Supplementary documentation

unCOVERApp documentation

June 14, 2020
v1.0.0

Index

1 unCOVERApp

unCOVERApp is a shiny graphical application for clinical assessment of sequence coverage. unCOVERApp allows:

- to display interactive plots showing sequence gene coverage down to base-pair resolution and functional/ clinical annotations of sequence positions within coverage gaps (**Coverage Analysis** page)
- to calculate the maximum credible population allele frequency (AF) to be applied as AF filtering threshold tailored to the model of the disease-of-interest instead of a general AF cut-off (e.g. 1 % or 0.1 %) (**Calculate AF by allele frequency app** page)
- to calculate the 95 % probability of the binomial distribution to observe at least N variant-supporting reads (N is the number of successes) based on a user-defined allele fraction that is expected for the variant (which is the probability of success). Especially useful to obtain the range of variant-supporting reads that is most likely to occur at a given DoC (which is the number of trials) for somatic variants with low allele fraction (**Binomial distribution** page).

2 Prerequisites

Install **unCOVERApp** by downloading the GitHub repository. It requires:

- **R** version $>= 3.5$
- **java** installed
- The annotation files can be downloaded from Zenodo and must be loaded on the R environment before launching unCOVERApp. Alternatively, unCOVERApp can be installed as R package in development version.

The final tree of directory unCOVERpp is as follows:

```
.
├── CONTACTS.md
├── LICENSE.md
├── README.md
├── intro.md
├── preprocessing.md
├── script
│   ├── POLG.example.bed
│   ├── Rpreprocessing.R
│   ├── contact.png
│   ├── dependencies.R
│   ├── example_POLG.bam
│   ├── example_POLG.bam.bai
│   ├── gene.txt
│   └── script.js
├── server-annotation.R
├── server-binomial.R
├── server-maxAF.R
├── server-plots.R
├── server-preprocess.R
├── server-reactiveDF.R
├── server-tables.R
├── server.R
├── sorted.bed.gz
├── sorted.bed.gz.tbi
└── ui.R
```

Figure 1: Tree structure of unCOVERApp folder

3 Instructions

All unCOVERApp functionalities are based on the availability of a BED-style formatted input file containing tab-separated specifications of genomic coordinates (chromosome, start position, end position), the coverage value, and the reference:alternate allele counts for each position. In the first page **Preprocessing**, users can

prepare the input file by specifying the genes to be examined and the BAM file(s) to be inspected. Users should be able to provide:

- a text file, named with .txt extension, containing HGNC official gene name(s) one per row, that is to be uploaded on `Load a gene(s) file` box. An example of text file is included in **script** directory `mygene.txt`.

- a text file, named with .list extension, containing the absolute paths to BAM file(s) one per row, that is to be uploaded on `Load bam file(s) list` box. In the output file, sample 1,2,3,.. correspond to the samples to which BAM files written in rows 1,2,3,... of the .list-extended file. An example BAM file is included in the repository. Users can move to the unCOVERApp directory and follow the commands below to retrieve the BAM file absolute path and write it in the `bam.list` file.

```
bam.path= paste(getwd(),"/script/example_POLG.bam", sep = "")
write.table(bam.path, file= "./script/bam.list", quote= F, row.names = F, col.names = F)
```

4 Usage

Open RStudio and set-up the R environment with Rscript dependencies.R . To run unCOVERApp, do the following steps to open the shiny app in your default browser:

```
library(shiny)
runApp()
```

In the first page, Preprocessing, users should load `mygene.txt` in: `Load a gene(s) file` and `bam.list` in: `Load bam file(s) list`.

