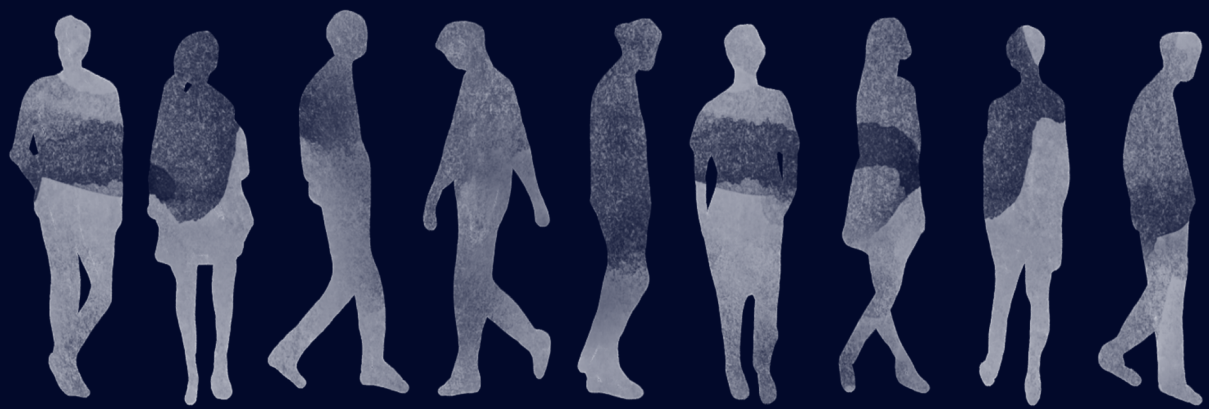
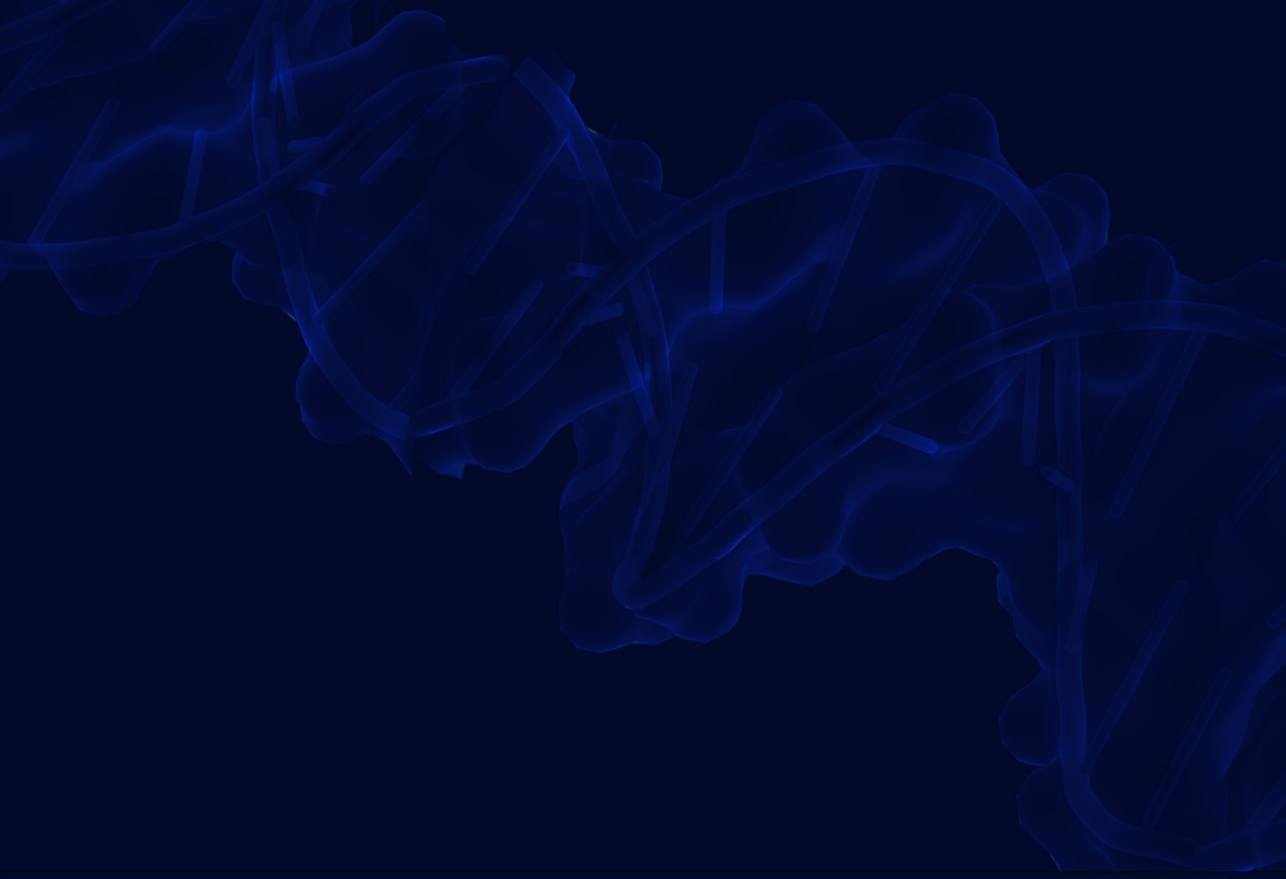
Getting personal: advancing personalized oncology through computational analysis of membrane proteins

