Supplementary documentation

unCOVERApp documentation

Index

1 unCOVERApp

unCOVERApp allows users to visualize and annotate low-coverage genomic regions containing genes in sequencing data. In particular, users can obtain:

- interactive graphical DoC analysis from whole gene(s) to base-pair resolution

- clinically and functionally annotations of low-coverage site downloadable in spreadsheet format

- Calculator of maximum credible population Allele Frequency to allow user-defined AF thresholds rather gnomAD gnomAD generic filter.

- 95 % probability of the binomial distribution based on an expected allele fraction (probability of success) and a minimum number of variant reads (number of successes) for somatic variants.

The code of App is available on GitHub here.

2 Prerequisites

This app requires following dependencies:

- **samtools v.1.9**

- **R v.3.5.1** or **RStudio**, and run Rscript to set up the environment.

3 Instructions

To run locally unCOVERApp, users can clone or download unCOVERApp repository. Annotation files can be downloaded from googledrive and positioned in unCOVERApp directory. The md5sum of the bed and bed.tbi files can be retrieve in repository.

```
git clone https://github.com/Manuelaio/unCOVERApp.git
```

unCOVERApp directory must retain the following tree structure.

```
├── CONTACTS.md
├── LICENSE.md
├── README.md
├── dependencies.R
├── intro.md
├── md5sum.txt
├── server-annotation.R
├── server-binomial.R
├── server-maxAF.R
├── server-plots.R
├── server-reactiveDF.R
├── server-tables.R
├── sorted.bed.gz
├── sorted.bed.gz.tbi
├── uncoverApp.R
└── www
    ├── contact.png
    ├── make_bed.sh
    ├── preprocessing.R
    ├── script.js
    └── uncoverapp.config
```

Figure 1: Tree structure of unCOVERApp folder

4 Input file preparation

unCOVERApp input file is a BED file (tab-separated) containing depth of coverage (DoC) for each genomic position (one per row) of target genes for as many samples as many BAM files are listed in the ".list" file. In order to easily obtain a input format, users follow the instructions below:

- **write a file with ".txt"** extension containing HGNC gene name(s) (one per row)

- **write a file with ".list"** extension containing absolute path of BAM(s) file (one per row)

- **compile** a configuration file specifying absolute path of: unCOVERApp folder, txt file containing HGNC gene name(s), list file containing absolute path of BAM(s) and folder output location. Compile genome reference and chromosome notation BAM. (number refers to 1, 2, ... X,.M notation BAM, chr refers to chr1, chr2,... chrX, chrM notation BAM).

The following image shows an example of compiled configuration file.

```
########################################

pathscript=/myworkspace/unCOVERApp/www/
geneList=/myworkspace/my_sample/gene.txt
genome=hg19 #or hg38
notation_bam=number #or chr for Bam chromosome Notation
bamList=/myworkspace/my_sample/bam.list
output=/myworkspace/my_sample/

########################################
```

Once users have compiled the configuration file, run the following command through command line

```
bash www/make_bed.sh www/uncoverapp.config
```

The log file is a trouble shooter, so please revise when any problem happens .

Bash script creates a new directory named with current date in users-defined location, inside is stored input file named **multisample.bed.gz file**.

5 unCOVERApp example

Users can run the shiny app with just one command in R:

```
library(shiny)
runApp('uncoverApp.R')
```

The following example shows how unCOVERApp works and how it helped us to identify pathogenic low coverage position within a candidate gene, POLG, starting from of negative exome sequencing result.

Using bash attached script we have prepared a bed file containing the base-pair DoC across the POLG. We wrote a gene.txt file which contains HGNC official gene name, a file with ".list" extension containing absolute path to BAM sample and we had setup a configuration file.

The first page of unCOVERApp, **Coverage Analysis**, is shown in following figure. Firstly, just made input file was loaded in `Select input box` and visualized in `bed file` table.

