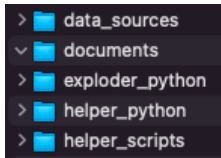


# Data management and processing instructions

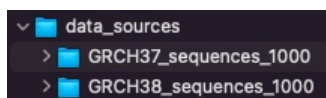
## Adding a new locus to the data management hierarchy

a) Locate the `$root` directory.

The Github repository [https://github.com/snowlizardz/rg\\_exploder\\_shared](https://github.com/snowlizardz/rg_exploder_shared), has four directories at `$root` level



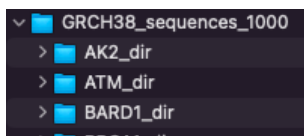
b) In `$root/data_sources` these are the first-level data-source directories



GRCH37\_sequences\_1000: holds data files for build GRCH37

GRCH38\_sequences\_1000: holds data files for build GRCH38

c) Within each of these, there is one data directory for each locus:



Eg: **GRCH38\_sequences\_1000/AK2\_dir** holds downloaded AK2 data from Ensembl; initially as a file called **ensembl.txt.gz**, and later processed versions of these files (see ‘sequence file processing’ below)

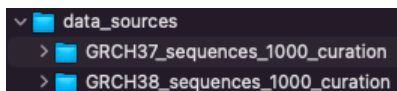
c) To add a new `$locus`, create a new folder in the appropriate directory

`GRCH38_sequences_1000/$locus_dir` **or** `GRCH37_sequences_1000/$locus_dir`

d) Follow “Downloading a sequence file ...” instructions below for the new `$locus`

e) Follow “Automated data processing...” instructions below for the new `$locus`

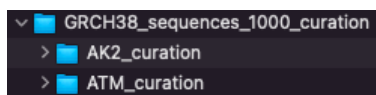
f) Also in `$root/data_sources` are the second two main data-curation directories



GRCH37\_sequences\_1000\_curation: these hold curated files for build GRCH37

GRCH38\_sequences\_1000\_curation: these hold curated files for build GRCH38

g) Within each of these, there is one data directory for each locus, eg:



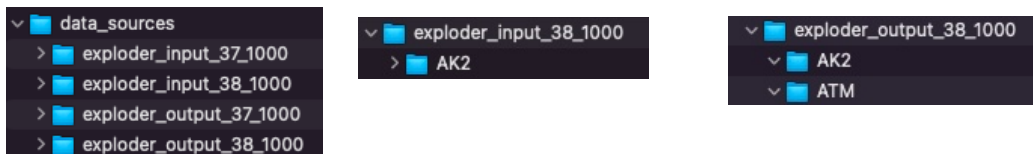
h) As with step c, add a new `$locus` folder in the appropriate curation directory eg:

`GRCH38_sequences_1000_curation/$locus_curation`

i) Follow “Maintaining the curation data” instructions below for the new `$locus`

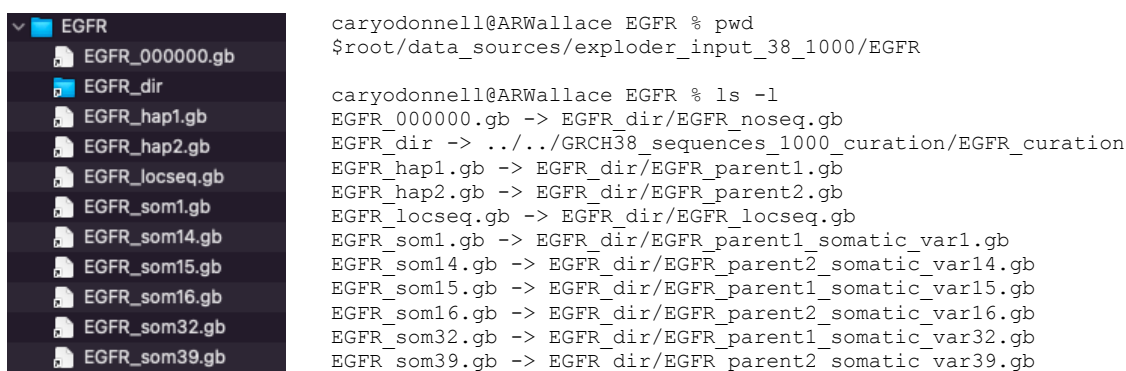
j) Follow “Mashup time” instructions below to include the new `$locus` in an updated version of the lookup file `$root/data_sources/GRCH38_sequences_1000_curation/loci.json`

k) In `$root/data_sources` are further directories to hold the input & output data when running the application. Create a directory for each locus in both the input and output directories:



l) In `$root/data_sources/exploder_input_38_1000/$locus` create soft links to files in the `GRCH38_sequences_1000_curation/$locus_curation` directory.

If this seems like overkill, it allows the culling or renaming of individual curated haplotype definitions, so separating the curation area from the input area. This is best illustrated by the EGFR set:



To set up new folders and links automatically, see “Adding multiple new input and output folders”

m) In the folder `$root/data_sources/exploder_input_38_1000/`

Check that a soft link exists here to the file created previously

`$root/data_sources/GRCH38_sequences_1000_curation/loci.json`

n) Locate `$root/data_sources/exploder_python`

Create soft links to the desired input and output directories

```
caryodonnell@ARWallace exploder_python % pwd
$root/exploder_python

input -> $root/data_sources/exploder_input_38_1000/
output -> $root/data_sources/exploder_output_38_1000/
```

Use `$root/data_sources/helper_scripts/switch_links.sh` to flip quickly between different sets; 37 and 38, for example

o) To create the lookup file `$root/data_sources/exploder_python/input/config.json`

Run the python script (check values in `set_config_consts`)

`$root/data_sources/exploder_python/RG_exploder_globals_make.py`

p) Finally, run the application `$root/data_sources/exploder_python/RG_exploder_gui.py`

# Downloading a sequence file for a locus from Ensembl

These instructions are suitable for downloading a new sequence, or when updating an existing one

Starting at [https://www.ensembl.org/Homo\\_sapiens/Info/Index](https://www.ensembl.org/Homo_sapiens/Info/Index):

- Find the gene of interest using Search & go to the Summary eg: [ATM](#)
  - Use the chosen gene name as *locus* below.
- Click on “export data” (LH menu)
- Select output: Flatfile/Genbank
- Select Forward Strand (preferred)
  - Alternatives are:
    - Feature Strand (This will be Forward or Reverse depending on the transcript)
    - Reverse Strand
- In “5' Flanking sequence (upstream)” and “3' Flanking sequence (downstream)”: enter **1000**
  - A minimum value of 1000 is essential for supporting ‘paired end reads’
- In “Options for Genbank”:
  - Deselect all
  - Reselect: “variation features” and “gene information” (exon, mRNA & CDS definitions)
- Press “Next”

The screenshot shows the 'Export data' window in Ensembl. The 'Export Configuration - Feature List' section on the left includes a tip, a 'Gene to export' field with 'ENSG00000149311.20 (ATM)', an 'Output' dropdown set to 'GenBank', a 'Strand' dropdown set to 'Forward strand', and two input fields for '5' Flanking sequence (upstream)' and '3' Flanking sequence (downstream)', both set to '1000'. A 'Next >' button is at the bottom. The 'Options for GenBank' section on the right has checkboxes for 'Select/deselect all:', 'Similarity features:', 'Repeat features:', 'Prediction features (genscan):', 'Contig Information:', 'Variation features:' (checked), 'Marker features:', 'Gene Information:' (checked), 'Vega Gene Information:', and 'EST Gene Information:'.