Gorostiola González, M.

Appendix A

GDC SQL implementation v22.0
(release 16th January 2020)
Quick start guide





What is the GDC?

The NCI Genome Data Commons (GDC)[1] is a publicly available cancer knowledge network to provide the cancer research community with a harmonized and curated data service for genome/transcriptome sequence data and standardized analyses for derived data from different cancer studies. The GDC is currently the official repository of data from the TCGA project[2] and other more recent and ongoing whole genome cancer sequencing projects such as the TARGET and CGCI projects[3,4].

GDC conventional data channels versus GDC SQL local implementation

The data provided by the GDC is available through three different channels: a data portal, a data transfer tool, and an API. From the data portal, part of the data (e.g. cases, genes, mutations, clinical data) is accessible for visualization and analysis, with a number of tools available for this purpose (i.e. Oncogrid, survival plots, cohort comparison). The data portal also allows exploration of the repository, where the data is stored in different file formats. The data comprised in those files, however, is only accessible upon download, which can be done through the data portal (for a small number of files), or through the data transfer tool (from the command line, for a larger amount of files), providing a data manifest previously generated in the data portal. Furthermore, there is an API that grants access to all of the data available on the data portal through different endpoints, as well as to the generation of data manifests for data downloads. Although the conventional GDC data channels are extremely useful for data visualization and retrieval of specific - limited - queries, it is not the most appropriate tool for big data analysis, since the links between different data types are sometimes unclear, and not all the data types are available from the same channels. Moreover, the GDC repository is updated every 2-3 months with new entries, and the conventional data channels only allow data retrieval from the most recent release, which can be prejudicial for projects running for a longer time. Due to these factors, I made the decision to develop a local SQL implementation for GDC using data acquired from all conventional sources. This implementation aims to streamline access to all data from a specific release. The data model that I formulated was carefully crafted to facilitate large-scale data queries and incorporates relevant data types essential for cancer research in both my thesis and related collaborations. Several tests showed that the data contained in the SQL local implementation is almost the same as that of the data portal (for the data types available in the data portal), although some minor differences are found due to errors in the database. The conventional channels, however, are still very useful tools for data visualization and analysis, but the release version needs to always be taken into account.

