HOWARD Help Parameters Databases

Contents

1 Introduction

HOWARD Parameters JSON file defines parameters for databases' downloads.

Examples:

Example of simple databases parameters JSON file for downloads

```
{
   "databases": {
         "genomes": {
             "download_genomes": "~/howard/databases/genomes/current",
             "download_genomes_contig_regex": "chr[0-9XYM]+$"
         },
         "snpeff": {
             "download_snpeff": "~/howard/databases/snpeff/current"
         },
         "annovar": {
             "download_annovar": "~/howard/databases/annovar/current",
             "download_annovar_files": "refGene,cosmic70,nci60"
         },
         "refseq": {
             "download_refseq": "~/howard/databases/refseq/current"
         },
         "dbnsfp": {
             "download_dbnsfp": "~/howard/databases/dbnsfp/current",
             "download_dbnsfp_release": "4.4a"
         },
         "alphamissense": {
             "download_alphamissense": "~/howard/databases/alphamissense/current"
         },
         "exomiser": {
             "download_exomiser": "~/howard/databases/exomiser/current"
         },
         "dbsnp": {
             "download_dbsnp": "~/howard/databases/dbsnp/current",
             "download_dbsnp_releases": "b156",
             "download_dbsnp_vcf": true,
             "download_dbsnp_parquet": true
         },
         "assemblies": [
             "hg19"
         ]
   }
}
```

Example of a full Databases parameters JSON file for downloads

```
{
   "databases": {
         "genomes_folder": "~/howard/databases/genomes/current",
         "genome": "~/howard/databases/genomes/current/hg19/hg19.fa",
         "genomes": {
             "download_genomes": "~/howard/databases/genomes/current",
             "download_genomes_provider": "UCSC",
             "download_genomes_contig_regex": "chr[0-9XYM]+$"
         },
         "snpeff": {
             "download_snpeff": "~/howard/databases/snpeff/current"
```

```
},
"annovar": {
    "download_annovar": "~/howard/databases/annovar/current",
    "download_annovar_files": "refGene,cosmic70,nci60",
    "download_annovar_url": "http://www.openbioinformatics.org/annovar/download"
},
"refseq": {
    "download_refseq": "~/howard/databases/refseq/current",
    "download_refseq_url": "http://hgdownload.soe.ucsc.edu/goldenPath",
    "download_refseq_prefix": "ncbiRefSeq",
    "download_refseq_files": "ncbiRefSeq.txt,ncbiRefSeqLink.txt",
    "download_refseq_format_file": "ncbiRefSeq.txt",
    "download_refseq_include_utr5": false,
    "download_refseq_include_utr3": false,
    "download_refseq_include_chrM": false,
    "download_refseq_include_non_canonical_chr": false,
    "download_refseq_include_non_coding_transcripts": false,
    "download_refseq_include_transcript_version": false
},
"dbnsfp": {
    "download_dbnsfp": "~/howard/databases/dbnsfp/current",
    "download_dbnsfp_url": "https://dbnsfp.s3.amazonaws.com",
    "download_dbnsfp_release": "4.4a",
    "download_dbnsfp_parquet_size": 100,
    "download_dbnsfp_subdatabases": true,
    "download_dbnsfp_parquet": false,
    "download_dbnsfp_vcf": false,
    "download_dbnsfp_no_files_all": false,
    "download_dbnsfp_add_info": false,
    "download_dbnsfp_uniquify": false,
    "download_dbnsfp_row_group_size": 100000
},
"alphamissense": {
    "download_alphamissense": "~/howard/databases/alphamissense/current",
    "download_alphamissense_url": "https://storage.googleapis.com/dm_alphamissense"
},
"exomiser": {
    "download_exomiser": "~/howard/databases/exomiser/current",
    "download_exomiser_url": "http://data.monarchinitiative.org/exomiser",
    "download_exomiser_remm_url": "https://kircherlab.bihealth.org/download/ReMM",
    "download_exomiser_cadd_url": "https://kircherlab.bihealth.org/download/CADD",
    "download_exomiser_cadd_url_snv_file": "whole_genome_SNVs.tsv.gz",
    "download_exomiser_cadd_url_indel_file": "InDels.tsv.gz"
},
"dbsnp": {
    "download_dbsnp": "~/howard/databases/dbsnp/current",
    "download_dbsnp_releases": "b156",
    "download_dbsnp_url": "https://ftp.ncbi.nih.gov/snp/archive",
    "download_dbsnp_url_files_prefix": "GCF_000001405",
    "download_dbsnp_assemblies_map": {
        "hg19": "25",
        "hg38": "40"
    },
    "download_dbsnp_vcf": true,
    "download_dbsnp_parquet": false
},
```

```
            "parameters": {
                "generate_param_releases": "current",
                "generate_param_formats": "parquet",
                "generate_param_bcftools": false
            },
            "assemblies": [
                "hg19"
            ]
        }
    }
```