- In the new “Export data” window, click the “compressed text (gz)” link
- The downloaded file is named **ensembl.txt.gz**
  - Move this file, into a data directory called `$root/data_sources/$locus_dir` eg: the `GRCH38_sequences_1000/AK2_dir` example above
  - Rename it to `$locus_ensembl` eg: `AK2_ensembl`

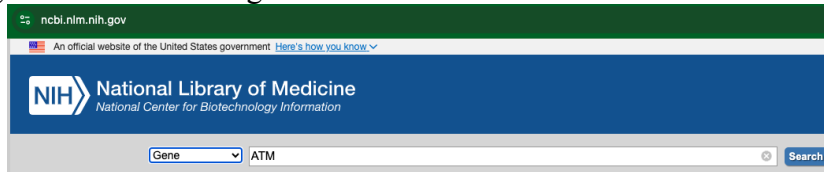
The screenshot shows the 'Export data' window in Ensembl, specifically the 'Export Configuration - Output Format' section. It asks the user to 'Please choose the output format for your export' and provides three radio button options: 'HTML', 'Text', and 'Compressed text (.gz)'. The 'Compressed text (.gz)' option is selected.

# Downloading a sequence file for a locus from NCBI

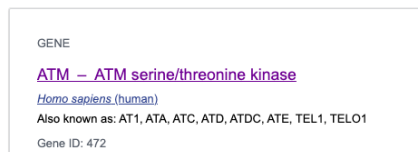
These instructions are suitable for downloading a new sequence, or when updating an existing one

Starting at <https://www.ncbi.nlm.nih.gov/>

- Find the gene of interest using Search



- Click on the link in the gene card



- Which shows the gene summary

ATM ATM serine/threonine kinase [ *Homo sapiens* (human) ]

Gene ID: 472, updated on 5-Jan-2025

Download Datasets

Table of contents

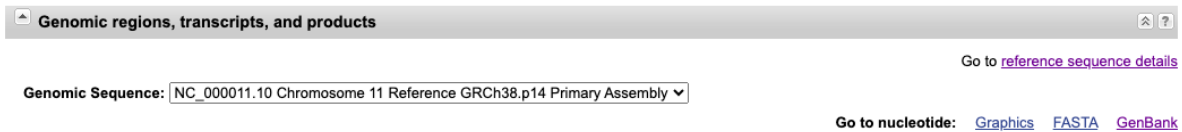
Summary

Genomic context

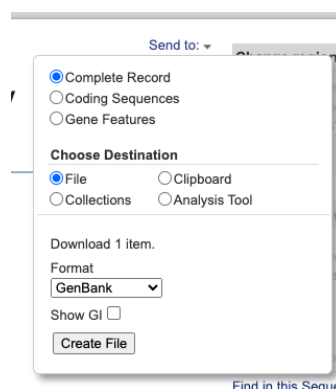
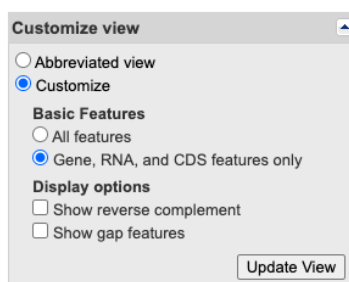
Genomic regions, transcripts, and products

Summary

- Click on "Genomic regions, transcripts, and products", then scroll down



- Click on "Go to nucleotide ... Genbank" eg:
  - [https://www.ncbi.nlm.nih.gov/nuccore/NC\\_000011.10?report=genbank&from=108223067&to=108369102](https://www.ncbi.nlm.nih.gov/nuccore/NC_000011.10?report=genbank&from=108223067&to=108369102)
- Subtract 1000 from the start and add 1000 to the end:
  - [https://www.ncbi.nlm.nih.gov/nuccore/NC\\_000011.10?report=genbank&from=108222067&to=108370102](https://www.ncbi.nlm.nih.gov/nuccore/NC_000011.10?report=genbank&from=108222067&to=108370102)
- Top of page; modify "Customize view"; pulldown "Send to"; click "File" & "Create File"



- The output is a file called sequence.db

# Automated data processing for downloaded Ensembl data

## Introduction to data processing

There are two main objectives:

- a) Create a `config.json` file as a lookup list for the application.
- b) Remove, from `ensembl.txt.gz`, all the data unnecessary for this application.
- c) Optionally gzip the original: `gzip ATM_Ensembl`

Adding new data into the `config.json` file could be done manually by looking at the existing examples. The helper scripts automate the filtering, `config.json` generation, and other fiddly bits.

## Using a feature-filter script

A Python script, `$root/helper_python/embl_feature_filter_revise.py` can be used to process the `$locus_ensembl` file in either gz or uncompressed format eg:

```
> cd $root/data_sources/GRCH38_sequences_1000/$locus_dir
> python3 $root/helper_python/embl_feature_filter_revise.py -i $locus_ensembl -a
```

The output files, which are used by the application, are:

`$locus_locseq.gb`: Contains a cleaned-up feature table: retaining only minimal `db_xref` identifiers; mRNA and CDS join data. Also holds the DNA sequence.

`$locus_noseq.gb`: Contains **no** DNA sequence and the bare minimum definition data. T In the application it is used to define the `$locus_00000` haplotype. It is also used as a template for defining the variants in haplotypes, in the curation directory.

`$locus_transcripts.json`: Holds lookup data for the GUI application.

Other output files, useful for curation and checking:

`$locus_filtered.gb`: The ‘original file’ with all the unwanted data taken out. Same content as `$locus_locseq.gb`, with the inclusion of all the variation features from the original source.