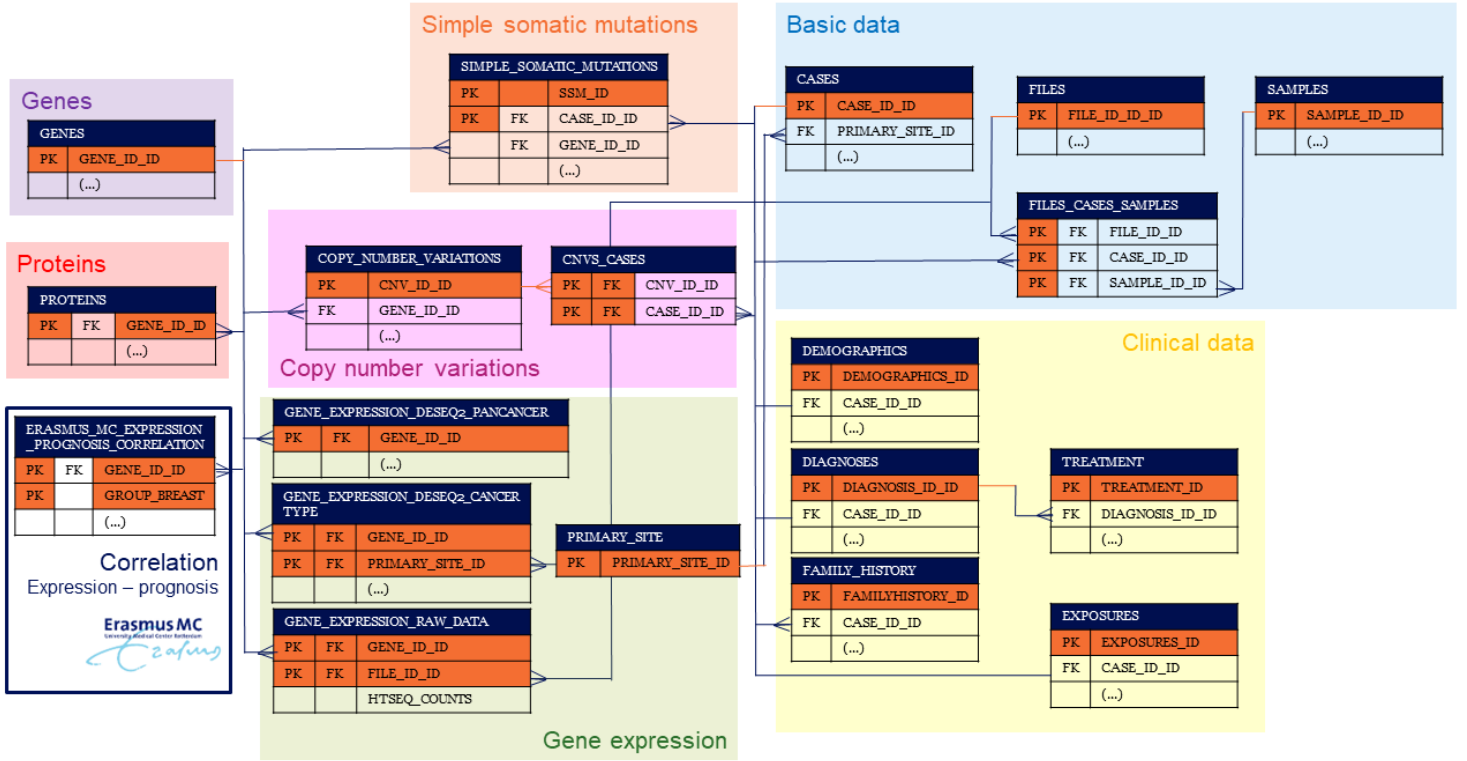


Figure A.1. The basic architecture of the GDC SQL local implementation. Only primary and foreign keys are depicted in the diagram, as well as the connections between them.

The GDC SQL local implementation's basic structure

The SQL local implementation features 19 tables organized into eight fields connected by a network of primary (PK) and foreign (FK) keys to optimize storage and query processing, as shown in **Figure A.1**. Unique numerical values are used for all PKs, and FKs reference PKs in parent tables. There is only one exception to this rule, explained in more detail in the section *The connection between cases, samples, and files*. Some tables (*files_cases_samples, cnvs_cases,* and *primary_site*) serve mainly as connection tables and lack additional properties. The full database schema, including all properties, can be found in the associated online repository for **Chapter 5**[5].

Description of the data fields and their source

The seven data fields in **Figure A.1** depict diverse data types gathered from GDC conventional data channels.

a) Basic data

   The tables in this field contain basic data properties for cases, samples, and files. "Cases" is the term used in GDC for patients. The connection between cases, samples, and files is crucial for analyzing the data in the database. The data was obtained through API queries to cases and files endpoints. More detailed relationships and their implications are discussed in *The connection between cases, samples, and files* section.

b) Clinical data

   Some of the patients in the GDC have associated clinical data, depending on the cancer project. For those cases, five different tables are available (demographics, family history, exposures, diagnoses, and treatments). The data contained in these tables was obtained by querying the API's cases endpoint.

c) Simple somatic mutations

   This table contains all the data associated with genomic sequencing. The data was obtained by querying the API's ssms endpoint and filtered as in the data portal to only keep the canonical transcript's data when several transcripts were available.

d) Copy number variations

   Most data is available in the data portal, but copy number variation data can be found only through the API's cnvs endpoint.

e) Gene expression

   This field includes raw transcriptomic data (RNA seq HTSeq counts) and analyzed gene expression annotations, making it one of the most challenging fields due to the lack of availability in the data portal and API. I obtained and analyzed the files using

a specific pipeline detailed in the *Analysis of gene expression data* section.

f) Genes

This is a field that provides gene information for different tables. The data for this field was obtained from the API's ssms endpoint, and extracted from the Simple somatic mutations table to be able to provide information to a larger number of tables.

g) Proteins

Similarly to the field *Genes*, this field provides protein information for different tables.

h) Erasmus MC expression-progression correlation

This field provides information derived from an Erasmus MC (CC. J.W.M. Martens) correlation analysis between breast cancer patient's gene expression data and their cancer progression profiles.