2 assembly

Genome Assembly (e.g. 'hg19', 'hg38').

Type: `str`

Default: `hg19`

Examples:

Default assembly for all analysis tools

```
{
   "assembly": "hg19"
}
```

List of assemblies for databases download tool

```
{
   "assembly": "hg19,hg38"
}
```

3 genomes_folder

Folder containing genomes. (e.g. '~/howard/databases/genomes/current'

Type: `Path`

Default: `~/howard/databases/genomes/current`

4 genome

Genome file in fasta format (e.g. 'hg19.fa', 'hg38.fa').

Type: `Path`

Default: `~/howard/databases/genomes/current/hg19/hg19.fa`

5 genomes

Genomes download.

5.1 download_genomes

Path to genomes folder with Fasta files, indexes, and all files generated by pygenome module. (e.g. '~/howard/databases/genomes/current').

Type: `Path`

Default: `None`

5.2 download_genomes_provider

Download Genome from an external provider. Available: GENCODE, Ensembl, UCSC, NCBI.

Type: `str`

Choices: `['GENCODE', 'Ensembl', 'UCSC', 'NCBI']`

Default: `UCSC`

5.3 download_genomes_contig_regex

Regular expression to select specific chromosome (e.g 'chr[0-9XYM]+$').

Type: `str`

Default: `None`

6 snpeff

snpEff download.

6.1 download_snpeff

Download snpEff databases within snpEff folder

Type: `Path`

Default: `None`

7 annovar

Annovar download.

7.1 download_annovar

Path to Annovar databases (e.g. '~/howard/databases/annovar/current').

Type: `Path`

Default: `None`

7.2 download_annovar_files

Download Annovar databases for a list of Annovar file code (see Annovar Doc). Use None to donwload all available files, or Annovar keyword (e.g. 'refGene', 'cosmic70', 'clinvar_202*'). Note that refGene will at least be downloaded, and only files that not already exist or changed will be downloaded.

Type: `str`

Default: `None`

7.3 download_annovar_url

Annovar databases URL (see Annovar Doc).

Type: `str`

Default: `http://www.openbioinformatics.org/annovar/download`

8 refseq

refSeq download.

8.1 download_refseq

Path to refSeq databases (e.g. '~/howard/databases/refseq/current').

Type: `Path`

Default: `None`

8.2 download_refseq_url

refSeq databases URL (see refSeq WebSite) (e.g. 'http://hgdownload.soe.ucsc.edu/goldenPath') • /n

Type: `str`

Default: `http://hgdownload.soe.ucsc.edu/goldenPath`

8.3 download_refseq_prefix

Check existing refSeq files in refSeq folder.

Type: `str`

Default: `ncbiRefSeq`

8.4 download_refseq_files

List of refSeq files to download.

Type: `str`

Default: `ncbiRefSeq.txt,ncbiRefSeqLink.txt`

8.5 download_refseq_format_file

Name of refSeq file to convert in BED format (e.g. 'ncbiRefSeq.txt'). Process only if not None.

