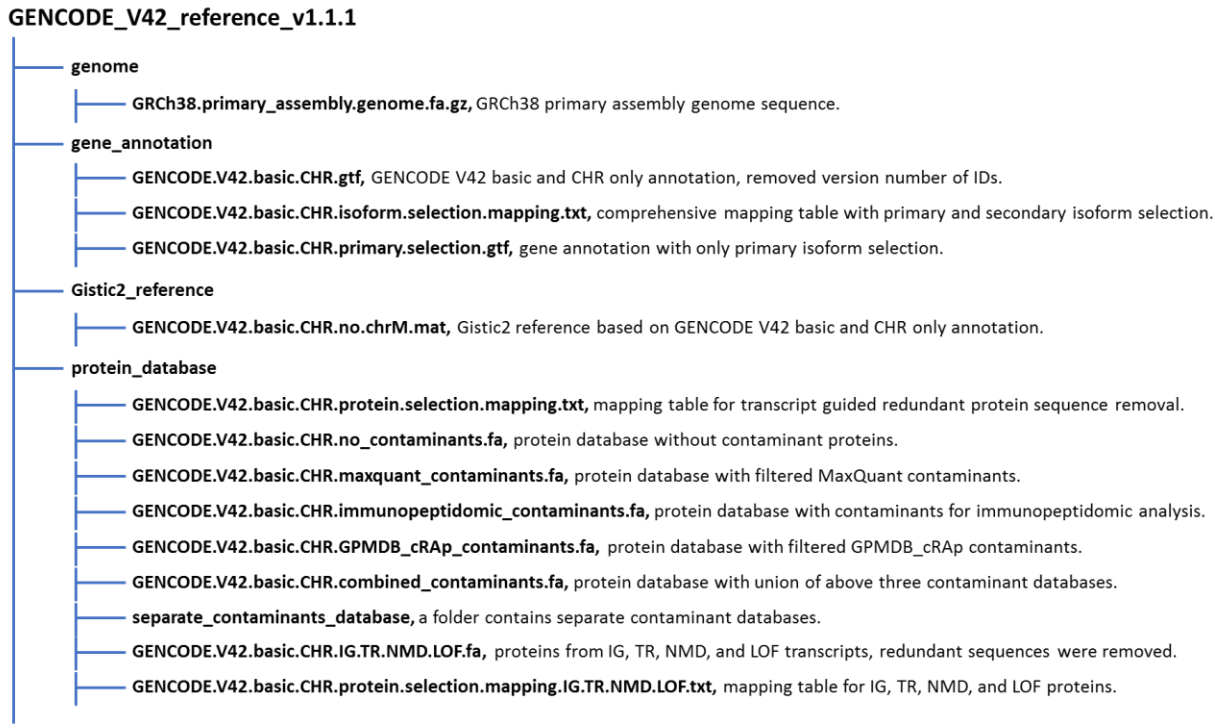
# GENCODE V42 reference preparation for proteogenomics data processing and downstream analysis

Yongchao Dou, Bing Zhang, and the reference working group

## 1. Structure of the ENCODE_V42_reference_v1.1.1 folder




## 2. Reference files

Reference files were download from GENCODE and the human release 42 basic and CHR only annotation was selected. Links of reference files used in reference preparation are listed below.

**GENCODE V42**
https://www.gencodegenes.org/human/release_42.html

**Genome sequence: GRCh38 primary assembly**
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/GRCh38.primary_assembly.genome.fa.gz

**Gene annotation: GENCODE V42, basic, CHR only**
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/gencode.v42.basic.annotation.gtf.gz

1. It contains the basic gene annotation on the reference chromosomes only
2. This is a subset of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene

**Entrez gene ids mapping**
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/gencode.v42.metadata.EntrezGene.gz

**Entrez gene ids mapping**
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/gencode.v42.metadata.EntrezGene.gz

**Gene symbol mapping**
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/gencode.v42.metadata.HGNC.gz

**RefSeq mapping**
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/gencode.v42.metadata.RefSeq.gz

**SwissProt mapping**
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/gencode.v42.metadata.SwissProt.gz

**Protein sequence**
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/gencode.v42.pc_translations.fa.gz

**HGNC gene description**
Download on 11/15/2022 from https://www.genenames.org/download/custom/, approved symbols only.

# 3. Statistics of GENCODE Release 42 Basic, CHR

This annotation includes 62,696 genes and 117,681 transcripts. Statistics of biotypes at gene and transcript levels are shown in Table 1.