Guide to the most useful data properties

All properties obtained from the GDC API conserve their original names, except for the PKs and FKs, which were manually created to give numerical references. Therefore, the description for some of the properties is available in the GDC data dictionary online. In general, the description of the properties is intuitive. The properties in the tables corresponding to the *Gene expression* field are derived from differential expression analysis and are further detailed in the section *Analysis of gene expression data*. The tables *files_cases_samples, cnvs_cases,* and *primary_site* are mainly connection tables, therefore they do not contain useful properties for other purposes than linking tables. The most useful properties in the most relevant tables are described in **Tables A.1-A.11** with definitions based on the GDC data dictionary (https://docs.gdc.cancer.gov/Data_Dictionary/viewer).

Table A.1. Description of the most useful properties in table cases.

Table: cases	
Property	**Meaning**
primary_site	Primary site or the general location of the cancer, as categorized by the World Health Organization (WHO). Can be used as a replacement for cancer type. E.g. Adrenal gland, Breast, Bronchus and lung.
disease_type	Type of malignant disease as categorized by the World Health Organization's (WHO) International Classification of Diseases for Oncology (ICD-O). E.g. Blood Vessel Tumors, Mesothelial Neoplasms.

Table A.2. Description of the most useful properties in table samples.

Table: samples	
Property	**Meaning**
sample_type	Origin of a biological sample utilized in a laboratory analysis. E.g. Blood Derived Cancer - Bone Marrow, Primary Tumor, Metastatic, Solid Tissue Normal, RNA, Slides
tissue_type	Type of tissue based on its disease status or proximity to tumor tissue. E.g. Tumor, Normal, Peritumoral *NOTE: even though this would be the perfect property to define whether a sample is derived from a tumor or normal tissue, it is often unknown or not described, so for that purpose is better to use sample_type

Table A.3. Description of the most useful properties in table files.

Table: files	
Property	**Meaning**
data_category	Category of data included in a file. E.g. Simple Nucleotide Variation, Clinical
data_type	Detailed data type included in a file. E.g. Gene Expression Quantification, Slide Image
experimental_strategy	Experimental strategies employed for molecular characterization of the cancer. E.g. WGS, miRNA-Seq
workflow_type	Bioinformatic workflow used for analysis of the data. E.g. DNAcopy, HTSeq - Counts, GENIE Simple Somatic Mutation

Table A.4. Description of the most useful properties in table diagnoses.

Table: diagnoses	
Property	**Meaning**
age_at_diagnosis	Age at the time of diagnosis as number of days since birth.
last_known_disease_status	Last known condition of an individual's disease. E.g. Tumor free, Distant met recurrence/progression
progression_or_recurrence	Yes/No/Unknown indicator to identify whether a patient has had a new tumor after initial treatment.
ajcc_clinical_stage	Stage group determined from clinical information on the tumor, regional node, and metastases to group patients with similar prognosis for cancer. E.g. Stage 0, Stage IA2
ajcc_pathologic_stage	Spread of the disease through the body based on cancer staging using AJCC criteria.
igcccg_stage	Staging according to the International Germ Cell Cancer Collaborative Group (IGCCCG), used to further classify metastatic testicular tumors. E.g. Good Prognosis, Poor Prognosis

Table A.5. Description of the most useful properties in table treatments.

Table: treatments	
Property	**Meaning**
therapeutic_agents	Individual agent(s) used in treatment. E.g. 10-Deacetyltaxol
treatment_effect	Effect a treatment had on the tumor. E.g. complete Necrosis (No Viable Tumor), No Necrosis
treatment_type	Type of treatment used. E.g. Chemotherapy, Immunotherapy (Including Vaccines)

Table A.6. Description of the most useful properties in table simple_somatic_mutations.