Type: `str`

Default: `None`

8.6 download_refseq_include_utr5

Formating BED refSeq file including 5'UTR.

Default: `False`

8.7 download_refseq_include_utr3

Formating BED refSeq file including 3'UTR.

Default: `False`

8.8 download_refseq_include_chrM

Formating BED refSeq file including Mitochondiral chromosome 'chrM' or 'chrMT'.

Default: `False`

8.9 download_refseq_include_non_canonical_chr

Formating BED refSeq file including non canonical chromosomes.

Default: `False`

8.10 download_refseq_include_non_coding_transcripts

Formating BED refSeq file including non coding transcripts.

Default: `False`

8.11 download_refseq_include_transcript_version

Formating BED refSeq file including transcript version.

Default: `False`

9 dbnsfp

dbNSFP download.

9.1 download_dbnsfp

Download dbNSFP databases within dbNSFP folder(e.g. '~/howard/databases').

Type: `Path`

Default: `None`

9.2 download_dbnsfp_url

Download dbNSFP databases URL (see dbNSFP website) (e.g. https://dbnsfp.s3.amazonaws.com').

Type: `str`

Default: `https://dbnsfp.s3.amazonaws.com`

9.3 download_dbnsfp_release

Release of dbNSFP to download (see dbNSFP website) (e.g. '4.4a').

Default: `4.4a`

9.4 download_dbnsfp_parquet_size

Maximum size (Mb) of data files in Parquet folder. Parquet folder are partitioned (hive) by chromosome (sub-folder), which contain N data files.

Type: `int`

Default: `100`

9.5 download_dbnsfp_subdatabases

Generate dbNSFP sub-databases. dbNSFP provides multiple databases which are split onto multiple columns. This option create a Parquet folder for each sub-database (based on columns names).

Default: `False`

9.6 download_dbnsfp_parquet

Generate a Parquet file for each Parquet folder.

Default: `False`

9.7 download_dbnsfp_vcf

Generate a VCF file for each Parquet folder. Need genome FASTA file (see --download-genome).

Default: `False`

9.8 download_dbnsfp_no_files_all

Not generate database Parquet/VCF file for the entire database ('ALL'). Only sub-databases files will be generated. (see '--download-dbnsfp-subdatabases').

Default: `False`

9.9 download_dbnsfp_add_info

Add INFO column (VCF format) in Parquet folder and file. Useful for speed up full annotation (all available columns). Increase memory and space during generation of files.

Default: `False`

9.10 download_dbnsfp_only_info

Add only INFO column (VCF format) in Parquet folder and file. Useful for speed up full annotation (all available columns). Decrease memory and space during generation of files. Increase time for partial annotation (some available columns).

Default: `False`

9.11 download_dbnsfp_uniquify

Uniquify values within column (e.g. "D,D" to "D", "D,.,T" to "D,T"). Remove transcripts information details. Usefull to reduce size of the database. Increase memory and space during generation of files.

Default: `False`

9.12 download_dbnsfp_row_group_size

Minimum number of rows in a parquet row group (see duckDB doc). Lower can reduce memory usage and slightly increase space during generation, speed up highly selective queries, slow down whole file queries (e.g. aggregations).

Type: `int`

Default: `100000`

10 alphamissense

AlphaMissense download.

10.1 download_alphamissense

Path to AlphaMissense databases

Type: `Path`

Default: `None`

10.2 download_alphamissense_url

Download AlphaMissense databases URL (see AlphaMissense website) (e.g. 'https://storage.googleapis.com/dm_alphamissense').

Type: `str`

Default: `https://storage.googleapis.com/dm_alphamissense`

11 exomiser

Exomiser download.

11.1 download_exomiser

Path to Exomiser databases (e.g. ~/howard/databases/exomiser/current).