**Table 1**: Statistics of biotypes at gene and transcript levels

| idx | Gene | Transcript |
|---|---|---|
| protein_coding | 20030 | 63723 |
| lncRNA | 18879 | 29132 |
| processed_pseudogene | 10149 | 10150 |
| unprocessed_pseudogene | 2607 | 2608 |
| misc_RNA | 2212 | 2212 |
| snRNA | 1901 | 1901 |
| miRNA | 1879 | 1879 |
| TEC | 1054 | 1056 |
| transcribed_unprocessed_pseudogene | 961 | 961 |
| snoRNA | 942 | 942 |
| transcribed_processed_pseudogene | 512 | 512 |
| rRNA_pseudogene | 497 | 497 |
| IG_V_pseudogene | 187 | 187 |
| transcribed_unitary_pseudogene | 153 | 154 |
| IG_V_gene | 145 | 145 |
| TR_V_gene | 107 | 107 |
| unitary_pseudogene | 99 | 98 |
| TR_J_gene | 79 | 79 |
| scaRNA | 49 | 49 |
| rRNA | 47 | 47 |
| IG_D_gene | 37 | 37 |
| TR_V_pseudogene | 33 | 33 |
| Mt_tRNA | 22 | 22 |
| artifact | 19 | 19 |
| IG_J_gene | 18 | 18 |
| pseudogene | 15 | 15 |
| IG_C_gene | 14 | 14 |
| IG_C_pseudogene | 9 | 9 |
| ribozyme | 8 | 8 |
| TR_C_gene | 6 | 6 |
| sRNA | 5 | 5 |
| TR_D_gene | 4 | 4 |
| TR_J_pseudogene | 4 | 4 |
| IG_J_pseudogene | 3 | 3 |
| translated_unprocessed_pseudogene | 3 | 3 |
| Mt_rRNA | 2 | 2 |
| translated_processed_pseudogene | 2 | 2 |
| IG_pseudogene | 1 | 1 |
| scRNA | 1 | 1 |
| vault_RNA | 1 | 1 |
| nonsense_mediated_decay | NA | 305 |
| processed_transcript | NA | 601 |
| protein_coding_LoF | NA | 71 |
| retained_intron | NA | 58 |

## 4. Representative isoform selection for protein coding genes

Swiss-Prot and MANE Select are two major efforts that aim to select one high-quality representative transcript for each protein-coding gene. Leveraging these two resources, we developed a workflow (Figure 1) and selected one representative primary isoform for each protein coding gene (see section 6. Comprehensive mapping table). Moreover, 28 genes are associated with multiple Swiss-Prot proteins, and all non-primary proteins and their longest transcripts were designated as secondary isoforms for these genes. The MANE Plus Clinical isoforms were added as supplements of MANE Selects for clinical variant reporting. If a MANE Plus Clinical and a secondary isoform from Swiss-Prot correspond to the same proteins but different transcripts of a gene, the MANE Plus Clinical selected isoform was used as the secondary selected isoform.