Table: simple_somatic_mutations	
Property	**Meaning**
mutation_type	General type of mutation. E.g. Substitution, Deletion, Insertion
mutation_subtype	Detailed subtype of mutation. E.g. Missense, Frameshift, Stop Gained, Intron
gene_id	Ensembl gene id
aa_change	Amino acid change in the protein affected by a mutation in a protein-coding gene. E.g. V600E, K15Rfs*5, empty (for deletions)
sift_impact	Sorting Intolerant From Tolerant (SIFT[6]) predicted category respect to the likelihood of a phenotypic effect upon mutation: • tolerated: Not likely • tolerated_low_confidence: More likely than tolerated • deleterious: Likely • deleterious_low_confidence: Less likely than deleterious
vep_impact	Ensembl Variant Effect Predictor (VEP[7]) predicted category respect to the extent of the impact on protein function upon mutation : • HIGH (H): Disruptive impact on the protein, e.g. truncation, loss of function • MODERATE (M): Non-disruptive but might change protein effectiveness • LOW (L): Mostly harmless • MODIFIER (MO): Non-coding variants or variants affecting non-coding genes, therefore the impact is difficult to predict
polyphen_impact	Polymorphism Phenotyping (Polyphen[8]) predicted category respect to the possibility to affect protein structure or function: • probably damaging (PR): Highly possible • possibly damaging (PO): Possible • benign (BE): Not likely • unknown (UN): Difficult to make a prediction

Table A.7. Description of the most useful properties in table copy_number_variations.

Table: copy_number_variations	
Property	**Meaning**
gene_id	Ensembl gene id
cnv_change	Copy number estimation based on the GDC *Copy Number Variation Analysis Pipeline*. Three categories are defined based on the focal CNV values: • loss (-1): focal CNV values smaller than -0.3 • gain (+1): focal CNV values larger than 0.3 • neutral (0): focal CNV values between -0.3 and 0.3

Table A.8. Description of the most useful properties in table genes.

Table: genes	
Property	**Meaning**
gene_id	Ensembl gene id
symbol	HGNC symbol for the gene analyzed

Table A.9. Description of the most useful properties in tables gene_expression_deseq2_pancancer and gene_expression_deseq2_cancertype.

Table: gene_expression_deseq2_pancancer / gene_expression_deseq2_cancertype	
Property	**Meaning**
gene_id	Ensembl gene id
expression_status	Gene expression estimation upon differential expression analysis of tumor vs. normal samples with DESeq2 as detailed in section *Analysis of gene expression data*. • Genes with log2_fold_change values larger than 2 and p_value values lower than 0.05 are categorized as "overexpressed" • Genes with log2_fold_change values lower than -2 and p_value values lower than 0.05 are categorized as "underexpressed" • Genes with log2_fold_change values between -2 and 2 and p_value values lower than 0.05 are categorized as "neutral" • Genes with p_value values higher than 0.05 are categorized as "not significant"

Table A.10. Description of the most useful properties in table proteins.

Table: proteins	
Property	**Meaning**
gene_id	Ensembl gene id
SwissProt	UniProt accession code for the protein corresponding to the gene analyzed

Table A.11. Description of the most useful properties in erasmus_mc_expression_prognosis_correlation table.

Table: erasmus_mc_expression_prognosis_correlation	
Property	**Meaning**
gene_id	Ensembl gene id
group_breast	Group to which the breast cancer patients belong to: • ERpos: ER positive • ERneg: ER negative • TN: triple negative (negative for ER, PR and ERBB2) • ALL: all patients
expression_prog_corr	The existence or not of correlation between gene expression and negative prognosis (rapid progression leading to metastasis): • "True" for hr_value > 1 and p_value < 0.05 • "False" for hr_value < 1 and pvalue < 0.05 • "ns" or not statistically significant for p_value > 0.05 and any hr_value

The connection between cases, samples, and files

The basic information field is crucial to understand the patient's data. All the rest of the fields are connected to this one, either by the *case_id_id* or the *file_id_id*. The cases table provides important information, like the primary site of the tumor, while the samples table provides information about the type of sample (e.g. normal, tumor). Even though the data directly connected to *case_id_id* (e.g. simple somatic mutations or clinical data) cannot be directly connected to a specific sample, the information provided by that link can be still very useful.

It is important to note, however, that the relationship between cases, files, and samples is complicated. A single case can be associated to different files and to different samples. At the same time, a file can be created with data from different samples, even from different cases. This can be confusing when the same case has samples of different types (**Figure A.2**). All these possible scenarios need to be considered when querying the database.

Moreover, these relationships can be retrieved from the GDC API through two different ways: a) using the cases endpoint, the cases - samples, and the cases - files relationships can be retrieved, and b) using the files endpoint, the files - cases - samples triple relationship can be retrieved. Here, I used both ways to obtain data for the *files_cases_samples* table. This means that some file-sample relationships are not defined. In order to be able to add the additional cases - samples and cases -files relationships, they were linked to an empty value in the files and samples tables, respectively.