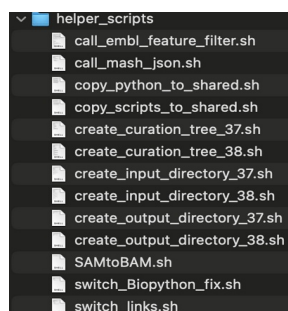
`$locus_filtvar.gb`: As for `$locus_noseq.gb`, but including the the variation features. These can be useful for extracting a subset to make haplotype definition files.

Parameters for `embl_feature_filter_revise.py`:

- `-i` is the downloaded-from-Ensembl input file eg: `ensembl.txt.gz`,
- `-a` is necessary to produce “all” output files (described below)
- `-j` removes mRNA and CDS *join* data to `$locus_transcripts.json`
  - `-j` may be used **only** for the Python GUI; the browser version **requires** the join data

## Filtering multiple data sets at once

To automate this one step further, by (re-)processing multiple locus directories at once, using scripts in `$root/helper_scripts/`



```
> cd $root/data_sources/GRCH38_sequences_1000/
> sh $root/helper_scripts/call_embl_feature_filter.sh
```

Or

```
> cd $root/data_sources/GRCH37_sequences_1000/
> sh $root/helper_scripts/call_embl_feature_filter.sh
```

If this is just data-update, go to the section below “Mashup time”

## Alternative feature-filter scripts

A previously-used script that does the same task as `embl_feature_filter_revise.py`, but does not depend on Biopython or modules used in the main application, is `embl_feature_filter8.py`

This script relies on the elimination of numerous 'dbxref= database' lines:

```
unwanted=['xref="RefSeq_mRNA_predicted:', ... 'db_xref="RefSeq_ncRNA_predicted:']
```

Note that new, unnecessary data appears over time, so this list needs to be maintained

## Amendments for attention in these scripts prior to execution

When adding a new locus, corresponding identifiers need to be added to two dictionary lists within `embl_feature_filter_revise.py` or `embl_feature_filter8.py`

- `MANE_Select_dict` and `LRG_id_dict`
  - Each item in `MANE_Select_dict` is the transcript identifier assigned MANE\_Select status for each locus.
  - Likewise items in `LRG_id_dict` are the LRG identifiers for each locus.
  - A link to [LRG](#) was an early user-requirement for a link from “Locus”, but LRG now appears to be being phased out. Links now go to the LRG page in Ensembl.
  - To include the correct date requires a code change in `make_transcriptconstants()`, eg: `Release="Ensembl Release 105 (Dec 2021)"`

Within `call_embl_feature_filter.sh` you may change which of the feature-filter scripts is used

```
#Choose one of two filter programs
python_filter1="/Users/caryodonnell/mytools/embl_feature_filter8.py"
# embl_feature_filter8.py is a difficult-to-follow line-by-line parsing of the input file
python_filter2="/Users/caryodonnell/mytools/embl_feature_filter_revise.py"
...

# Pick one!
# python_filter=$python_filter1
python_filter=$python_filter2
```

## Data processing for downloaded NCBI data

Automated processing is currently not available for data downloaded directly from NCBI

There are significant differences between the content of Ensembl & Genbank downloads. Listing these, incompletely, here:

- There is no link between Transcript ID and CDS in Genbank output., unlike Ensembl
- The CDS in Genbank are clustered together in Genbank, and not interlaced with mRNA, as in Ensembl
- There are far fewer tags that would need excluding in `noref.sh`
  - The major things to remove are:
    - Any information from genes not in the intended gene set (ie: overlaps or other-strand)
    - `/translation=`
- The terms for the transcript id differ between Ensembl & Genbank eg:

### Ensembl

```
gene          1001..147059
               /gene=ENSG00000149311.20
               /locus_tag="ATM"
               /note="ATM serine/threonine kinase [Source:HGNC
               Symbol;Acc:HGNC:795]"

mRNA ...      /gene="ENSG00000149311.20"
               /
               standard_name="ENST00000675843.1"
```

### Genbank

```
gene          1001..147036
               /gene="ATM"
               /db_xref="GeneID:472"
               /db_xref="HGNC:HGNC:795"

mRNA ...      /gene="ATM"
               /transcript_id="XM_011542840.4"
```

Advantage:

a) Far less removal of unwanted content is required

Disadvantages:

a) Transcript ID tags are different

b) Cannot link a CDS to a given transcript with the information in this file, it would have to be inferred.

## Maintaining the curation data

### Introduction to curation data

The purpose of the curation area is to maintain a working area where the variant haplotype data can be manipulated separately from the reference-data. The curation directory initially holds soft links to the processed files in the curation directories. Occasionally an edited copy, instead of soft-link is used instead (eg: AK2 for GRCH37).

### Adding a new curation folder

```
> cd $root/data_sources/GRCH38_sequences_1000_curation/$locus_curation
```

Simply soft link to the source data locus directory

```
> ln -s ../ ../GRCH38_sequences_1000/$locus_dir .
```

Then soft link the following 2 files:

```
> ln -s $locus_dir/$locus_locseq.gb
```

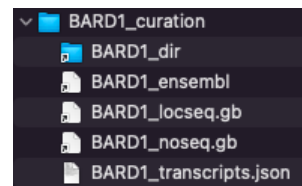
```
> ln -s $locus_dir/$locus_noseq.gb
```

Then copy this one:

```
> cp -p $locus_dir/$locus_transcripts.json .
```

For a simple setup, such as for BARD1, which has no haplotype definitions apart from the reference, nothing more needs to be done.

```
BARD1_dir -> ../../GRCH38_sequences_1000/BARD1_dir/  
BARD1_ensembl -> BARD1_dir/BARD1_ensembl  
BARD1_locseq.gb -> BARD1_dir/BARD1_locseq.gb  
BARD1_noseq.gb -> BARD1_dir/BARD1_noseq.gb  
BARD1_transcripts.json
```



A link to BARD1\_ensembl is not required in this directory, but its presence can be useful.

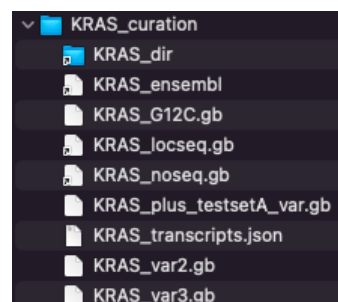
Other loci are described below; chosen to demonstrate both the standard and less-obvious maintenance options available.