Users should also specify the reference genome in `Genome` box and the chromosome notation of their BAM file(s) in `Chromosome Notation` box. In the BAM file, the `number` option refers to 1, 2, ..., X,.M chromosome notation, while the `chr` option refers to chr1, chr2, ... chrX, chrM chromosome notation. Users can specify the **minimum mapping quality (MAPQ)** value in `minum Mapping Quality (MAPQ)` box and **minimum base quality (QUAL)** value in `minimum Base Quality` box. Default values for both mapping and base qualities is 1. To run the example, choose `chr` chromosome notation, `hg19` genome reference and leave minimum mapping and base qualities to the default settings, as shown in the following screenshot of the Preprocessing page:

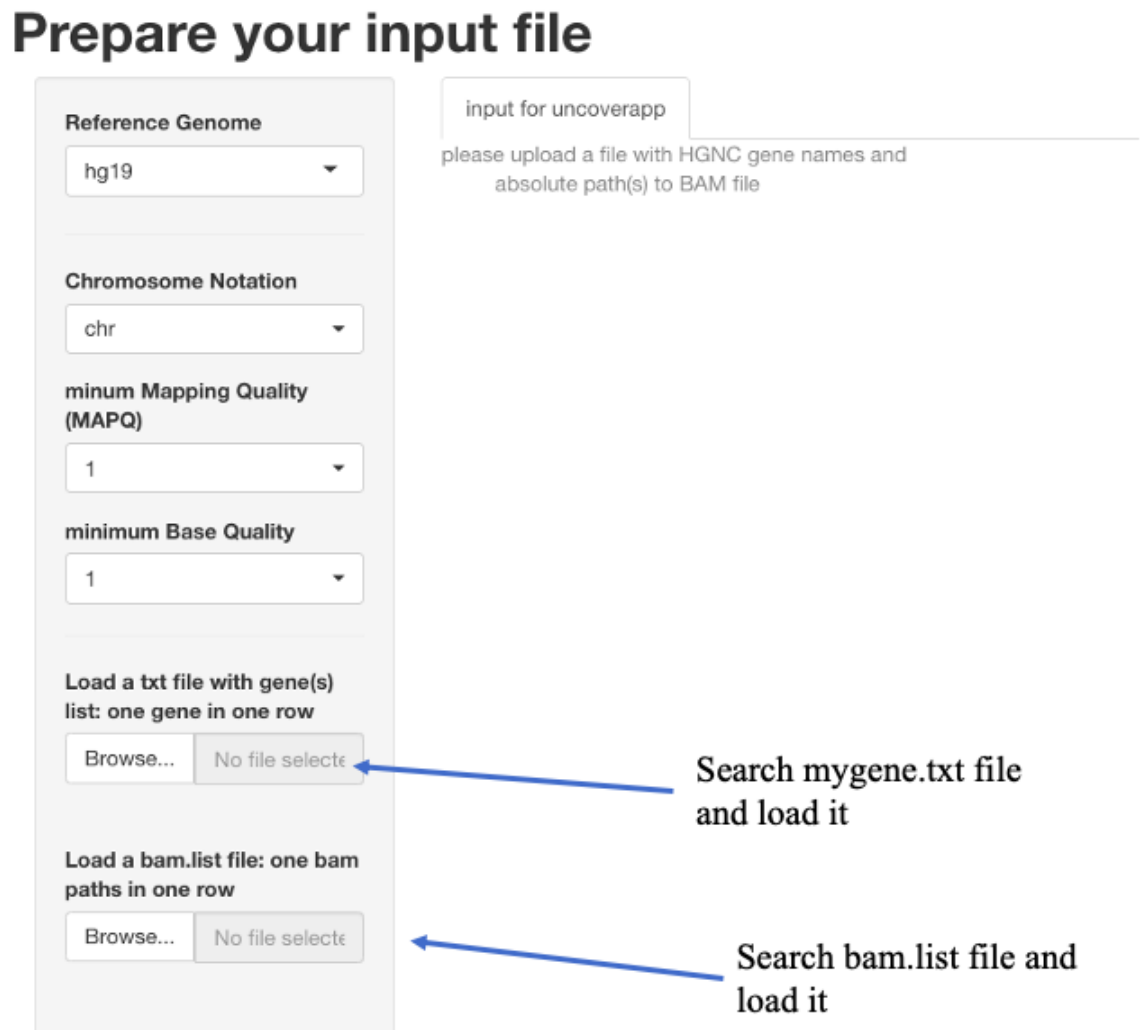


Figure 2: Screenshot of Preprocessing page.

unCOVERApp input file generation fails if incorrect gene names are specified. An "unrecognized gene name(s)" table is displayed if such a case occurs.