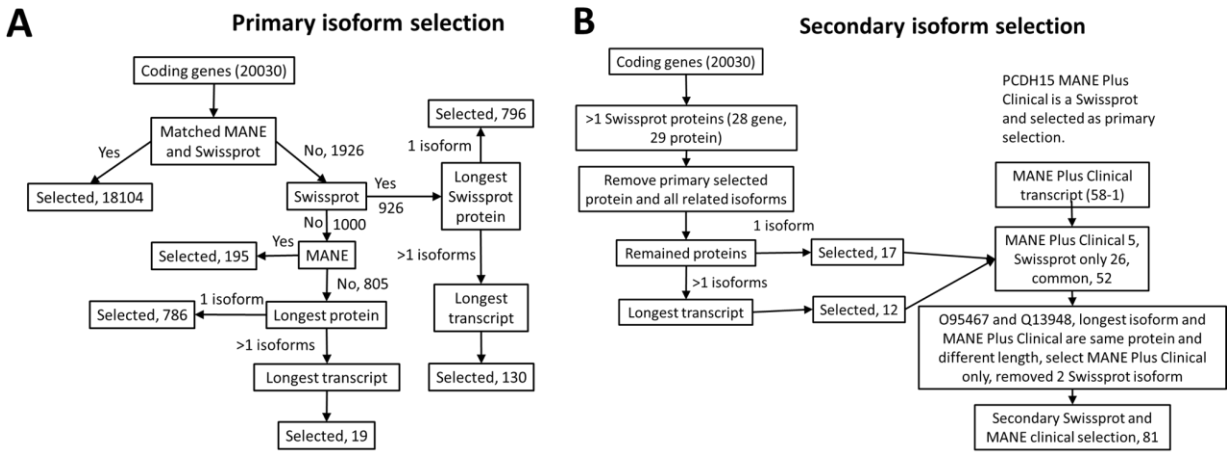



**Figure 1:** Workflow for the selection of representative isoforms for each protein coding gene.

## 5. Representative isoform selection for non-coding genes

Longest isoform was selected as primary isoform for noncoding genes. There are 5 non-coding genes with multiple isoforms with the same length which are ELOA3CP, LINC02277, ENSG00000256916, ENSG00000284719, and ENSG00000291232. These isoforms were alphabetically ordered to select primary isoform for each gene.

## 6. Unique gene name assignment for display

To facilitate human understanding of the Ensembl gene IDs, we used corresponding HUGO gene symbols as display names. When multiple Ensembl gene IDs are mapped to the same gene symbol, each ID was assigned a unique gene name for display using the following method. First, all primary selected isoforms were ordered by presence of a SwissProt ID that mapped to the

Ensembl protein ID, transcript ID listed in the MANE plus Clinical annotation, longer CCDS length, longer transcript length, and finally, alphabetic order of the Ensembl gene ID. The first Ensembl gene ID was assigned the gene symbol (e.g., MATR3 for ENSG00000015479) and all following had the Ensembl gene ID appended (e.g., MATR3_ENSG00000280987 for ENSG00000280987). For PAY_Y genes (e.g., ENSG00000002586_PAR_Y), a tag "_PAR_Y" was added to the end of gene names (e.g., CD99_PAR_Y).

## 7. Comprehensive mapping table

A comprehensive mapping table was generated for use's convenience. The table is based on selected gene annotation and matched mapping files. Details of each column are shown here.

1. Gene_id: ensembl gene id from GENCODE V42 Basic (CHR)
2. Transcript_id: ensembl transcript id from GENCODE V42 Basic (CHR)
3. Protein_id: ensembl protein id from GENCODE V42 Basic (CHR)
4. Gene_name: gene name from GENCODE V42 Basic (CHR)
5. Unique_gene_name_for_display: curated unique gene name for display
6. Gene_type: biotype of gene from GENCODE V42 Basic (CHR)
7. Transcript_type: biotype of transcript from GENCODE V42 Basic (CHR)
8. CCDS_id: CCDS id from GENCODE V42 Basic (CHR)
9. MANE_select: MANE selected transcript from GENCODE V42 Basic (CHR)
10. MANE_Plus_Clinical: MANE Plus Clinical transcript from GENCODE V42 Basic (CHR)
11. EntrezGene: matched gene Entrez gene ID from Entrez gene ids mapping table
12. RefSeq_RNA: matched RefSeq transcript ID from RefSeq mapping table
13. RefSeq_protein: matched RefSeq protein ID from RefSeq mapping table
14. SwissProt: matched SwissProt protein ID from SwissProt mapping table
15. Transcript_length: length of transcript
16. CCDS_length: length of CCDS
17. Primary_select: primary representative isoforms for each gene
18. Secondary_select: secondary representative isoforms for each gene
19. Gene_id_versioned: ensembl gene id with version number
20. Transcript_id_versioned: ensembl transcript id with version number
21. Protein_id_versioned: ensembl protein id with version number
22. Gene_description: gene description from HGNC

## 8. Matched Gistic2 reference

A rg table of transcripts, except transcripts from chrM, were extracted from GENCODE V42 Basic (CHR) annotation (Figure 2). Hg38 cytoBand were downloaded from UCSC. Then rg and cytoBand tables were reformat following Gistic2 reference format by inhouse Matlab script.