Type: `Path`

Default: `None`

11.2 download_exomiser_application_properties

Exomiser Application Properties configuration file (see Exomiser website). This file contains configuration settings for the Exomiser tool. If this parameter is not provided, the function will attempt to locate the application properties file automatically based on the Exomiser. Configuration information will be used to download expected releases (if no other parameters). CADD and REMM will be downloaded only if 'path' are provided.

Type: `Path`

Default: `None`

11.3 download_exomiser_url

URL where Exomiser database files can be downloaded from (e.g. 'http://data.monarchinitiative.org/exomiser').

Type: `str`

Default: `http://data.monarchinitiative.org/exomiser`

11.4 download_exomiser_release

Release of Exomiser data to download. If "default", "auto", or "config", retrieve from Application Properties file. If not provided (None), from Application Properties file (Exomiser data-version) or default '2109'.

Type: `str`

Default: `None`

11.5 download_exomiser_phenotype_release

Release of Exomiser phenotype to download. If not provided (None), from Application Properties file (Exomiser Phenotype data-version) or Exomiser release.

Type: `str`

Default: `None`

11.6 download_exomiser_remm_release

Release of ReMM (Regulatory Mendelian Mutation) database to download. If "default", "auto", or "config", retrieve from Application Properties file.

Type: `str`

Default: `None`

11.7 download_exomiser_remm_url

URL where ReMM (Regulatory Mendelian Mutation) database files can be downloaded from (e.g. 'https://kircherlab.bihealth.org/d

Type: `str`

Default: `https://kircherlab.bihealth.org/download/ReMM`

11.8 download_exomiser_cadd_release

Release of CADD (Combined Annotation Dependent Depletion) database to download. If "default", "auto", or "config", retrieve from Application Properties file.

Type: `str`

Default: `None`

11.9 download_exomiser_cadd_url

URL where CADD (Combined Annotation Dependent Depletion) database files can be downloaded from (e.g. 'https://kircherlab.bihealth.org/download/CADD').

Type: `str`

Default: `https://kircherlab.bihealth.org/download/CADD`

11.10 download_exomiser_cadd_url_snv_file

Name of the file containing the SNV (Single Nucleotide Variant) data for the CADD (Combined Annotation Dependent Depletion) database.

Type: `str`

Default: `whole_genome_SNVs.tsv.gz`

11.11 download_exomiser_cadd_url_indel_file

Name of the file containing the INDEL (Insertion-Deletion) data for the CADD (Combined Annotation Dependent Depletion) database.

Type: `str`

Default: `InDels.tsv.gz`

12 dbsnp

dbSNP download.

12.1 download_dbsnp

Path to dbSNP databases (e.g. '~/howard/databases/exomiser/dbsnp').

Type: `Path`

Default: `None`

12.2 download_dbsnp_releases

Release of dbSNP to download (e.g. 'b152', 'b152,b156').

Type: `str`

Default: `b156`

12.3 download_dbsnp_release_default

Default Release of dbSNP ('default' symlink) (e.g. 'b156'). If None, first release to download will be assigned as default only if it does not exists.

Type: `str`

Default: `None`

12.4 download_dbsnp_url

URL where dbSNP database files can be downloaded from. (e.g. 'https://ftp.ncbi.nih.gov/snp/archive').

Type: `str`

Default: `https://ftp.ncbi.nih.gov/snp/archive`

12.5 download_dbsnp_url_files

Dictionary that maps assembly names to specific dbSNP URL files. It allows you to provide custom dbSNP URL files for specific assemblies instead of using the default file naming convention.

Type: `str`

Default: `None`

12.6 download_dbsnp_url_files_prefix

String that represents the prefix of the dbSNP file name for a specific assembly. It is used to construct the full URL of the dbSNP file to be downloaded.

Type: `str`

Default: `GCF_000001405`

12.7 download_dbsnp_assemblies_map

dictionary that maps assembly names to their corresponding dbSNP versions. It is used to construct the dbSNP file name based on the assembly name.