Below is a snippet of a the unCOVERApp input file generated as a result of the preprocessing step performed for the example:

```
chr15   89859516        89859516        68      A:68
chr15   89859517        89859517        70      T:70
chr15   89859518        89859518        73      A:2;G:71
chr15   89859519        89859519        73      A:73
chr15   89859520        89859520        74      C:74
chr15   89859521        89859521        75      C:1;T:74
```

The preprocessing time depends on the size of the BAM file(s) and on the number of genes to investigate. In general if many (e.g. > 50) genes are to be analysed, we would recommend to download the Rscript from the unCOVERApp **Preprocessing** page and run it separately. Alternatively, other tools do a similar job and can be used to generate the unCOVERApp input file (for instance: bedtools, samtools, gatk). Then users can upload the unCOVERApp input file directly on **Coverage Analysis** page in `Select input file` box. Once preprocessing is done, users can move to the Coverage Analysis page and push the `load prepared input file` button.

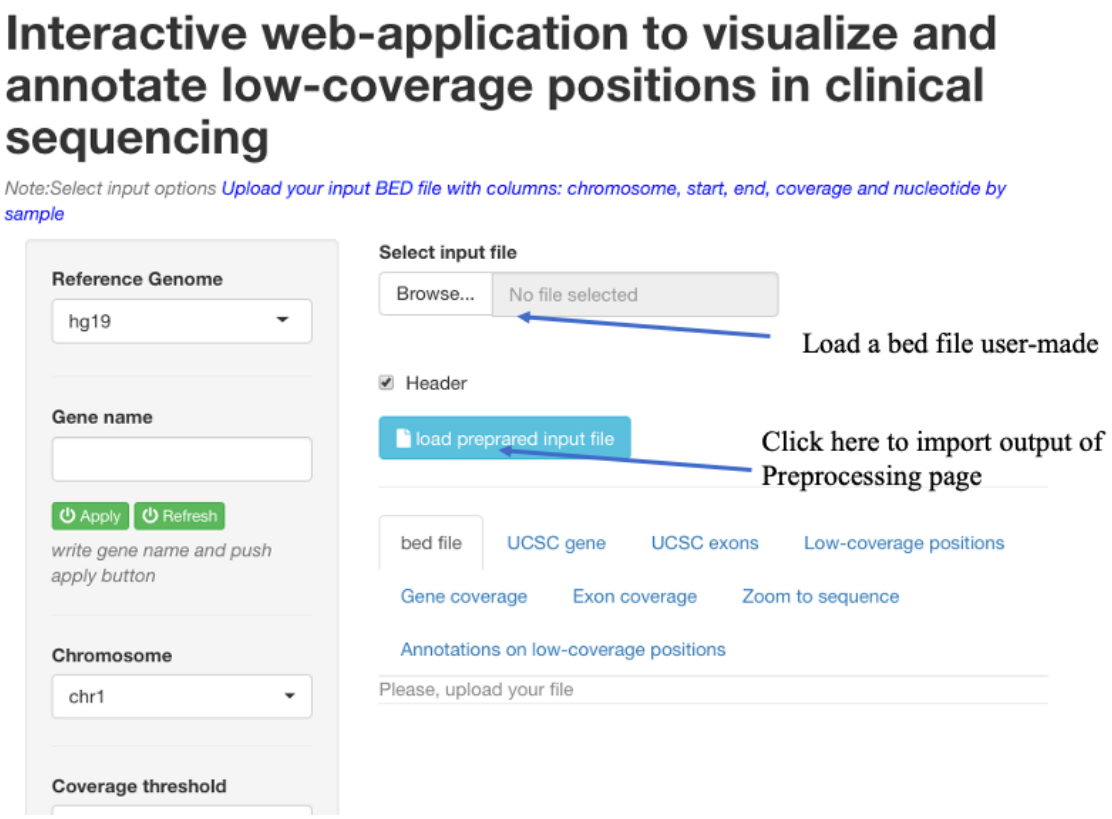


Figure 3: Screenshot of Coverage Analysis Page.