### For AK2

```
> cd $root/data_sources/GRCH38_sequences_1000_curation/AK2_curation
```

```
> ls -l
```

```
AK2_EB0001.gb  
AK2_EB0002.gb  
AK2_EB0003.gb  
AK2_EB0004.gb  
AK2_EB0005.gb  
AK2_EB0006.gb  
AK2_dir -> ../../GRCH38_sequences_1000/AK2_dir/  
AK2_ensembl -> AK2_dir/AK2_ensembl  
AK2_locseq.gb -> AK2_dir/AK2_locseq.gb  
AK2_noseq.gb -> AK2_dir/AK2_noseq.gb  
AK2_transcripts.json
```



Each of the haplotype definition files, in **bold**, contain these essential components:



A) The Header section, note the **0 bp** definition, as there is no sequence in the file:

```
LOCUS      1                      0 bp      DNA      HTG 19-AUG-2022
DEFINITION Homo sapiens chromosome 1 GRCh38 partial sequence 33006986..33081996
            reannotated via Ensembl.

ACCESSION  chromosome:GRCh38:1:33006986:33081996:1
VERSION    chromosome:GRCh38:1:33006986:33081996:1
FEATURES   Location/Qualifiers
            source          1..75011
                        /organism="Homo sapiens"
                        /db_xref="taxon:9606"
            gene            complement(1001..74011)
                        /gene="ENSG00000004455.18"
                        /locus_tag="AK2"
                        /note="adenylate kinase 2 [Source:HGNC
                        Symbol;Acc:HGNC:362]"
```

This section should be the same as found in the file `AK2_noseq.gb` which can be used as a template for their creation.

B) A definition of the variation(s) in the Feature table, eg:

**AK2\_EB0001.gb** has just one:

```
variation      7582..7583
                /replace="GG/G"
```

**AK2\_EB0001.gb** also has one:

```
variation      14469..14471
                /replace="TCA/-"
```

C) To create a file defining a new variant, options include:

1) Use a SNP or other identifier and look in the file `AK2_ensembl` (the uncompressed, unfiltered source file) or in `AK2_dir/AK2_filtervar.gb`

eg: dbSNP:[rs1553151177](#) (the `AK2_EB0001` variant) can be found alongside other definitions

```
variation      7582..7582
                /replace="G/T"
                /db_xref="dbSNP:rs1241229733"
variation      7582..7582
                /replace="HGMD_MUTATION"
                /db_xref="HGMD-PUBLIC:CD090014"
variation      7582..7583
                /replace="GG/G"
                /db_xref="dbSNP:rs1553151177"
```

Edit the required lines into a copy of `AK2_noseq.gb`

Just be certain NOT to include overlapping definitions; the application does not recognise these overlaps and is likely to give incorrect output.

2) You may be able to edit a new definition from other markers you recognise, or using offsets. This manual method is very error-prone. The next option gives a way of creating a *validated* position and sequence-modification.

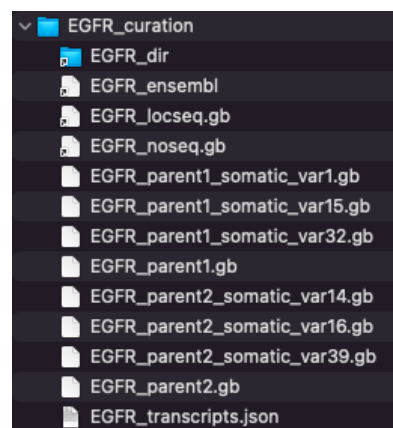
3) Leave it as a simple setup for now; complete the other “adding a new locus” steps; run the GUI and use the “Create a new Variations Source” feature. That will generate an output file with a name you supply that can be copied straight back into this directory. See “Additional variants” in EGFR.

## For EGFR

```
> cd $root/data_sources/GRCH38_sequences_1000_curation/EGFR_curation
```

```
> ls -l
```

```
EGFR_dir -> ../../GRCH38_sequences_1000/EGFR_dir/  
EGFR_ensembl -> EGFR_dir/EGFR_ensembl  
EGFR_locseq.gb -> EGFR_dir/EGFR_locseq.gb  
EGFR_noseq.gb -> EGFR_dir/EGFR_noseq.gb  
EGFR_parent1.gb  
EGFR_parent1_somatic_var1.gb  
EGFR_parent1_somatic_var15.gb  
EGFR_parent1_somatic_var32.gb  
EGFR_parent2.gb  
EGFR_parent2_somatic_var14.gb  
EGFR_parent2_somatic_var16.gb  
EGFR_parent2_somatic_var39.gb  
EGFR_transcripts.json
```



The variant definitions here are hierarchical, it uses two different haplotype definitions, each extracted originally from \*EGFR\_ensembl, with extra commentary added from other sources:

### EGFR\_parent1.gb

```
variation      155518..155518  
                /replace="G/A"  
                /db_xref="dbSNP:rs55959834"  
                /consequence="dbSNP:synonymous_variant,genic_downstream_transcript_variant"  
                /consequence="dbSNP:coding_sequence_variant"  
                /comment="ensembl:minor allele exon18, synonymous variant at v low freq 0.001"  
variation      160011..160011  
                /replace="G/A"  
                /db_xref="dbSNP:rs62457092"  
                /consequence="dbSNP:genic_downstream_transcript_variant,intron_variant"  
                /comment="ensembl:intron_variant 19_20 minor allele at 0.32"
```

### EGFR\_parent2.gb

```
variation      159298..159298  
                /replace="A/G"  
                /db_xref="dbSNP:rs845552"  
                /consequence="dbSNP:intron_variant,genic_downstream_transcript_variant"  
                /comment="ensembl:intron_variant 19_20 minor allele at 0.45"  
variation      162854..162854  
                /replace="G/A/"  
                /db_xref="dbSNP:rs1050171"  
                /consequence="dbSNP:genic_downstream_transcript_variant,synonymous_variant"  
                /consequence="dbSNP:missense_variant,non_coding_transcript_variant"  
                /consequence="dbSNP:coding_sequence_variant"  
                /comment="ensembl:minor allele exon 20, synonymous variant at 0.43"
```

The other files have further variants added onto these basic haplotypes

*\*an older version of*

## IMPORTANT:

The header sections of the 'parent1' files must agree with the source of the variant.