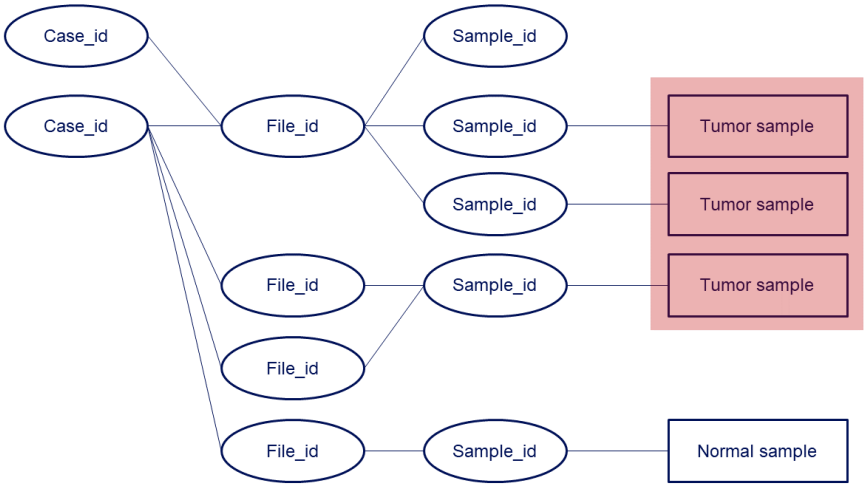


Figure A.2. Relationships between cases, files, and samples available in the GDC.

Analysis of gene expression data

Even though transcriptomic data is available in files from the GDC data transfer tool, these need further analysis in order to be properly interpreted. From files with *workflow_type* equal to "HTSeq - Counts", I performed differential expression analyses with DESeq2 in order to assess the over- and under-expression of genes in tumor vs. normal samples. I made use of the files-cases-samples relationships in order to define the origin of the different RNA sequencing files. I performed two types of analyses: a) pan-cancer differential expression analysis and b) per cancer type differential expression analysis. The cancer type was defined based on the *primary_site* property. The potential batch effect introduced by samples from different projects was accounted for by introducing it as the covariate in the analysis. The DESeq2 analysis was performed using the Leiden University supercomputer facilities (ALICE), and the results were uploaded to the GDC SQL local implementation, together with an interpretation of the results (property *expression_status*). Moreover, in the GDC SQL local implementation, there is a *gene_expression_raw_data* table, where the raw counts from the HTSeq - Counts files are included, in order to be able to perform differential expression analyses a posteriori from raw data on custom cohorts.

Erasmus MC prognosis analysis

The data from Erasmus MC was provided by J.W.M. Martens for breast tumors and breast cell lines. Regarding the tumors, this is a cohort of their own data (n = 344) supplemented with publicly available samples that all run on the same chip type (867 samples in total). Clinically, the samples are similar as well, all are lymph-node negative and have not been adjuvantly treated (no chemo / hormonal therapy after surgery to

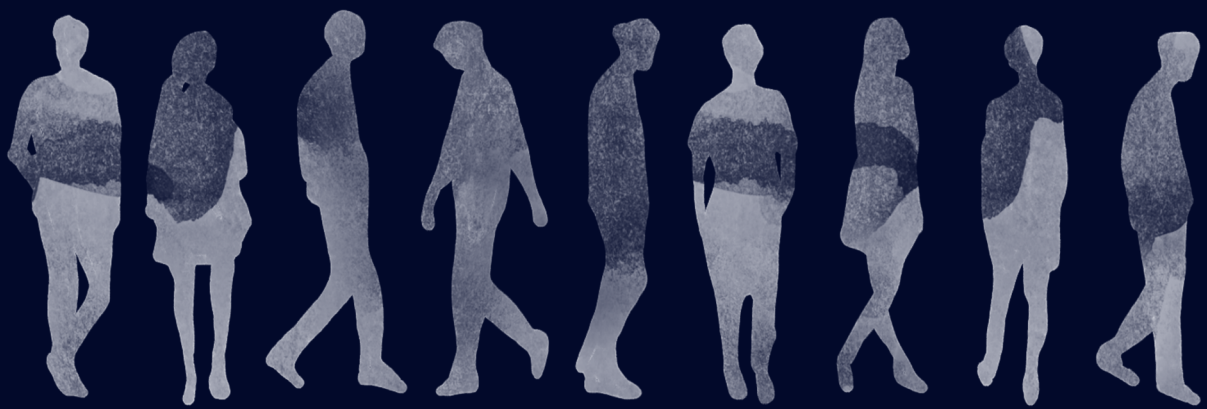
remove the primary tumor). They also know the metastasis-free survival (MFS) of these patients, and they then view the prognosis in all samples or separately for ER negatives, ER positives, and Triple negatives (negative for ER, PR, and ERBB2). With a Cox regression, they calculated a Hazard Ratio ("hr_value") with a p-value. This was done on the expression data as a continuous value. A HR> 1 means a correlation exists between high expression and poor prognosis (short time between primary and metastasis).