| refseq | gene | symb | locus_id | chr | strand | start | end | cds_start | cds_end | status | chrn |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ENST00000456328 | ENSG00000290825 | ENSG00000290825 | 62187 | chr1 | 1 | 11869 | 14409 | 14409 | 14409 | reviewed | 1 |
| ENST00000450305 | ENSG00000223972 | ENSG00000223972 | 24985 | chr1 | 1 | 12010 | 13670 | 13670 | 13670 | reviewed | 1 |
| ENST00000488147 | ENSG00000227232 | ENSG00000227232 | 27224 | chr1 | 0 | 14404 | 29570 | 29570 | 29570 | reviewed | 1 |
| ENST00000619216 | ENSG00000278267 | ENSG00000278267 | 54568 | chr1 | 0 | 17369 | 17436 | 17436 | 17436 | reviewed | 1 |
| ENST00000473358 | ENSG00000243485 | ENSG00000243485 | 36700 | chr1 | 1 | 29554 | 31097 | 31097 | 31097 | reviewed | 1 |
| ENST00000469289 | ENSG00000243485 | ENSG00000243485 | 36700 | chr1 | 1 | 30267 | 31109 | 31109 | 31109 | reviewed | 1 |
| ENST00000607096 | ENSG00000284332 | ENSG00000284332 | 57280 | chr1 | 1 | 30366 | 30503 | 30503 | 30503 | reviewed | 1 |
| ENST00000417324 | ENSG00000237613 | ENSG00000237613 | 34243 | chr1 | 0 | 34554 | 36081 | 36081 | 36081 | reviewed | 1 |
| ENST00000606857 | ENSG00000268020 | ENSG00000268020 | 49550 | chr1 | 1 | 52473 | 53312 | 53312 | 53312 | reviewed | 1 |
| ENST00000642116 | ENSG00000290826 | ENSG00000290826 | 62188 | chr1 | 1 | 57598 | 64116 | 64116 | 64116 | reviewed | 1 |
| ENST00000492842 | ENSG00000240361 | ENSG00000240361 | 35434 | chr1 | 1 | 62949 | 63887 | 63887 | 63887 | reviewed | 1 |
| ENST00000641515 | ENSG00000186092 | ENSG00000186092 | 16146 | chr1 | 1 | 65419 | 71585 | 65565 | 70005 | reviewed | 1 |
| ENST00000466430 | ENSG00000238009 | ENSG00000238009 | 34517 | chr1 | 0 | 89295 | 120932 | 120932 | 120932 | reviewed | 1 |
| ENST00000495576 | ENSG00000239945 | ENSG00000239945 | 35267 | chr1 | 0 | 89551 | 91105 | 91105 | 91105 | reviewed | 1 |
| ENST00000610542 | ENSG00000238009 | ENSG00000238009 | 34517 | chr1 | 0 | 120725 | 133723 | 133723 | 133723 | reviewed | 1 |
| ENST00000442987 | ENSG00000233750 | ENSG00000233750 | 31627 | chr1 | 1 | 131025 | 134836 | 134836 | 134836 | reviewed | 1 |
| ENST00000494149 | ENSG00000268903 | ENSG00000268903 | 49829 | chr1 | 0 | 135141 | 135895 | 135895 | 135895 | reviewed | 1 |
| ENST00000595919 | ENSG00000269981 | ENSG00000269981 | 50199 | chr1 | 0 | 137682 | 137965 | 137965 | 137965 | reviewed | 1 |
| ENST00000493797 | ENSG00000239906 | ENSG00000239906 | 35248 | chr1 | 0 | 139790 | 140339 | 140339 | 140339 | reviewed | 1 |
| ENST00000484859 | ENSG00000241860 | ENSG00000241860 | 36043 | chr1 | 0 | 141474 | 149707 | 149707 | 149707 | reviewed | 1 |

**Figure 2:** rg table for all transcripts.

# 9. Matched protein database

## 9.1 Redundant protein sequence removal

Matched protein database, gencode.v42.pc_translations.fa, was download from GENCODE. Only proteins from **coding transcripts** in GENCODE V42 Basic (CHR) annotation were retained (63,723 proteins) and others were discarded (47,330 proteins). To remove redundant protein sequences, we developed a transcript guided workflow to pick representable protein IDs for proteins with the same sequences (Figure 3). The coding transcripts were grouped into 50,827 transcript groups with the same protein sequences. If a transcript group has only one transcript, the transcript is selected (43,419). If a transcript group has multiple transcripts, all transcripts were ordered by presence of primary selection, secondary selection, Swissprot mapped, MANE selected, longest transcript and alphabetic order of transcript IDs. If multiple transcripts passed a criterion, the longest transcript will be selected. Then only proteins encoded from selected transcripts were kept.