### EGFR\_parent1.gb header:

```
LOCUS      7 0 bp DNA HTG 21-APR-2021
ACCESSION  chromosome:GRCh38:7:55018517:55212128:1
VERSION    chromosome:GRCh38:7:55018517:55212128:1
COMMENT     /consequence and /comment annotation by Replicon Genetics from public domain sources
FEATURES    Location/Qualifiers
             source          1..193612
                               /organism="Homo sapiens"
                               /db_xref="taxon:9606"
             gene            501..193112
                               /gene=ENSG00000146648.19
                               /locus_tag="EGFR"
```

### EGFR\_parent1\_somatic\_var1.gb:

```
LOCUS      7 0 bp DNA HTG 21-APR-2021
ACCESSION  chromosome:GRCh38:7:55018517:55212128:1
VERSION    chromosome:GRCh38:7:55018517:55212128:1
COMMENT     /consequence and /comment annotation by Replicon Genetics from public domain sources
FEATURES    Location/Qualifiers
             source          1..193612
                               /organism="Homo sapiens"
                               /db_xref="taxon:9606"
             gene            501..193112
                               /gene=ENSG00000146648.19
                               /locus_tag="EGFR"
```

They DO NOT need to be *exactly the same* as the header in the reference-sequence file

### EGFR\_locseq.gb:

```
LOCUS      7 194612 bp DNA HTG 27-FEB-2022
DEFINITION  Homo sapiens chromosome 7 GRCh38 partial sequence 55018017..55212628 reannotated
            via Ensembl
ACCESSION  chromosome:GRCh38:7:55018017:55212628:1
VERSION    chromosome:GRCh38:7:55018017:55212628:1
FEATURES    Location/Qualifiers
             source          1..194612
                               /organism="Homo sapiens"
                               /db_xref="taxon:9606"
             gene            1001..193612
                               /gene=ENSG00000146648.20 ← The difference is gene-id version is acceptable
                               /locus_tag="EGFR"
```

The range for the reference definition is ~~GRCh38:7:55018017:55212628:1~~

The range for the variant definition is ~~GRCh38:7:55018517:55212128:1~~

- ~~Both the GRCh version and strand are the same: this is essential~~
- ~~The sequence range of the variant is wholly contained within the reference range~~
  - ~~If not, the application will generate a warning message and should ignore the haplotype completely.~~
  - ~~The application detects the disparity and recalculates the offset~~

In this way, variant definition files do **not** need to be regenerated each time the source files are updated, and where only a small redefinition of the gene range takes place.

The inspiration, and naming, for this set of EGFR variants comes from table 2 of [\*“Molecular characteristics and clinical outcomes of EGFR exon 19 indel subtypes to EGFR TKIs in NSCLC patients”\*](#) by Su et al *Oncotarget*.8(67); 2017 Dec 19

Subtypes are defined at CDS like this: Subtype 21 - c.2239\_2247del19

### Additional variants:

Here’s how to add two of the subtypes described using the application GUI rather than alternative, painstaking methods.

### Subtype 21

**Reference Source**

[Locus](#)

EGFR

[Template](#)

EGFR-275493(MANE\_Select)

CDS only ☒

**Create a new Variations Source**

Source	Local_pos	Extension	Global_pos
CDS_Begin	1	0	55019278
CDS_End	3633	0	55205617

**Create a new Variations Source**

Source	Local_pos	Extension	Global_pos
CDS_Begin	2239	0	55174776
CDS_End	2247	0	55174784

**Retrieve Reference Sequence**

**Reference Sequence**

9 bases: TTAAGAGAA

**Variant Sequence**

TTAAGAGAA

**Haplotype Definition Name**

hap01

**Variant Name**

c.2239\_2247

**Variant Sequence**

TTAAGAGAA

**Haplotype Definition Name**

som21

**Variant Name**

c.2239\_2247

**Save New Variations Source**

**Haplotype Definition**

Haplotype Definition	Source Ratio
EGFR_00000	0
EGFR_hap1	50
EGFR_hap2	50
EGFR_som1	30
EGFR_som14	30
EGFR_som15	30
EGFR_som16	30
EGFR_som32	30
EGFR_som39	30
EGFR_som21	50

Set this before “GO”

**Save Source Features** ☒

In the output will be a file **EGFR-locus\_som21.gbout**

```
LOCUS      7                               0 bp      DNA                HTG 27-FEB-2022
DEFINITION Homo sapiens chromosome 7 GRCh38 partial sequence 55018017..55212628
            reannotated via Ensembl.
ACCESSION  chromosome:GRCh38:7:55018017:55212628:1
VERSION    chromosome:GRCh38:7:55018017:55212628:1
KEYWORDS   .
SOURCE     .
ORGANISM   .
FEATURES   Location/Qualifiers
            source          1..194612
                               /organism="Homo sapiens"
                               /db_xref="taxon:9606"
            gene           1001..193612
                               /gene="ENSG00000146648.20"
                               /locus_tag="EGFR"
                               /note="epidermal growth factor receptor [Source:HGNC
            variation      156760..156768
                               /replace="TTAAGAGAA/-"
                               /db_xref="som21_1:c.2239_2247"
                               /global_range="GRCh38:7:55174776:55174784:1"

ORIGIN
```

Note that the location calculated by the application matches the assigned [COSM6218](#) entry

Somatic mutation: COSV51780076

**COSV51780076** SOMATIC DELETION

Most severe consequence	<b>coding sequence variant</b>   <a href="#">See all predicted consequences</a>
Alleles	<b>COSMIC_MUTATION</b>   Ancestral: TTAAGAGAA
Change tolerance	GERP: 3.54
Location	<a href="#">Chromosome 7:55174776-55174784</a> (forward strand)   VCF: 7

NB: The header is different from the existing curated variants, but matches that of the reference **EGFR\_locseq.gb**

**Subtype 24 - c.2239\_2253>aat [COSM51503](#)**

Reference Sequence
15 bases: TTAAGAGAAGCAACA
Variant Sequence
AAT
Haplotype Definition Name
som24
Variant Name
c. 2239_2253

```
variation      156760..156774
               /replace="TTAAGAGAAGCAACA/AAT"
               /db_xref="som24_1:c.2239_2253"
               /global_range="GRCh38:7:55174776:55174790:1"
```

Somatic mutation: COSV51779474

**COSV51779474** SOMATIC SEQUENCE ALTERATION

Most severe consequence	<b>coding sequence variant</b>   <a href="#">See all predicted consequences</a>
Alleles	<b>COSMIC_MUTATION</b>   Ancestral: TTAAGAGAAGCAACG
Change tolerance	GERP: 3.54
Location	<a href="#">Chromosome 7:55174776-55174790</a> (forward strand)

A preferable method would be to use VCF files, but the application is not currently set up for this.