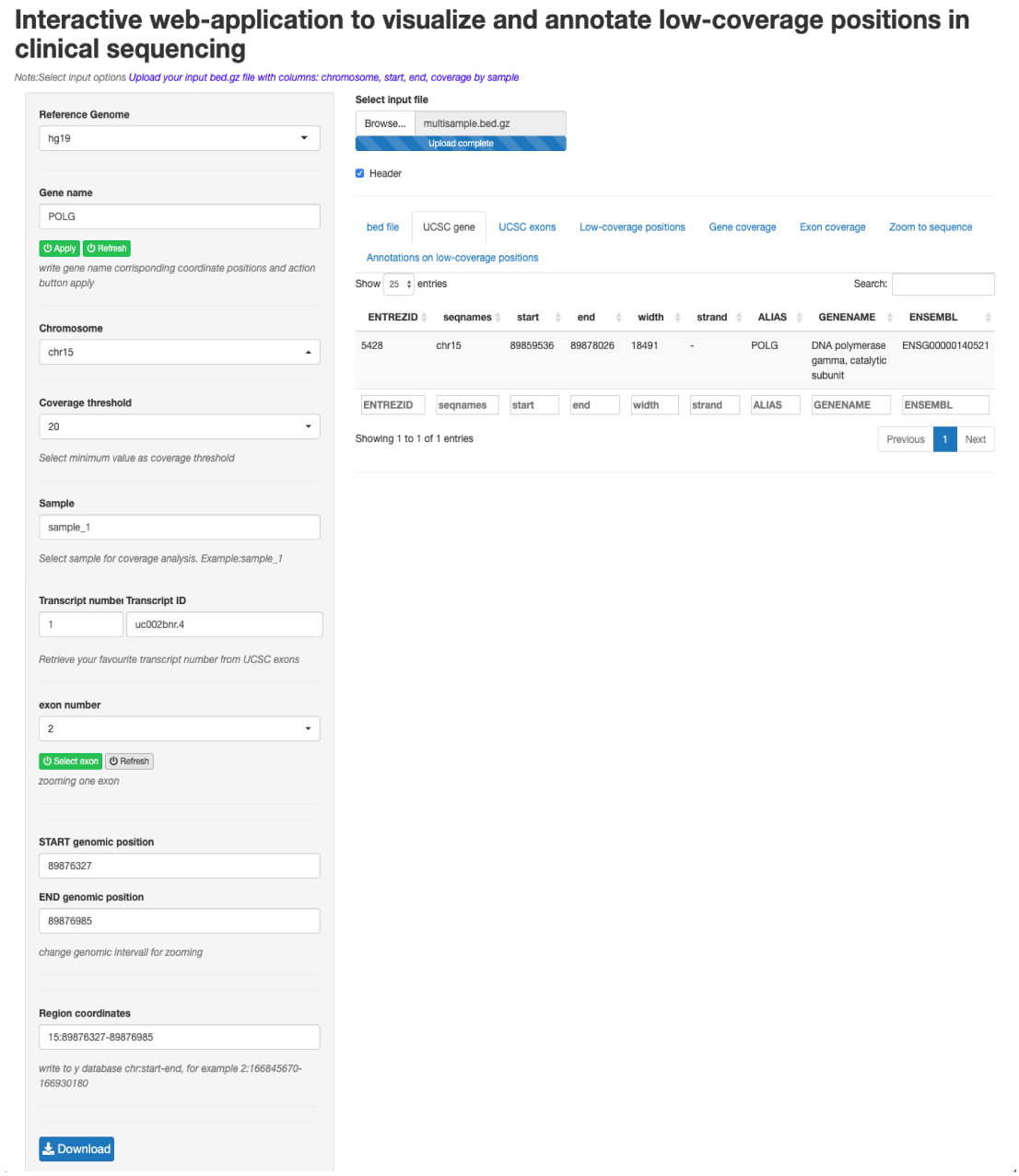


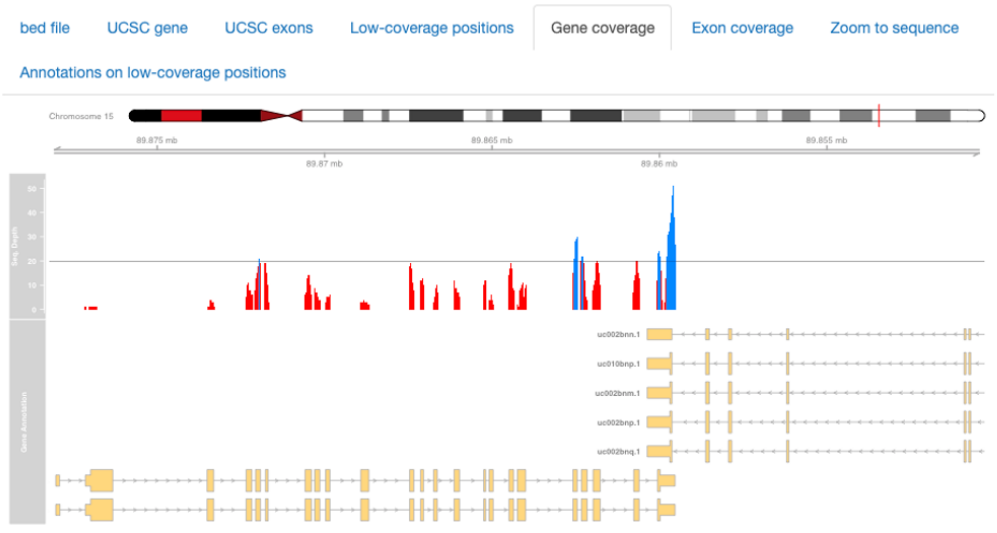
Figure 2: The figure shows first page of unCOVERApp using for coverage analysis in which all required input are filled. As it shows, all required inputs are located in sidebar on the left one by one.

Filling inputs as **Reference genome** and **gene name** unCOVERApp returns two outputs:

- `UCSC gene` table, that returns genome coordinates, chromosome number and other useful information based on gene user-defined

- `UCSC exons` table, that returns genome coordinates of each exon for each transcript.

Based on informations provided in outputs, we filled `Chromosome` box and `transcript number` box, moreover we have chosen a coverage threshold and the sample to analyze. Then, unCOVERApp return a plot for graphical inspection of POLG DoC in `Gene coverage` box and a related table with the number of low-coverage positions in each exon given a transcript user-defined. (Figure 3)