To assess sequence coverage of the example, the following **input** parameters must be specified in the sidebar of the **Coverage Analysis** section:

- `Reference Genome`: reference genome (hg19 or hg38); choose hg19
- `Gene name` and push `Apply` button: write the HGNC official gene name **POLG**
- `Chromosome`: The chromosome on which the gene is locate. choose chr15
- `coverage threshold` : specify coverage threshold (e.g. 20x)
- `Sample` : sample to be analyzed, choose 1
- `Transcript number` : transcript number, choose 1 item `exon number` : option to zoom in a specific exon, choose 10

  Other input sections, such as `Transcript ID`, `START genomic position`, `END genomic position` and `Region coordinate`, are dynamically filled.

unCOVERApp generates the following **outputs**:

- unfiltered BED file in `bed file` and the corresponding data-set pruned of high-coverage positions in `Low coverage positions`

- information about POLG in `UCSC gene` table

	ENTREZID	seqnames	start	end	width	strand	ALIAS	GENENAME	ENSEMBL
1	5428	chr15	89859536	89878026	18491	-	POLG	DNA polymerase gamma, catalytic subunit	ENSG00000140521

Showing 1 to 1 of 1 entries Previous 1 Next

Figure 4: Screenshot of output of UCSC gene table.

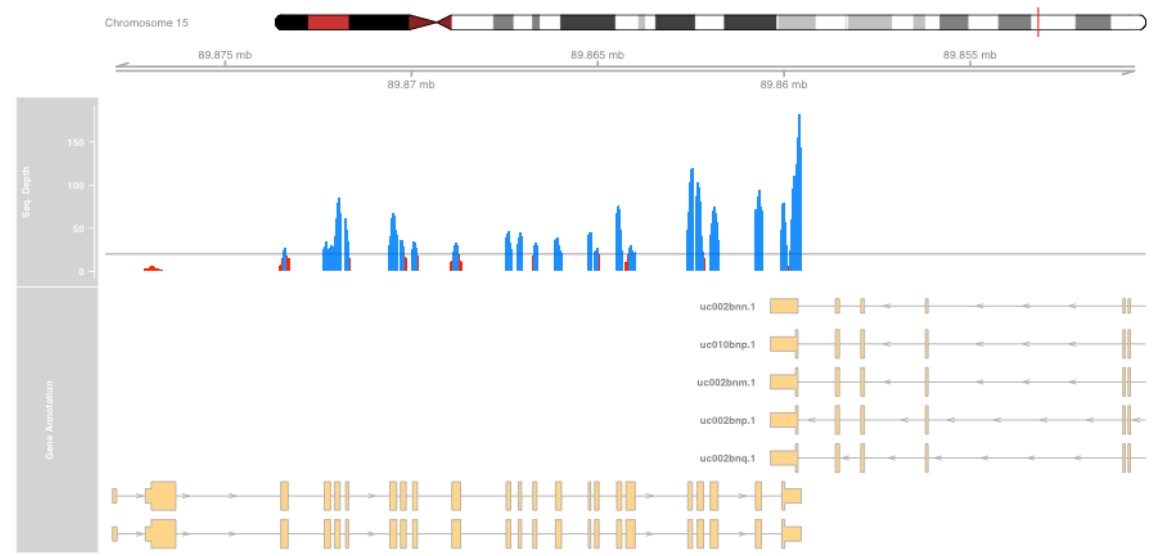
- information about POLG exons in `Exon genomic coordinate positions from UCSC` table

	number_of_transcript	type_of_transcript	chrom	start	end	length_of_exon	cds_id	exon_rank
1	1	uc002bnr.4	chr15	89876327	89876985	659	165612	2
2	1	uc002bnr.4	chr15	89873312	89873507	196	165611	3
3	1	uc002bnr.4	chr15	89872174	89872341	168	165610	4
4	1	uc002bnr.4	chr15	89871916	89872062	147	165609	5
5	1	uc002bnr.4	chr15	89871687	89871766	80	165608	6
6	1	uc002bnr.4	chr15	89870398	89870580	183	165607	7
7	1	uc002bnr.4	chr15	89870143	89870294	152	165606	8
8	1	uc002bnr.4	chr15	89869843	89869969	127	165605	9
9	1	uc002bnr.4	chr15	89868681	89868917	237	165604	10
10	1	uc002bnr.4	chr15	89867338	89867458	121	165603	11