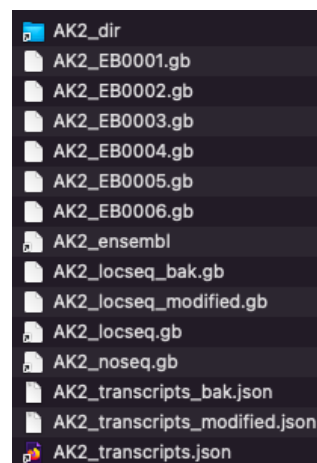
## For AK2 in GRCh37

The MANE\_Select transcript for AK2 is not defined in GRCh37, but exists in GRCh38 with the identifier `ENST00000672715`

After manually identifying the differing offset-positions between GRCh37 and GRCh38 for exons shared with other transcripts, it was possible to create mRNA and CDS join features for the missing features in GRCh37. To verify this, the transcripts created from each version by the application were aligned. The sequence for this transcript differs by a single base substitution at one location between the two genome builds.

```
AK2_locseq.gb -> AK2_locseq_modified.gb
AK2_locseq_bak.gb
AK2_locseq_modified.gb
AK2_noseq.gb -> AK2_dir/AK2_noseq.gb
AK2_transcripts.json -> AK2_transcripts_modified.json
AK2_transcripts_bak.json
AK2_transcripts_modified.json

AK2_locseq_bak.gb is the original version of AK2_locseq_bak
```



The new file is maintained as `AK2_locseq_modified.gb` with `AK2_locseq.gb` now a soft-link to it and the original saved as `AK2_locseq_bak.gb`

Comparing the feature tables between the two builds:

### **AK2\_locseq\_modified.gb for GRCh37**

Note the modified transcript identifier: from `"ENST00000672715.1"` to `"ENST00000672715m.1"`

The 'm' modifier is recognised in the GUI to generate a URL to link to Ensembl GRCh38 instead of GRCh37.

```
LOCUS      1 75013 bp DNA HTG 12-AUG-2022
DEFINITION Homo sapiens chromosome 1 GRCh37 partial sequence 33472585..33547597 reannotated
            via Ensembl
ACCESSION  chromosome:GRCh37:1:33472585:33547597:1
VERSION    chromosome:GRCh37:1:33472585:33547597:1
FEATURES   Location/Qualifiers
            source          1..75013
                               /organism="Homo sapiens"
                               /db_xref="taxon:9606"
            gene            complement(1001..74013)
                               /gene=ENSG00000004455.12
                               /locus_tag="AK2"
            mRNA             /note="adenylate kinase 2 [Source:HGNC Symbol;Acc:362]"
                               join(complement(29753..29887),complement(17459..17584),
                               complement(14610..14720),complement(14384..14478),
                               complement(7539..7611),complement(1003..6419))
                               /gene="ENSG00000004455.12"
                               /standard_name="ENST00000672715m.1"
                               /comment="RG:copied from GRCh38 as not present in 37 download"
            CDS              join(complement(29753..29845),complement(17459..17584),
                               complement(14610..14720),complement(14384..14478),
                               complement(7539..7611),complement(6198..6419))
                               /gene="ENSG00000004455.12"
                               /protein_id="ENSP00000499935.1"
                               /note="transcript_id=ENST00000672715m.1"
```

## AK2\_locseq.gb for GRCh38:

```
LOCUS       1 75011 bp DNA HTG 19-AUG-2022
DEFINITION  Homo sapiens chromosome 1 GRCh38 partial sequence 33006986..33081996 reannotated
            via Ensembl
ACCESSION   chromosome:GRCh38:1:33006986:33081996:1
VERSION     chromosome:GRCh38:1:33006986:33081996:1
FEATURES             Location/Qualifiers
     source          1..75011
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
     gene            complement(1001..74011)
                     /gene=ENSG00000004455.18
                     /locus_tag="AK2"
                     /note="adenylate kinase 2 [Source:HGNC
                     Symbol;Acc:HGNC:362]"
...
     mRNA            join(complement(29751..29898),complement(17457..17582),
                     complement(14608..14718),complement(14382..14476),
                     complement(7537..7609),complement(1001..6417))
                     /gene="ENSG00000004455.18"
                     /standard_name="ENST00000672715.1"
     CDS              join(complement(29751..29843),complement(17457..17582),
                     complement(14608..14718),complement(14382..14476),
                     complement(7537..7609),complement(6196..6417))
                     /gene="ENSG00000004455.18"
                     /protein_id="ENSP00000499935.1"
                     /note="transcript_id=ENST00000672715.1"
```

The file `AK2_transcripts.json` has been modified to incorporate the additional, modified transcript identifier. The original, modified and soft-linked files are managed in the same manner as for `AK2_locseq.gb`.

## Adding multiple new curation folders

If you are starting from scratch and do **not** already have the directory

`$root/GRCH38_sequences_1000_curation`, then

a) Modify the path definition at the top of the script

```
$root/helper_scripts/create_curation_tree_38.sh:
```

```
root="/Users/caryodonnell/Documents/repositories/rg_exploder_shared/
data_sources/"
datadir="GRCH38_sequences_1000"
```

b) Execute this script and it will build a set of directories in

`$root/GRCH38_sequences_1000_curation`, with the same names as in

`$root/GRCH38_sequences_1000`

It also creates soft links from `GRCH38_sequences_1000` as described previously

c) If you *already have* a **curation** directory, and execute the script anyway: you will get many error-reports about files already created.

You should find that any new locus directories in `$datadir` will be created within the curation directory.



## ***Adding multiple new input and output folders***

If you are starting from scratch and do **not** already have the directories

`$root/exploder_input_38_1000` and `$root/exploder_output_38_1000` then

a) Modify the path definition at the top of the script

```
$root/helper_scripts/create_input_directory_38.sh:
root="/Users/caryodonnell/Documents/repositories/rg_exploder_shared/
data_sources/"
datadir="GRCH38_sequences_1000"
base_input_seq="$root"exploder_input_38_1000"
base_output_seq="$root"exploder_output_38_1000"
```

b) Execute this script and it will build a set of directories in

`$root/exploder_input_38_1000` and `root/exploder_output_38_1000`

with the same names as in `$root/GRCH38_sequences_1000`

It also creates soft links from `GRCH38_sequences_1000_curation` as described previously

c) If you *already have* an input and output directory, and execute the script anyway: you will get many error-reports about files already created.

You should find that any new locus directories in `$datadir` will be created within the input and output directories.

## ***Mashup time***

`$root/data_sources/GRCH38_sequences_1000_curation/loci.json`

is a concatenation of all the individual

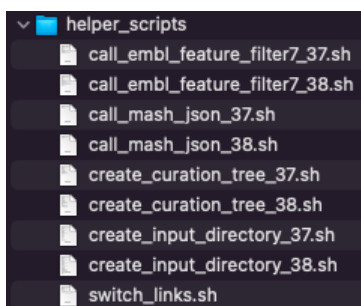
`GRCH38_sequences_1000_curation/$locus_curation/$locus_transcripts.json` files

To do this:

```
sh $root/helper_scripts/call_mash_json.sh
```

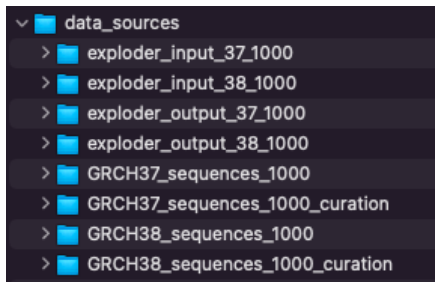
```
dataroot="/Users/caryodonnell/Documents/repositories/rg_exploder_shared/data_sources/"
pythonroot="/Users/caryodonnell/Documents/repositories/rg_exploder_shared/helper_python"
targetdir="GRCH38_sequences_1000"
```