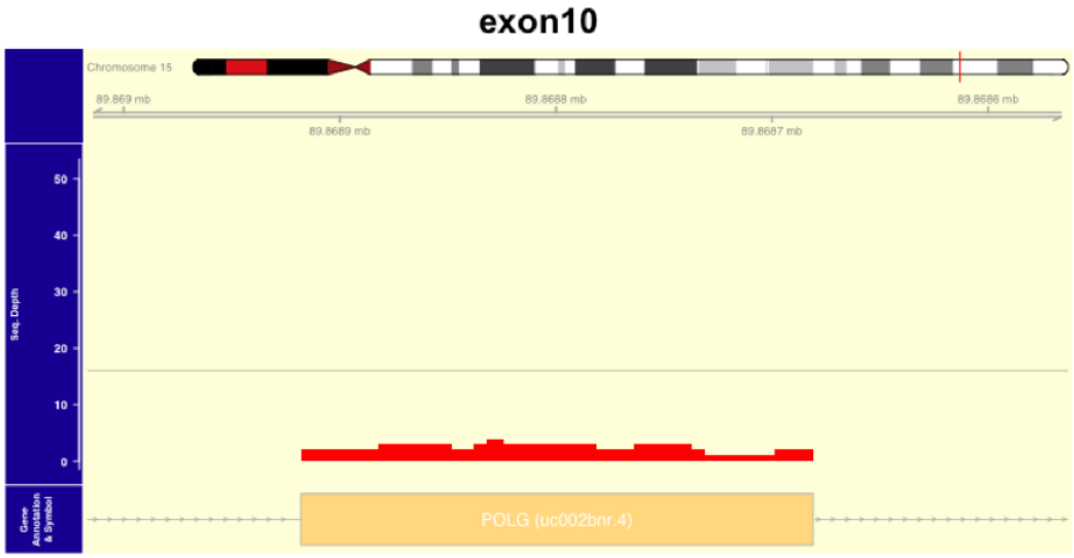


exon	ucovered positions
2	153
3	176
4	168
5	142
6	80
7	183

Figure 3: The figure shows DoC of POLG. On top panel it is viewed information about chromosome, genome coordinates and below a DoC information in form of histogram with a dynamically drawn line given a users-threshold cuff off and lastly the different gene transcripts tracks. The table shows the number of uncovered positions for each exon given a chosen transcript.

Table and graphical inspection had shown that the majority of POLG exons are uncovered. Moreover, unCOVERApp provides two zoom function in order to expand the histogram plot in to user-selected intervals, from exon (`exon number` box) to base-pair level (`zoom to sequence`). Inspecting each exon, we have found some low-DoC positions with functional e clinical annotations in exon 10. (Figure 4)

This is a Gviz function and it plots exon with ideogram, genome coordinates, coverage information, Ensembl and UCSC gene annotation. The annotation for the databases are directly fetched from Ensemb and all tracks will be plotted in a 3' -> 5' direction.



Show 25 entries Search:

Low coverage positions in ClinVar	Low coverage positions with High or Medium impact
31	351

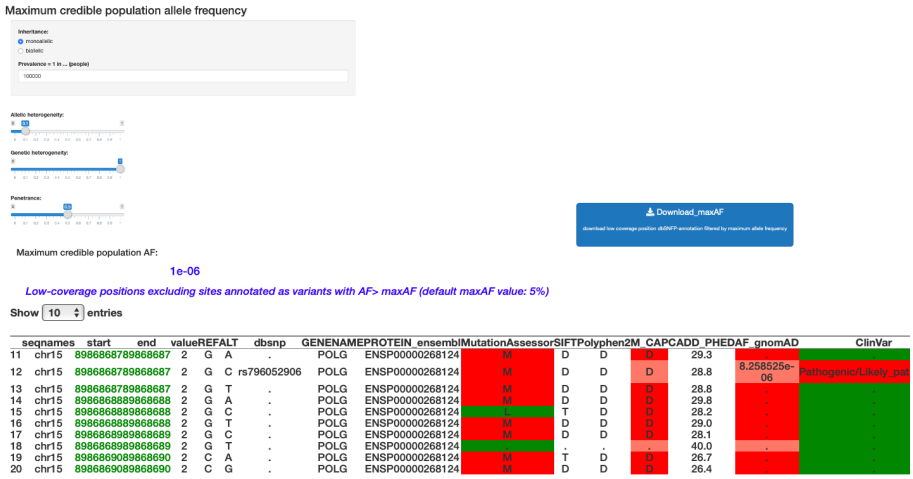
Figure 4: The figure shows DoC of exon 10, instead the table summarizes the number of positions known in ClinVar and with a High or medium impact.

dbNSFP-based annotation of all potential nucleotide changes across low-DoC POLG are available in `Annotation on low-coverage positions` box. The output is a downloadable table (Figure 5) displaying low-DoC positions at base-pairs level in which cells highlighted according to several criteria as high impact, clinical annotation (ClinVar), low gnomAD allelic frequency (<0.5), a damaging M-CAP score and CADD-score <20. Moreover, a low-DoC genomic position is yellow highlighted when a damaging score is found in all considered dbSNP-predictor.