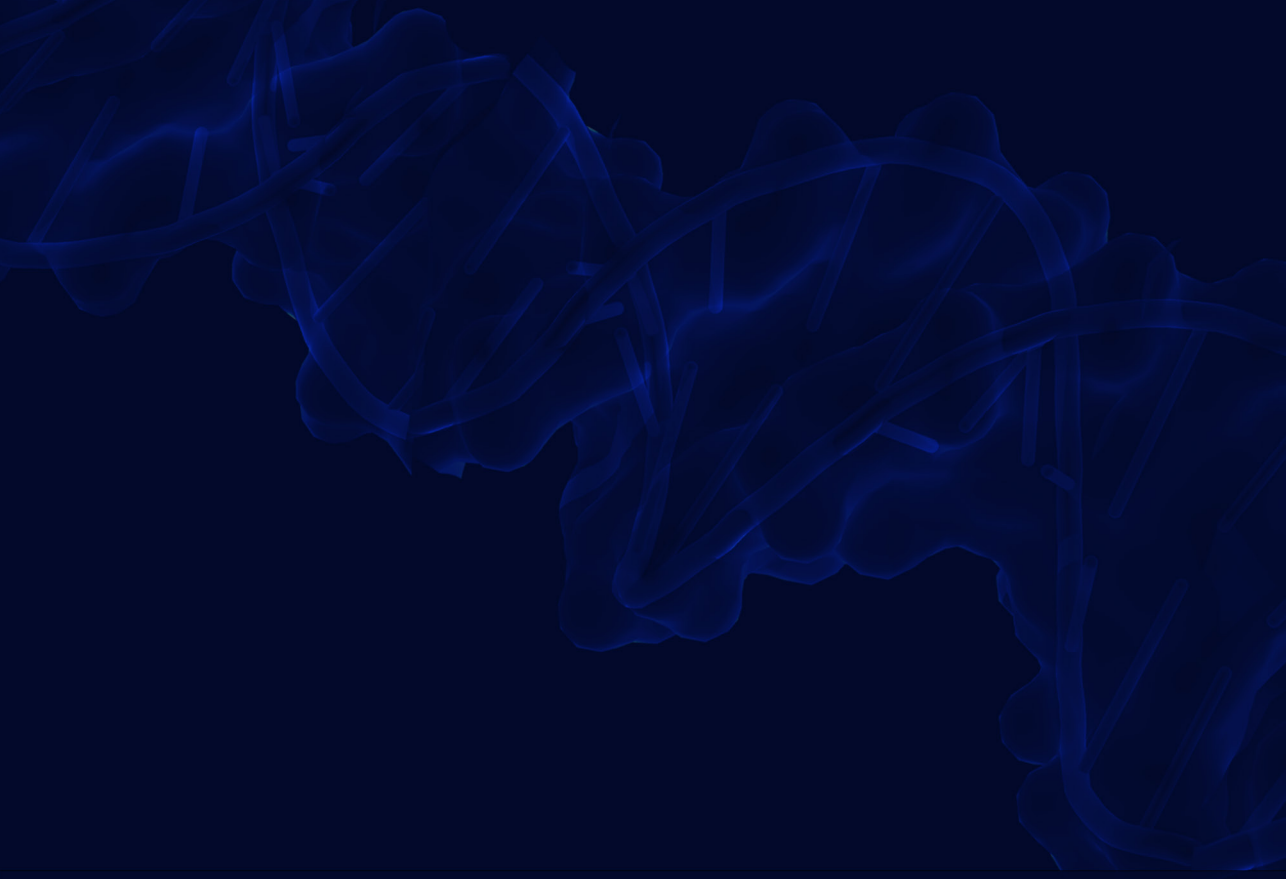
References

1. Grossman, R. L. *et al.* Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
2. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* **19**, A68 (2015).
3. Ma, X. *et al.* Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).
4. Thomas, N. *et al.* Genetic subgroups inform on pathobiology in adult and pediatric Burkitt lymphoma. *Blood* **141**, 904–916 (2023).
5. Bongers, B. *et al.* Data underlying the article: Pan-cancer in silico analysis of somatic mutations in G-protein coupled receptors: The effect of evolutionary conservation and natural variance. Available at https://doi.org/10.4121/15022410.V1 (2021).
6. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
7. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 1–14 (2016).
8. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

