S1. Building a novel dataset: the ordering procedure

The two databases searched are UniProt (https://www.uniprot.org/), the central hub for the collection of functional information on proteins (https://www.uniprot.org/help/uniprotkb) [9] and Ensembl (https://www.ensembl.org/index.html), a genome browser for vertebrate genomes [10]. The queries in UniProt allow the download and analysis of all proteins codified by the same chromosome. Ensembl genomic complex datasets can be retrieved using the Biomart data-mining tool (http://www.ensembl.org/biomart/martview/ff9ecfb63b2cf534ed20a16879eaebb8) based on Ensembl Genes 98 (June 2019) and human genes GRCh38.p13

Step 1. UniProt Protein Table

UniProt chromosome table was downloaded from UniProt Swiss-Prot, which contains all reviewed sequences.

UniProt has a Proteome – Homo sapiens (Human) section with the list of all human chromosomes available for download (https://www.uniprot.org/proteomes/UP000005640). We selected the chromosome and then "View all proteins", "View proteins from 1 selected component", filtered by "Reviewed".

We selected the following UniProt Table attributes from Proteome Homo Sapiens (Human) UniProt section:

1. Entry: AC, primary (citable) accession number (https://www.uniprot.org/help/accession_numbers)
2. Entry name: mnemonic identifier for a UniProt entry (https://www.uniprot.org/help/entry_name)
3. Protein names: names of the protein to allow unambiguous identification of a protein (https://www.uniprot.org/help/protein_names)
4. Gene names: list of gene names assigned to a specific gene (https://www.uniprot.org/help/gene_name)
5. Length: sequence length
6. Sequence: the canonical protein sequence (https://www.uniprot.org/help/sequences)
7. Gene names (primary): recommended gene name (official gene symbol) (https://www.uniprot.org/help/gene_name)

UniProt chooses for each entry a canonical sequence based on at least one of the following criteria (https://www.uniprot.org/help/canonical_and_isoforms):

1. It is the most prevalent.
2. It is the most similar to orthologous sequences found in other species.
3. By virtue of its length or amino acid composition, it allows the clearest description of domains, isoforms, polymorphisms, post-translational modifications, etc.
4. In the absence of any information, the longest sequence is chosen.

The UniProt chromosome table gives all the information about proteins, including length and amino acid composition, but without reference to the position of the coding gene on chromosome.

Step 2. Biomart Chromosome Table

The construction of the chromosome table using Biomart [11] requires first the creation of a new dataset by selecting from the main menu:

1. the database: Ensembl Genes 102;
2. the dataset: human genes (GRCh38.p13);
3. the filters of interest, and in particular the Region Chromosome/scaffold (the number of the chromosome to be analysed).
4. the attributes-sequence: the sequence of the proteins codified by genes are obtained by selection of the button Sequences and then Peptide
5. the attributes-header information: information about peptide and gene

Step 2.1 Creation of the dataset with Biomart

From the attributes section we went into *HEADER INFORMATION* and selected a gene information list as follows:

1. Gene stable Id: Ensembl Gene Stable Identity, unambiguous and consistent across Ensembl releases, created in the form ENS[species prefix][feature type prefix][a unique eleven digit number], e.g., ENSG00000090470

2. Gene name: the official gene symbol approved by the HUGO Gene Nomenclature Committee (HGNC), which is typically a short form of the gene name. Symbols are approved in accordance with the Guidelines for Human Gene Nomenclature (https://www.genenames.org/about/), e.g., PDCD7
3. UniProtKB/Swiss-Prot ID: AC UniProt of the codified protein, a unique accession number, which is called 'Primary (citable) accession number' (https://www.UniProt.org/help/accession_numbers)
4. Gene start (bp): the starting point of the gene measured in base pair
5. Gene end (bp): the ending point of the gene measured in base pair
6. Peptide: the amino acid sequence just selected in Step 2 item 4.

All results are available in FASTA Format. We chose to export the tables in a separate file with "Unique results only" checked. The file exported was originally named mart_export.txt and was renamed following our Nomenclature System as described below.

Step 2.2 Conversion of the dataset to Excel format
We wrote a tool in Perl language (https://www.perl.org/) named CHR+chromosome number (two digits)_fasta_elab.pl to convert the original Biomart FASTA format file (called CHR+ chromosome number (two digits)_Biomart.txt) to TAB format. We obtained a file called CHR+ chromosome number (two digits)_Biomart_TAB.fa. This file was used to import data in the Excel file called CHR+chromosome number (two digits)_gene_Biomart_excel.xls.

Step 2.3 Ensembl protein columns table manipulation
The columns of the Ensembl table were manipulated to obtain a new table comparable with UniProt table. We made changes in the following columns:
1. AC protein UniProt
   Lines with empty AC UniProt (accession number) were deleted because we decided to keep only the UniProt annotated and reviewed sequences.
2. Sequence
   The character '*' at the end of the sequence was deleted where it was present
3. Length: we calculated the length of the protein based on its amino acid composition, creating the new column length. We obtained the file called CHR+chromosome number (two digits)_ gene_Biomart_excel_elab.xls

Step 3. Joining the tables: canonical UniProt sequences enriched with Ensembl gene information
UniProt table contains the canonical protein sequences but not the genes/proteins molecular position necessary to support many biological analysis, such as the walking procedure and then the Surfing analysis described below. To implement the uniprot list with this added value, we matched the information from UniProt and Ensembl building a new table.

Using Microsoft Access we generated a new table through a sql query joining the Ensembl data (step 2.3) and UniProt data (step 1) based on the fields "AC UniProt" and "length". The new table was called CHR+chromosome number (two digits) canonical.xls.

Finally, we ordered the proteins by "gene_start", which is the position on the chromosome.

S2. Nomenclature

All files were named using the following rules:
Original UniProt table (step 1): CHR+ chromosome number (two digits)+_uniprot.xls; example CHR01_uniprot.xls

Original Biomart table (step 2.1): CHR+chromosome number (two digits)_Biomart.txt; example CHR01_Biomart.txt FASTA format

Perl Tool file name (step 2,2): CHR+chromosome number (two digits)_fasta_elab.pl

Perl Tool output file (step 2.2): CHR+ chromosome number (two digits)_Biomart_TAB.fa; example CHR01_Biomart_TAB.fa

Excel file from perl Tool output file (step 2.2): CHR+chromosome number (two digits)_ _gene_Biomart_excel.xls; example CHR01_gene_Biomart_excel.xls

Excel file from perl Tool output file (step 2.3): CHR+chromosome number (two digits)_ _gene_Biomart_excel_elab.xls; example CHR01_gene_Biomart_excel_elab.xls

Canonical table (step 3) : CHR+chromosome number (two digits)_canonical.xls; example CHR01_canonical.xls