seqnames	start	end	value	REF	ALT	dbsnp	GENENAME	PROTEIN_ensembl	MutationAssessor	SIFT	Polyphen2	M_CAP	CADD_PHRED	AF_gnomAD	ClinVar	clinvar_MedGen_id	clinvar_OMIM_id	HGVSc_VEP	HGVSp_VEP
chr15	89868743	89868743	3	G	T	.	POLG	ENSP00000268124	L	T	B	D	22.8	.	.	.	.	c.1887C>A	p.Asp629Glu
chr15	89868744	89868744	3	T	A	.	POLG	ENSP00000268124	M	D	P	D	33.0	.	.	.	.	c.1886A>T	p.Asp629Val
chr15	89868744	89868744	3	T	C	.	POLG	ENSP00000268124	M	D	B	D	27.5	.	.	.	.	c.1886A>G	p.Asp629Gly
chr15	89868744	89868744	3	T	G	rs1039182766	POLG	ENSP00000268124	M	D	P	D	32.0	4.020618e-06	Uncertain_significance	CN169374	.	c.1886A>C	p.Asp629Ala
chr15	89868745	89868745	3	C	A	.	POLG	ENSP00000268124	M	D	D	D	29.5	.	.	.	.	c.1885G>T	p.Asp629Tyr
chr15	89868745	89868745	3	C	G	.	POLG	ENSP00000268124	M	D	P	D	29.1	.	Uncertain_significance	C0205710	203700	c.1885G>C	p.Asp629His
chr15	89868745	89868745	3	C	T	.	POLG	ENSP00000268124	M	T	B	D	25.0	.	.	.	.	c.1885G>A	p.Asp629Asn
chr15	89868747	89868747	3	C	A	.	POLG	ENSP00000268124	M	T	B	D	22.6	.	.	.	.	c.1883G>T	p.Arg628Leu
chr15	89868747	89868747	3	C	G	rs201871736	POLG	ENSP00000268124	L	T	P	D	23.7	.	.	.	.	c.1883G>C	p.Arg628Pro
chr15	89868747	89868747	3	C	T	rs201871736	POLG	ENSP00000268124	L	T	B	D	21.9	4.014871e-06	Likely_benign	CN169374	.	c.1883G>A	p.Arg628Gln

Figure 5: Low-DoC positions annotated with dbNSPF. dbSNP-annotation collects all consequences found in VEP-defined canonical transcripts.

However, in `Coverage Analysis` page, the default AF threshold for a variant to be annotated is 5%, then we used calculator of max AF available in `calculate AF by allele frequency app` in order to use a allele frequency based on genetic architecture of observed disease.



	seqnames	start	end	value	REF	ALT	dbsnp	GENENAME	PROTEIN_ensembl	MutationAssessor	SIFT	Polyphen2	M_CAP	CADD_PHRED	AF_gnomAD	ClinVar
11	chr15	89868678	89868687	2	G	A	.	POLG	ENSP00000268124	M	D	D	D	29.3	.	
12	chr15	89868678	89868687	2	G	C	rs796052906	POLG	ENSP00000268124	M	D	D	D	28.8	8.258525e-06	Pathogenic/Likely_pat
13	chr15	89868678	89868687	2	G	T	.	POLG	ENSP00000268124	M	D	D	D	28.8	.	
14	chr15	89868688	89868688	2	G	A	.	POLG	ENSP00000268124	M	D	D	D	29.8	.	
15	chr15	89868688	89868688	2	G	C	.	POLG	ENSP00000268124	L	T	D	D	28.2	.	
16	chr15	89868688	89868688	2	G	T	.	POLG	ENSP00000268124	M	D	D	D	29.0	.	
17	chr15	89868689	89868689	2	G	C	.	POLG	ENSP00000268124	M	D	D	D	28.1	.	
18	chr15	89868698	89868689	2	G	T	.	POLG	ENSP00000268124	L	.	.	.	40.0	.	
19	chr15	89868690	89868690	2	C	A	.	POLG	ENSP00000268124	M	T	D	D	26.7	.	
20	chr15	89868690	89868690	2	C	G	.	POLG	ENSP00000268124	M	D	D	D	26.4	.	