Appendix B

Data and software availability





This thesis was created with the aim to promote FAIR (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable) data management principles and open source software development practices. To this end, whenever possible, data was obtained from public databases and repositories, and open-source software was used. Similarly, novel datasets and code derived from the practical chapters of this thesis (Chapters 4-8) were made available through public repositories:

Chapter 4

Data availability

The ChEMBL 31 data used in this chapter is available online (https://doi.org/10.6019/CHEMBL.database.31). The bioactivity data from the Papyrus dataset is available online (https://doi.org/10.5281/zenodo.7373214). All protein structures used in this chapter are available on the RCSB Protein Data Bank (https://www.rcsb.org/). The bioactivity-enhanced bioactivity dataset derived from this chapter is available on Zenodo (https://doi.org/10.5281/zenodo.11236694).

Software availability

The Python 3.10 code used to compile and analyze the data for this chapter is available on Zenodo (https://doi.org/10.5281/zenodo.11236694) and maintained on GitHub (https://github.com/CDDLeiden/chembl_variants).

Chapter 5

Data availability

The protein structures used in this chapter are available on the RCSB Protein Data Bank (https://www.rcsb.org/). The ChEMBL 27 data used in this chapter is available online (https://doi.org/10.6019/CHEMBL.database.27). The G protein-coupled receptor information derived from the GPCRdb database is available online (https://gpcrdb.org/). The GDC v22.0 SQL implementation and the compilation of the 1000 Genomes dataset, as well as all datasets for analysis derived from this chapter are available on the 4TU repository (https://doi.org/10.4121/15022410).

Software availability

The source code used to produce the results in this chapter was generated using the commercial software package Accelrys Pipeline Pilot 2018 version 18. All Pipeline Pilot protocols, as well as the Python 3.8 code used to generate the figures for this chapter, are available on the 4TU repository (https://doi.org/10.4121/15022410).

Chapter 6

Data availability

The GDC v22.0 SQL implementation and the compilation of the 1000 Genomes dataset are available in online repositories (see Chapter 5 *Data availability*). All protein structures used in this chapter are available on the RCSB Protein Data Bank (https://www.rcsb.org/). The input files needed to generate the molecular dynamics simulations in this chapter using Desmond, as well as the results from Monte Carlo mutagenesis and 4D docking are available on Zenodo (https://doi.org/10.5281/zenodo.11236571).

Chapter 6

Software availability

The commercial software ICM-Pro version 3.9-2c and open source Desmond version 2021.1 were used in this chapter. The analysis of the molecular dynamics simulations was done with PyMol version 2.5.2 and Python 3.8. All the projects and scripts are available on Zenodo (https://doi.org/10.5281/zenodo.11236571).

Chapter 7

Data availability

The bioactivity data used in this chapter was obtained from the Papyrus dataset and is available online (https://doi.org/10.5281/zenodo.7373214). The wild-type molecular simulations were obtained from the GPCRmd database and are available online (https://submission.gpcrmd.org/home/). The G protein-coupled receptor information derived from the GPCRdb database is available online (https://gpcrdb.org/). The input files needed to generate the mutant molecular dynamics simulations in this chapter using AceMD are available on Zenodo (https://doi.org/10.5281/zenodo.7957235).

Software availability

The Python 3.8 code used to generate and analyze the results in this chapter is available on Zenodo (https://doi.org/ 10.5281/zenodo.8026883) and maintained at GitHub (https://github.com/CDDLeiden/3ddpd).

Chapter 8

Data availability

The data used in this chapter was previously compiled and made available in previous chapters, expect for the phosphorylation network, which is freely available online upon registration (https://cancer.ucsf.edu/phosphoatlas), and the kinome data from the KLIFS database, which is available online (https://klifs.net/browse.php). The input files needed to generate the kinome and receptor tyrosine kinase knowledge graphs, as well the pickle files to re-generate the graphs are available on Zenodo (https://doi.org/10.5281/zenodo.11236776). This repository also contains the HTML interactive visualization sessions described in the chapter.

Software availability

The Python 3.10 code used to compile and analyze the knowledge graphs is available on Zenodo (https://doi.org/10.5281/zenodo.11236776).