## ***The helper-scripts directory***

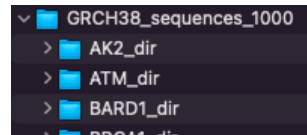


# Directory management

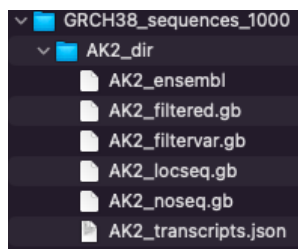
## A) Screen shots from MacOS Finder



At the root level are these directories



**GRCH38\_sequences\_1000/**  
**locus\_dir** initially hold the downloaded **ensembl.txt.gz** files of each locus



After automated processing of **ensembl.txt.gz**, these files remain in each **locus\_dir** sub directory:

**locus\_ensembl** – the original, now unzipped, downloaded file

**locus\_filtered.gb** – same as **locus\_ensembl**, minus unwanted annotation

**locus\_filtvar.gb** – as above, but without mRNA & CDS features, without sequence ; just the variations

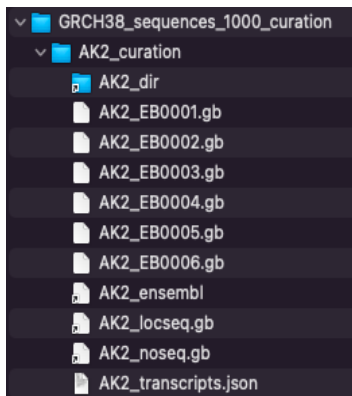
**locus\_locseq.gb** – everything in **locus\_filtered.gb**, without the variations; this is the Reference Source used in the application.

**locus\_noseq.gb** – A no-sequence, no-features version. This is used as the input for **locus\_00000** haplotype in the application. It is a blank template for curating other haplotypes. Find known variants in **locus\_locseq.gb**

**locus\_transcripts.json** – lookup information for the application. These are later concatenated into config.json

A curation directory is used to demarcate the automated-processed files from the hand-created haplotye definitions.

Here: **AK2\_EB0001.gb ... AK2\_EB0006.gb**



Other files in this **locus\_curation** directory are usually soft linked to the above. There may be exceptions to this, notably in

**GRCH37\_sequences\_1000\_curation** - the original **AK2\_locseq.gb** and **AK2\_transcripts.json** have been hand-edited to include the MANE\_Select transcript that is absent in GRCH37, but is defined in GRCH37.

### **exploder-input\_38\_1000**

Where all files in

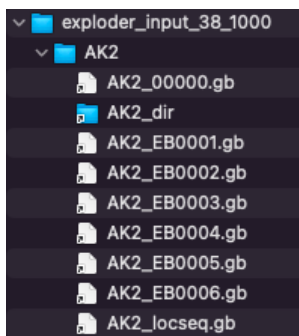
**exploder\_input\_38\_1000/locus**

are soft links to files in

**GRCH38\_sequences\_1000\_curation/locus\_curation**

This allows for hand-culling or renaming of links in the input directory whilst retaining them in the curation directory or under a different name. Eg: compare the EGFR directories.

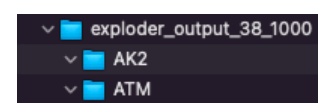
The **exploder\_python/input** directory is a soft link to here



### Finally – **exploder-output\_38\_1000**

Where the sub-directories for each locus are empty to receive the output from the application.

The **exploder\_python/output** directory is a soft link to here



## Recorded information on loci

NB: positions & annotation seems to change with Ensembl release versions. There may be inconsistency in the *numbers* stated within the URL ( $r=x:nnn-nnn$ ) for the Ensembl ID and the genomic *location* this actually links to, but the URL seems to correct itself over time.


Note that it's normal for Ensembl and NCBI to show different position mappings. It's typically small numerical-differences at start and end. All below are Ensembl mappings

