genome	gnomAD genome collection	20170311

hg19	abraom	2.3 million Brazilian genomic variants	20180312
hg38	abraom	liftOver from above	20180312

Atividade prática de anotação de VCF

- Construção do comando para download de base de dados adicionais;
- Converter arquivo VCF para o formato annovar;
- Construção do comando para anotação.

Comando para obter Base de Dados

/data/annovar/annotate_variation.pl -downdb -webfrom annovar -buildver VERSÃO_DO_GENOMA BASE_DE_DADOS /data/annovar/humandb/

/data/annovar/annotate_variation.pl -downdb -webfrom annovar -buildver hg38 refGene /data/annovar/humandb/

BAIXAR BASE DA DODOS abraom HG38?

Comando para obter Base de Dados

/data/annovar/annotate_variation.pl -downdb -webfrom annovar -buildver VERSÃO_DO_GENOMA BASE_DE_DADOS /data/annovar/humandb/

/data/annovar/annotate_variation.pl -downdb -webfrom annovar -buildver hg38 refGene /data/annovar/humandb/

BAIXAR BASE DA DODOS abraom HG38?

/data/annovar/annotate_variation.pl -downdb -webfrom annovar -buildver hg38 abraom /data/annovar/humandb/

Comando para converter VCF em annovar

- Converter o arquivo HG00096.vcf em HG00096.avinput

/data/annovar/convert2annovar.pl -format vcf4 /data/renan/HG00096.vcf --includeinfo -withzyg -out /data/renan/HG00096.vcf.avinput

Comando para anotar BD no annovar

- Anotar as seguintes bases de dados no arquivo HG00096.avinput:
  - refGene
  - exac03
  - abraom
  - dbnsfp35a

/data/annovar/table_annovar.pl HG00096.avinput /data/annovar/humandb/ -buildver hg38 -out HG00096 -remove -protocol refGene,exac03,abraom, dbnsfp35a -operation g,f,f,f