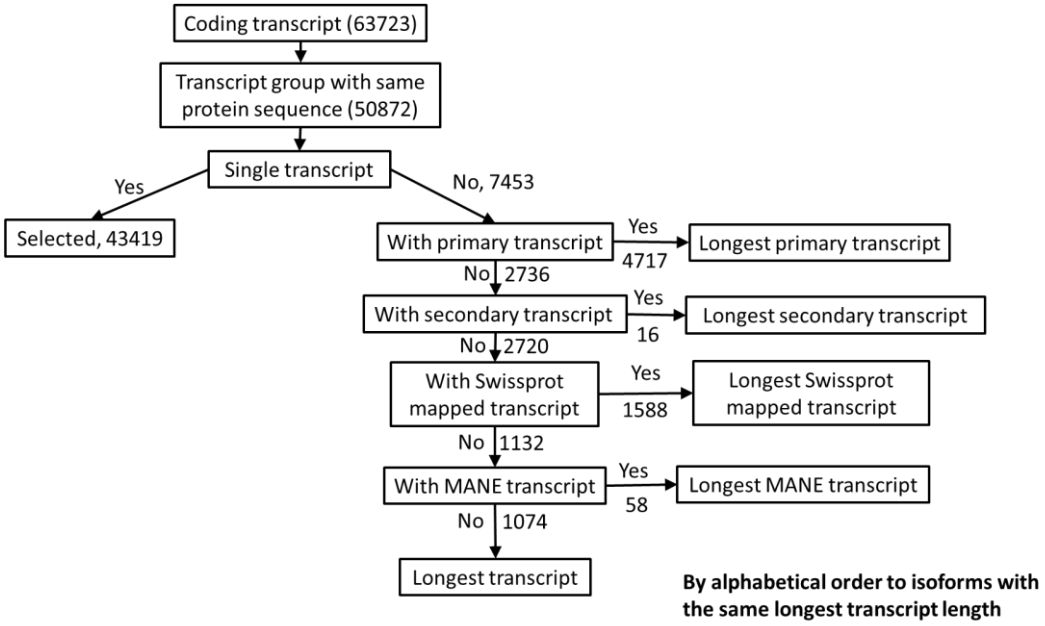



**Figure 3:** Transcript guided redundant protein sequence removal.

### 9.2 Evaluation of contaminants

We evaluated three protein contaminant databases which are GPMDB_cRAp (118), MaxQuant (246), and contaminants for immunopeptidomic from Karl (642), respectively.

The MaxQuant contaminant database has 246 unique IDs and 245 unique sequences. Q8N1N4-2 was removed as it has the same sequence with Q7RTT2. Then we removed 42 contaminant proteins which with the same sequences with regular proteins. There are 203 proteins remained in MaxQuant database after filtering. The GPMDB_cRAp contaminant database has 118 unique IDs and 116 unique protein sequences. P0DUB6|AMY1A_HUMAN, P0DTE8|AMY1C_HUMAN, and P0DTE7|AMY1B_HUMAN have the same protein sequence and only P0DUB6|AMY1A_HUMAN was kept. Then we removed 67 contaminant proteins which with the same sequences with regular proteins. The filtered GPMDB_cRAp database has 49 protein sequences remained. No proteins were filter out from Karl's database based on above method. While the overlap is still low between the three contaminant databases (Figure 4). Then we generated a combined contaminant database with union protein sequences (817) from the three filtered contaminant databases.