**Ensembl Release 96 Apr 2019, GRCh38.p12 – [GENCODE release 30](#)**

OMIM	HGNC symbol report	Ensembl ID	Genomic Location GRCh38.p12	Publication	NIH genetics home reference
<a href="#">ATM</a>	<a href="#">HGNC:795</a>	<a href="#">ENSG00000149311</a>	<a href="#">11: 108,219,552-108,372,034 (+)</a>		ATM serine/threonine kinase ataxia telangiectasia
<a href="#">BARD1</a>	<a href="#">HGNC:952</a>	<a href="#">ENSG00000138376</a>	<a href="#">2: 214,723,966-214,811,363 (-)</a>		<a href="#">BRCA1 associated RING domain 1</a>
<a href="#">BRCA1</a>	<a href="#">HGNC:1100</a>	<a href="#">ENSG00000012048</a>	<a href="#">17: 43,041,776-43,172,764 (-)</a>		DNA repair associated
<a href="#">BRCA2</a>	<a href="#">HGNC:1101</a>	<a href="#">ENSG00000139618</a>	<a href="#">13: 32,313,779-32,401,961 (+)</a>		<a href="#">DNA repair associated</a>
<a href="#">BRIP1</a>	<a href="#">HGNC:20473</a>	<a href="#">ENSG00000136492</a>	<a href="#">17: 61,677,621-61,867,166 (-)</a>		<a href="#">BRCA1 interacting protein C-terminal helicase 1</a>
<a href="#">CDH1</a>	<a href="#">HGNC:1748</a>	<a href="#">ENSG00000039068</a>	<a href="#">16: 68,735,328-68,837,505 (+)</a>		<a href="#">cadherin 1</a>
<a href="#">CDK12</a>	<a href="#">HGNC:24224</a>	<a href="#">ENSG00000167258</a>	<a href="#">17: 39,459,444-39,566,974 (+)</a>		cyclin dependent kinase 12
<a href="#">CHEK1</a>	<a href="#">HGNC:1925</a>	<a href="#">ENSG00000149554</a>	<a href="#">11: 125,624,114-125,677,277 (+)</a>		checkpoint kinase 1
<a href="#">CHEK2</a>	<a href="#">HGNC:16627</a>	<a href="#">ENSG00000183765</a>	<a href="#">22: 28,686,650-28,743,515 (-)</a>	<a href="#">Breast Cancer (Dove Med Press). 2017; 9: 331–335.</a>	<a href="#">checkpoint kinase 2</a>
<a href="#">EPCAM</a>	<a href="#">HGNC:11529</a>	<a href="#">ENSG00000119888</a>	<a href="#">2: 58,157,601-58,243,014 (+)</a>		epithelial cell adhesion molecule
<a href="#">FANCL</a>	<a href="#">HGNC:20748</a>	<a href="#">ENSG00000115392</a>	<a href="#">2: 58,159,243-58,241,372 (-)</a>		FA complementation group L
<a href="#">KRAS</a>	<a href="#">HGNC:6407</a>	<a href="#">ENSG00000133703</a>	<a href="#">12: 25,203,867-25,251,858 (-)</a>		KRAS proto-oncogene, GTPase
<a href="#">MLH1</a>	<a href="#">HGNC:7127</a>	<a href="#">ENSG00000076242</a>	<a href="#">3: 36,992,181-37,052,069 (+)</a>		<a href="#">mutL homolog 1 - DNA repair</a>
<a href="#">MSH2</a>	<a href="#">HGNC:7325</a>	<a href="#">ENSG00000095002</a>	<a href="#">2: 47,397,766-47,668,349 (+)</a>		mutS homolog 2
<a href="#">MSH6</a>	<a href="#">HGNC:7329</a>	<a href="#">ENSG00000116062</a>	<a href="#">2: 47,693,239-47,812,392 (+)</a>		mutS homolog 6
<a href="#">NBN</a>	<a href="#">HGNC:7652</a>	<a href="#">ENSG00000104320</a>	<a href="#">8: 89,931,939-90,004,625 (-)</a>		nibrin
<a href="#">NF1</a>	<a href="#">HGNC:7765</a>	<a href="#">ENSG00000196712</a>	<a href="#">17: 31,089,184-31,387,859 (+)</a>		Neurofibromin 1
<a href="#">PALB2</a>	<a href="#">HGNC:26144</a>	<a href="#">ENSG00000083093</a>	<a href="#">16: 23,602,397-23,642,073 (-)</a>		partner and localizer of BRCA2
<a href="#">PMS2</a>	<a href="#">HGNC:9122</a>	<a href="#">ENSG00000122512</a>	<a href="#">7: 5,970,162-6,009,869 (-)</a>		PMS1 homolog 2, mismatch repair system component
<a href="#">PPP2R2A</a>	<a href="#">HGNC:9304</a>	<a href="#">ENSG00000221914</a>	<a href="#">8: 26,289,885-26,374,303 (+)</a>		protein phosphatase 2 regulatory subunit Balpha
<a href="#">PTEN</a>	<a href="#">HGNC:9588</a>	<a href="#">ENSG00000171862</a>	<a href="#">10: 87,861,459-87,974,096 (+)</a>		phosphatase and tensin homolog
PTEN_a	<a href="#">HGNC:9588</a>	<a href="#">ENSG00000284792</a>	<a href="#">CHR_HG2334_PATCH: 87,861,382-87,968,399 (+)</a>		phosphatase and tensin homolog (alternative mapping)
<a href="#">RAD51B</a>	<a href="#">HGNC:9822</a>	<a href="#">ENSG00000182185</a>	<a href="#">14: 67,801,571-68,748,426 (+)</a>		RAD51 paralogs B
<a href="#">RAD51C</a>	<a href="#">HGNC:9820</a>	<a href="#">ENSG00000108384</a>	<a href="#">17: 58,691,713-58,736,471 (+)</a>		RAD51 paralogs C

<a href="#">RAD51D</a>	<a href="#">HGNC:9823</a>	<a href="#">ENSG00000185379</a>	<a href="#">17: 35,091,622-35,122,108 (-)</a>		RAD51 paralog D
RAD54L	<a href="#">HGNC:9826</a>	<a href="#">ENSG00000085999</a>	<a href="#">1: 46,247,073-46,279,088 (+)</a>		RAD54 like
STK11	<a href="#">HGNC:11389</a>	<a href="#">ENSG00000118046</a>	<a href="#">19: 1,176,541-1,229,452(+)</a>		serine/threonine kinase 11
TP53	<a href="#">HGNC:11998</a>	<a href="#">ENSG00000141510</a>	<a href="#">17: 7,661,264-7,688,065 (-)</a>		tumor protein p53

For clinical evidence on variations, from HGNC symbol report: click on Clinical Resources/Clinvar  
For locations: Clinical Resources/Genetic Testing Registry

Clinical resources ?		
OMIM	<a href="#">604373</a>	LRG <a href="#">LRG_302</a> 
Genetics Home Reference	<a href="#">Search via CHEK2</a>	DECIPHER <a href="#">Search via CHEK2</a>
ClinGen	<a href="#">Search via CHEK2</a>	Genetic Testing Registry <a href="#">Search via NCBI Gene ID 11200</a>
ClinVar	<a href="#">Search via NCBI Gene ID 11200</a>	dbVar <a href="#">Search via NCBI Gene ID 11200</a>
Orphanet	<a href="#">119394</a>	COSMIC <a href="#">CHEK2</a>

NB: Orphanet link has “diagnostic tests” link that lists laboratories testing for this gene.

## Alternative possibilities:

a) Use Entrez API as an alternative to a manual selection: direct extraction of locus into application?  
Entrez API requires a specific account, but is an obvious development

b) Ensembl Biomart

Do online search: tutorials etc, but beware broken links and not-working notices

Use Biomart to download only non-synonymous SNPs and indels?

Eg: for ATM

```
http://www.ensembl.org/biomart/martview/6f94c204c7331054b28277e0ed89d23c?
VIRTUALSCHEMANAME=default&ATTRIBUTES=hsapiens_snp.default.snp.refsnid|
hsapiens_snp.default.snp.refsnid_source|hsapiens_snp.default.snp.chr_name|
hsapiens_snp.default.snp.chrom_start|hsapiens_snp.default.snp.chrom_end|
hsapiens_snp.default.snp.consequence_type_tv|
hsapiens_gene_ensembl.default.snp.ensembl_gene_id_version&FILTERS=hsapiens_snp.d
efault.filters.chr_name."11"|
hsapiens_snp.default.filters.so_mini_parent_name."nonsynonymous_variant"&VISIBLE
PANEL=resultspanel
```

Your reference is martquery\_0704132240\_850.txt.gz.

## Modifications log

Date	Section	Changes/reasons
7 <sup>th</sup> February 2025	“Automated data processing for downloaded Ensembl data”	Documenting additional processing Python script, which supersedes previous.
28th January 2025		Corrections
27 <sup>th</sup> January 2025	New Sections: “Downloading a sequence file for a locus from NCBI” “Data processing for downloaded NCBI data”	Section similar to that for existing Ensembl download, re-written for NCBI-derived data
6 <sup>th</sup> May 2023	First version	