Data sources and datasets

Table 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Purpose | Access date | Filtering | source |
| Uniprot | Human Protein sequences + protein annotation | Feb 2025 | Reviewed sequences + proteins with valid Uniprot ids | https://www.uniprot.org/ |
| gnomAD v2.1.1 leftover to GRCh 38 | SNPs | 2025 | SNPs in exon regions; coding SNPs; AF >0.1; Filter = PASS; population = EUR\* | https://gnomad.broadinstitute.org/ |
| Ensembl v110 | VEP indexed cache | 2025 | https://ftp.ensembl.org/pub/release-110/variation/indexed\_vep\_cache/homo\_sapiens\_vep\_110\_GRCh38.tar.gz | |
| HPA v24 | HPA RNA dataset | 2025 | Valid cross-reference with Uniprot | https://www.proteinatlas.org/download/rna\_tissue\_hpa.tsv.zip |
| GTEx v8 | Transcript level gxp | 2025 | https://storage.googleapis.com/adult-gtex/bulk-gex/v8/rna-seq/GTEx\_Analysis\_2017-06-05\_v8\_RSEMv1.3.0\_transcript\_tpm.gct.gz |
| IEDB | Peptide annotation | 2025 | Human; linear peptide; peptide sequence | https://www.iedb.org/ |
| PEPREP | Peptide annotation | 2025 | Human; peptide sequence | BioCopy |

Table 2

|  |  |  |
| --- | --- | --- |
| Tools/package | Purpose | Reference |
| BLASTp | Sequence search |  |
| PEPMatch | Sequence search |  |
| tabix 1.9 | SNPs extraction |  |
| Bcftools |  |
| vep | SNPs annotation |  |
| NetMHCpan | HLA binding predictions |  |
| MHCflurry |  |
| R packages | |  |
| ggplot2 | Visualization | H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. |
| lessR | Visualization | Gerbing D (2025). lessR: Less Code, More Results. R package version 4.4.5, <https://CRAN.R-project.org/package=lessR>.  Gerbing DW (2021). “Enhancement of the Command-Line Environment for use in the Introductory Statistics Course and Beyond.”  Journal of Statistics and Data Science Education, 29(3), 251-256. doi:10.1080/26939169.2021.1999871  <https://doi.org/10.1080/26939169.2021.1999871>. |
| biomaRt | Cross-reference and gene annotation | Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Steffen Durinck, Paul  T. Spellman, Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).  BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Steffen Durinck, Yves  Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma and Wolfgang Huber, Bioinformatics 21, 3439-3440 (2005). |
| rmarkdown | Report generation | Allaire J, Xie Y, Dervieux C, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2024).  rmarkdown: Dynamic Documents for R. R package version 2.29, <https://github.com/rstudio/rmarkdown>. |
| tidyr | Data manipulation | Wickham H, Vaughan D, Girlich M (2024). tidyr: Tidy Messy Data. R package version 1.3.1,  <https://CRAN.R-project.org/package=tidyr>. |
| Python module | |  |
| Panda |  |  |
| sklearn |  |  |
| Numpy |  |  |
| Bio.SeqUtils |  |  |
| Scipy |  |  |
| peptides |  |  |
| CrossDome | An R package to measure cross-reactivity risk in the sequence-space | Fonseca A (2025). crossdome: An R package to measure cross-reactivity risk in the sequence-space. R package version <https://github.com/antuneslab/crossdome>. |
|  | |  |
| R 4.4.1 | R Core Team (2024). \_R: A Language and Environment for Statistical Computing\_. R  Foundation for Statistical Computing, Vienna, Austria.  <https://www.R-project.org/>. | |
| Python 3.9.16 | Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. | |
| Conda |  | |
| snakemake 7.24.2 | [Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., 2021. Sustainable data analysis with Snakemake. F1000Res 10, 33.](https://doi.org/10.12688/f1000research.29032.1) | |

\*: we followed gnomAD population’s description for selection (see https://gnomad.broadinstitute.org/news/2017-02-the-genome-aggregation-database/)