Showing 1 to 10 of 44 entries Previous 1 2 3 4 5 Next

Figure 5: Screenshot of output of UCSC exons table.

- sequence gene coverage plot in `Gene coverage`. The plot displays the chromosome ideogram, the genomic location and gene annotations from **Ensembl** and the transcript(s) annotation from **UCSC**. Processing time is few minutes. A related table shows the number of uncovered positions in each exon given a user-defined transcript number (here transcript number is 1), and the user-defined threshold coverage (here the coverage threshold is 20x). Table and plot both show the many genomic positions that display low-DoC profile in POLG.



Show 10 entries Search:

	exon	ucovered positions
1	2	285
2	3	147
3	4	28
4	6	10
5	8	68

Figure 6: Gene coverage output.

- plot of a specific exon, choose exon 10 in sidebar `Exon number`, push `Make exon` and view the plot in `Exon coverage`. Processing time is few minutes. A related table shows the number of low-DoC positions in **ClinVar** which have a high impact annotation. For this output to be generated, **sorted.bed.gz** and **sorted.bed.gz.tbi** are required to be in the unCOVERApp directory. Table and plot both show that 21 low-DoC genomic positions have ClinVar annotation, suggesting several clinically relevant positions that are not adequately represented in this experiment.

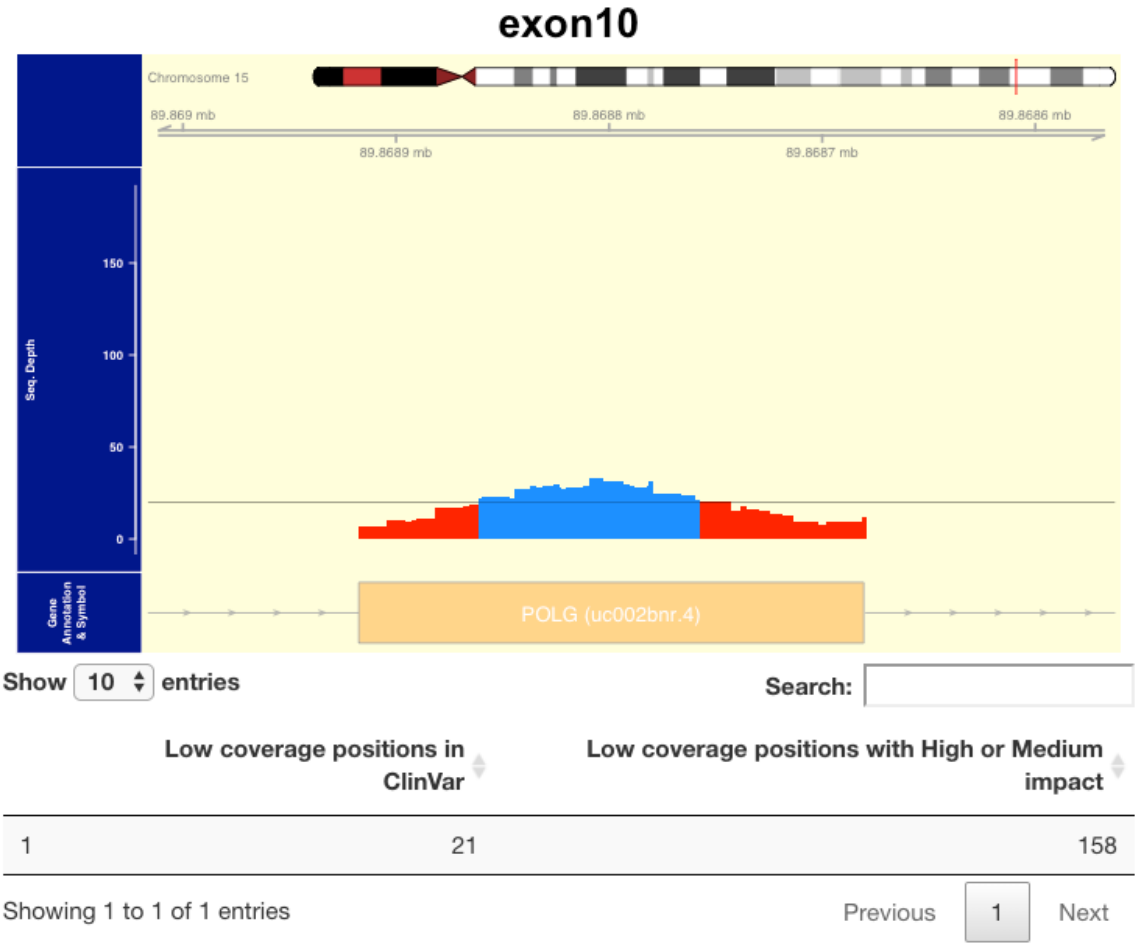


Figure 7: Exon coverage output.