**4 common proteins:**
P00761|TRYP_PIG Trypsin
P02769|ALBU_BOVIN Albumin
P0C1U8|SSPA_STAAU Glutamyl endopeptidase
P00766|CTRA_BOVIN Chymotrypsinogen A

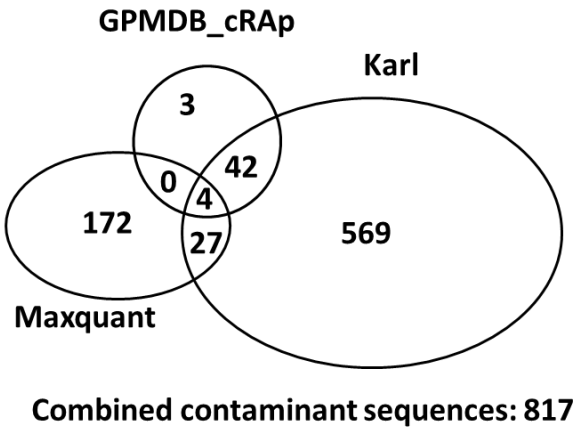



**Figure 4:** Protein contaminant database comparisons.

The headers of regular proteins were format as:
> protein_ID|transcript_ID|gene_ID|gene_symbol gene_description.
ENSP00000510254 of KRAS is shown here as an example:
>ENSP00000510254|ENST00000692768|ENSG00000133703|KRAS KRAS proto-oncogene, GTPase
The headers of contaminants for immunopeptidomic from Karl is not changed. While the headers of GPMDB_cRAp, MaxQuant, and combined contaminant proteins were format as:
>Cont|header_from_contaminant_database
P01966 from MaxQuant contaminants and P00761 from GPMDB_cRAp are shown here as examples:
>Cont|P01966 SWISS-PROT:P01966 (Bos taurus) Hemoglobin subunit alpha
>Cont|sp|P00761|TRYP_PIG Trypsin OS=Sus scrofa OX=9823 PE=1 SV=1

### 9.3 Protein selection mapping table

A mapping table of redundant protein sequence removal was generated for use's convenience. Many columns are the same with isoform selection table and only the protein selection part is unique. Details of each column are shown here.

1. Gene_id: ensembl gene id from GENCODE V42 Basic (CHR)
2. Transcript_id: ensembl transcript id from GENCODE V42 Basic (CHR)

3. Protein_id: ensembl protein id from GENCODE V42 Basic (CHR)
4. Gene_name: gene name from GENCODE V42 Basic (CHR)
5. Unique_gene_name_for_display: curated unique gene name for display
6. Gene_type: biotype of gene from GENCODE V42 Basic (CHR)
7. Transcript_type: biotype of transcript from GENCODE V42 Basic (CHR)
8. CCDS_id: CCDS id from GENCODE V42 Basic (CHR)
9. MANE_select: MANE selected transcript from GENCODE V42 Basic (CHR)
10. MANE_Plus_Clinical: MANE Plus Clinical transcript from GENCODE V42 Basic (CHR)
11. EntrezGene: matched gene Entrez gene ID from Entrez gene ids mapping table
12. RefSeq_RNA: matched RefSeq transcript ID from RefSeq mapping table
13. RefSeq_protein: matched RefSeq protein ID from RefSeq mapping table
14. SwissProt: matched SwissProt protein ID from SwissProt mapping table
15. Transcript_length: length of transcript
16. CCDS_length: length of CCDS
17. Primary_select: primary representative isoforms for each gene
18. Secondary_select: secondary representative isoforms for each gene
19. Gene_id_versioned: ensembl gene id with version number
20. Transcript_id_versioned: ensembl transcript id with version number
21. Protein_id_versioned: ensembl protein id with version number
22. Protein_group_id: Unique protein sequences were alphabetically ordered, and a unique number was assigned to each protein sequence
23. Number_of_proteins: The number of proteins in a protein group
24. Protein_selection: Selected proteins for each protein group
25. Protien_sequence: Protein sequences of each protein ID
26. Gene_description: gene description from HGNC

### 9.4 Proteins from IG/TR/NMD/LOF transcripts

In GENCODE V42 Basic (CHR) annotation, there are 705 proteins from transcripts labeled as IG_C/D/J/V/_gene (IG), TR_C/J/V/_gene, nonsense_mediated_decay (NMD), and protein_coding_LoF (LOF). There are 674 proteins retained after removing redundant sequences and sequences the same regular proteins. The filtered proteins can be appended to the regular protein database as needed.