Figure 6: unCOVERApp allows to draw AF thresholds based on genetic architecture of condition through integration of Calculator of maximum credible population Allele Frequency.

Analyzing functional and clinical annotation reports has allowed us to find a interesting low DoC position that hidden a probably causative variant. IGV graphical inspection had shown clinical relevant alternative allele in that low-covered position later confirmed by Sanger sequencing.

Importantly, unCOVERApp supports a binomial calculator expressing the probability that a variants is missed given its expected allelic fraction and sequencing coverage. Actually, the 20x minimum DoC threshold is reasonable for germline events where the expected fraction of variant alleles is around 0.5. Conversely, adequate DoC to detect somatic variants is paramount as that fraction can be substantially lower. unCOVERApp `Binomial distribution` page provides a simple statistical framework to evaluate if DoC is adequate to somatic variant detection. The user can set the allele fraction expected for the disease-related variant and the number of variant reads necessary to support variant calling.

In the below example, it unCOVERApp exploits binomial distribution to understand, with 95% probability, the number of reads that support a somatic variants given an `allele fraction` and the minimum number of `variant reads` required to support variant calling. The outcome in consideration box is marked in red when calculated number of reads are lower than variants reads user-defined(Figure 7), otherwise the letters are blue.

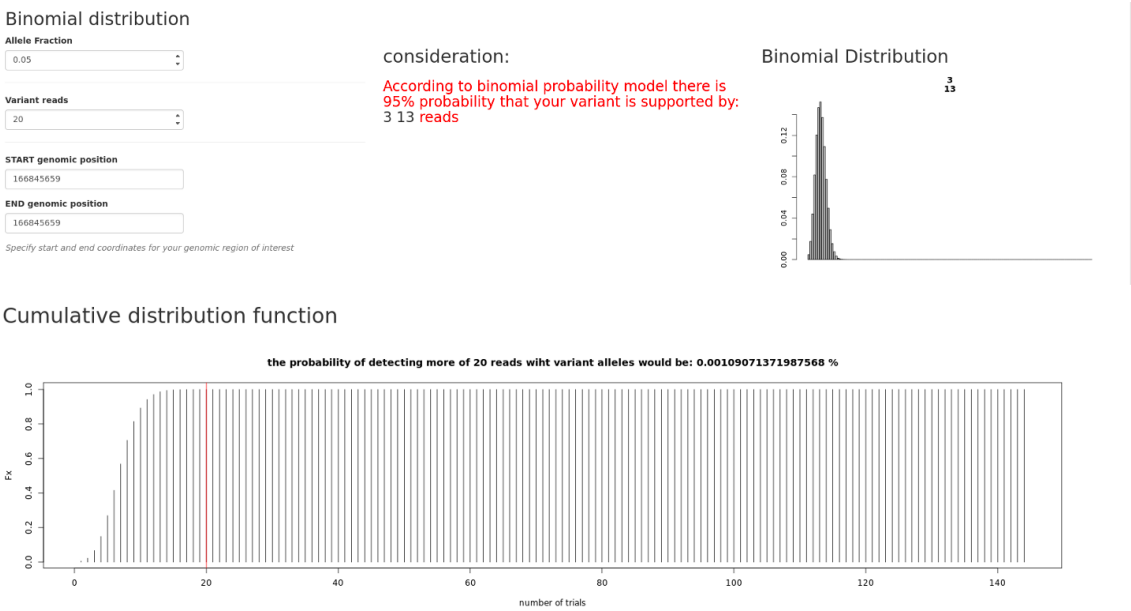


Figure 7: The figure shows binomial distribution analysis. The box consideration shows the number of reads that support variants based on users-defined inputs. Cumulative distribution function shows the probability of detecting less than or equal reads to the expected fraction of variant reads (probability of success) in user-defined reads (n trials).