Type: `str`

Default: `{'hg19': '25', 'hg38': '40'}`

12.8 download_dbsnp_vcf

Generate well-formatted VCF from downloaded file:

- Add and filter contigs associated to assembly
- Normalize by splitting multiallelics
- Need genome (see --download-genome)

Default: `False`

12.9 download_dbsnp_parquet

Generate Parquet file from VCF.

Default: `False`

13 hgmd

HGMD convert.

13.1 convert_hgmd

Convert HGMD databases. Folder where the HGMD databases will be stored. Fields in VCF, Parquet and TSV will be generated. If the folder does not exist, it will be created.

Type: `Path`

Default: `None`

13.2 convert_hgmd_file

File from HGMD. Name format 'HGMD_Pro___.vcf.gz'.

Type: `Path`

Default: `None`

13.3 convert_hgmd_basename

File output basename. Generated files will be prefixed by basename (e.g. 'HGMD_Pro_MY_RELEASE') By default (None), input file name without '.vcf.gz'.

Type: `str`

Default: `None`

14 from_Annovar

Annovar convert.

14.1 input_annovar

Input Annovar file path. Format file must be a Annovar TXT file, associated with '.idx'.

Type: `Path`

Default: `None`

14.2 output_annovar

Output Annovar file path. Format file must be either VCF compressesd file '.vcf.gz'.

Type: `Path`

Default: `None`

14.3 annovar_code

Annovar code, or database name. Usefull to name databases columns.

Type: `str`

Default: `None`

14.4 annovar_to_parquet

Parquet file conversion.

Type: `Path`

Default: `None`

14.5 annovar_reduce_memory

Reduce memory option for Annovar convert, either 'auto' (auto-detection), 'enable' or 'disable'.

Type: `str`

Choices: `['auto', 'enable', 'disable']`

Default: `auto`

14.6 annovar_multi_variant

Variant with multiple annotation lines on Annovar file. Either 'auto' (auto-detection), 'enable' or 'disable'.

Type: `str`

Choices: `['auto', 'enable', 'disable']`

Default: `auto`

15 from_extann

Extann convert (gene annotation).

15.1 input_extann

Input Extann file path. Format file must be a Extann TXT file or TSV file. File need to have at least the genes column.

Type: `Path`

Default: `None`

15.2 output_extann

Output Extann file path. Output extann file, should be BED or BED.gz.

Type: `Path`

Default: `None`

15.3 refgene

Path to refGene annotation file.

Type: `Path`

Default: `None`

15.4 transcripts

Transcripts TSV file, with Transcript in first column, optional Gene in second column.

Type: `Path`

Default: `None`

15.5 param_extann

Param extann file path. Param containing configuration, options to replace chars and bedlike header description, conf vcf specs. (e.g. '~/howard/config/param.extann.json')

Type: `Path`

Default: `None`

15.6 mode_extann

Mode extann selection. How to pick transcript from ncbi, keep all, keep the longest, or keep the chosen one (transcript_extann).

Type: `str`

Choices: `['all', 'longest', 'chosen']`

Default: `longest`

16 Parameters

Parameters generation.

16.1 generate_param

Parameter file (JSON) with all databases found. Databases folders scanned are defined in config file. Structure of databases follow this structure (see doc): .../////*.[parquet|vcf.gz|...]

Type: `Path`

Default: `None`

16.2 generate_param_description

Description file (JSON) with all databases found. Contains all databases with description of format, assembly, fields...

Type: `Path`

Default: `None`

16.3 generate_param_releases

List of database folder releases to check (e.g. 'current', 'latest').

Type: `str`

Default: `current`

16.4 generate_param_formats

List of database formats to check (e.g. 'parquet', 'parquet,vcf,bed,tsv').

Type: `str`

Default: `parquet`

16.5 generate_param_bcftools

Generate parameter JSON file with BCFTools annotation for allowed formats (i.e. 'vcf', 'bed').

Default: `False`