- annotations obtained from dbNSFP for low-DoC positions are collected in `Annotation on low-DoC positions`. Functional and clinical annotations of all potential non-synonymous single-nucleotide variants across the examined low-DoC sites are made available. Potential changes that have a clinical annotation, a high impact or deleterious prediction are highlighted in yellow. In the example, a low-DoC site (chr15:89868687) is predicted as pathogenic and could be potentially linked to disease.

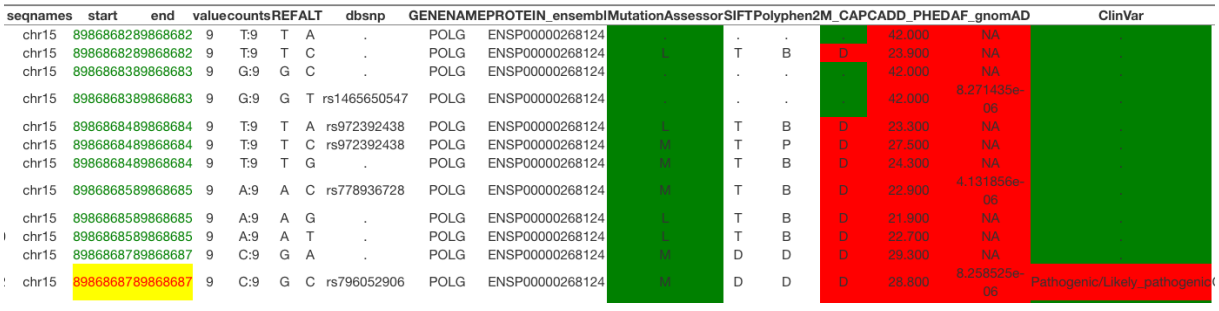


Figure 8: Example of uncovered positions annotate with dbNSFP.

By clicking on the "download" button, users can save the table as spreadsheet format with certain cells colored according to pre-specified thresholds for AF, CADD, MAP-CAP, SIFT, Polyphen2, ClinVar, OMIM ID, HGVSp and HGVSc, ...).

In **Calculate maximum credible allele frequency** page, users can set allele frequency cut-offs based on specific assumptions about the genetic architecture of the disease. If not specified, variants with allele frequency > 5 % will be instead filtered out. More details are available here . Moreover, users may click on the "download" button and save the resulting table as spreadsheet format.

The **Binomial distribution** page returns the 95 % binomial probability distribution of the variant-supporting reads on the input genomic position (`START genomic position` and `END genomic position`). Users should dfine the expected `allele fraction` (the expected fraction of variant reads, probability of success) and `Variant reads` (the minimum number of variant reads required by the user to support variant calling, number of successes).