

Gene Expression - Drug Relationship Datasets

Martina Betti
Stefania Sferragatta

Human Protein Atlas

[link to RNA expression level table](#)

THE HUMAN PROTEIN ATLAS

MENU HELP NEWS

TECHNICAL DATA DOWNLOADABLE DATA

cell), transcripts per million ("TPM") and protein-coding transcripts per million ("pTPM"). The data was obtained from Monaca publication and is based on The Human Protein Atlas version 20.1 and Ensembl version 92.38.

20 RNA Schindler blood cell gene data

Transcript expression levels summarized per gene in 15 blood cell types. The tab-separated file includes Ensembl gene identifier ("Gene"), analysed sample ("Blood cell") and transcripts per million ("TPM"). The data was obtained from Schindler publication and is based on The Human Protein Atlas version 20.1 and Ensembl version 92.38.

21 RNA HPA cell line gene data

Transcript expression levels summarized per gene in 69 cell lines. The tab-separated file includes Ensembl gene identifier ("Gene"), analysed sample ("Cell line"), transcripts per million ("TPM"), protein-coding transcripts per million ("pTPM") and normalized expression ("NK"). The data is based on The Human Protein Atlas version 20.1 and Ensembl version 92.38.

22 RNA TCGA cancer sample gene data

Transcript expression levels summarized per gene in 7932 samples from 17 different cancer types. The tab-separated file includes Ensembl gene identifier ("Gene"), analysed sample ("Sample"), cancer type ("Cancer") and fragments per kilobase million ("FPKM"). The data is based on The Human Protein Atlas version 20.1 and Ensembl version 92.38.

ma_blood_cell_schindler.tsv.zip
TSV-file (zip compressed), 1.5 MB

ma_cellline.tsv.zip
TSV-file (zip compressed), 10.2 MB

ma_cancer_sample.tsv.zip
TSV-file (zip compressed), 1.1 GB

HGNC

[link to converter](#)

HGNC

Gene data Tools Downloads VGNC Contact us More

Search symbols, keywords or IDs

Request symbol

Custom downloads

Select column data

Curated by the HGNC

☐ HGNC ID ☒ Approved symbol ☐ Approved name ☐ Status

☐ Locus type ☐ Locus group ☐ Previous symbols ☐ Previous name

☐ Alias symbols ☐ Alias names ☐ Chromosome ☐ Date approved

☐ Date modified ☐ Date symbol changed ☐ Date name changed ☐ Accession numbers

☐ Enzyme IDs ☐ NCBI Gene ID ☒ Ensembl gene ID ☐ Mouse genome database ID

☐ Specialist database links ☐ Specialist database IDs ☐ Pubmed IDs ☐ RefSeq IDs

☐ Gene group ID ☐ Gene group name ☐ CCDS IDs ☐ Vega IDs

☐ Locus specific databases

Select all

OncoKB

[link to actionability](#)
[link to oncogenes](#)

OncoKB

Levels of Evidence Actionable Genes Cancer Types API Access About Team News Terms FAQ

Level 1 FDA approved drugs 43 Genes Level 2 Standard case 13 Genes Level 3 Clinical evidence 20 Genes Level 4 Biological evidence 22 Genes Level R1 Standard case 8 Genes Level R2 Clinical evidence 5 Genes

Diagnoses Levels for hematologic malignancies only

Prognosis Levels for hematologic malignancies only

43 actionable genes Select a cancer type All drugs

Showing 134 clinical implications 143 genes, 27 cancer types, 1 level of evidence

Level	Gene	Alterations	Cancer Types	Drugs
1	ABL1	BCR-ABL1 Fusion	B-lymphoblastic Leukemia/Lymphoma	Dasatinib
1	ABL1	BCR-ABL1 Fusion	B-lymphoblastic Leukemia/Lymphoma	Imatinib
1	ABL1	BCR-ABL1 Fusion	B-lymphoblastic Leukemia/Lymphoma	Ponatinib
1	ABL1	BCR-ABL1 Fusion	Chronic Myelogenous Leukemia	Dasatinib
1	ABL1	BCR-ABL1 Fusion	Chronic Myelogenous Leukemia	Imatinib
1	ABL1	BCR-ABL1 Fusion	Chronic Myelogenous Leukemia	Nilotinib
1	ABL1	BCR-ABL1 Fusion	Chronic Myelogenous Leukemia	Ponatinib
1	ABL1	T3191	B-lymphoblastic Leukemia/Lymphoma	Ponatinib
1	ABL1	T3191	Chronic Myelogenous Leukemia	Ponatinib
1	ALK	Fusions	Anaplastic Large-Cell Lymphoma ALK Positive	Crizotinib
1	ALK	Fusions	Non-Small Cell Lung Cancer	Alectinib
1	ALK	Fusions	Non-Small Cell Lung Cancer	Crizotinib
1	ALK	Fusions	Non-Small Cell Lung Cancer	Crizotinib
1	ALK	Oncogenic Mutations	Non-Small Cell Lung Cancer	Brigatinib
1	ALK	Oncogenic Mutations	Non-Small Cell Lung Cancer	Lorlatinib

Gene Expression - Drug Relationship

Rationale

Martina Betti
Stefania Sferragatta

SQL

- **Continuous values for Gene expression.**
FPKM column specify level of expression from which thresholds are calculated specifically for each gene.
- **Alteration type is treated as any other column in the table.**
From the data structure one cannot infer that the alteration type is defines the relationship between gene and drug.

NEO4J

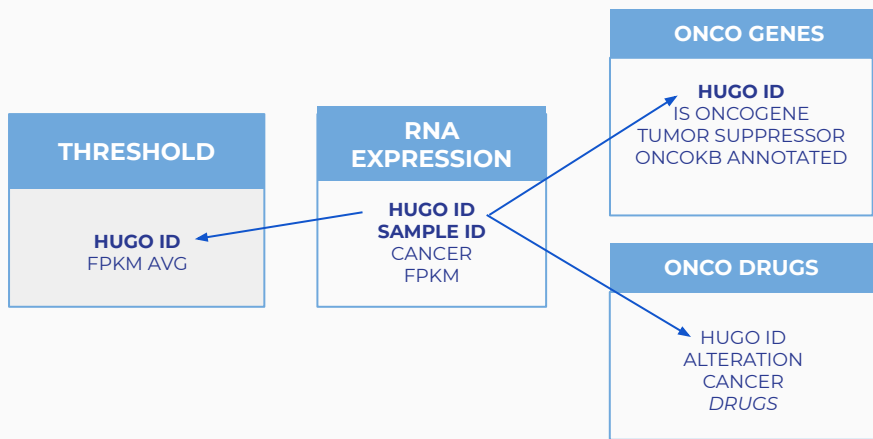
- **Binary information on Gene expression.**
No quantification is provided: a relation is established between a Gene ID and a Sample ID if a given gene is overexpressed in a given sample.
- **Alteration type is described by a relation property.**
The type of alteration for a given gene and a given drug is neither an attribute of the first or the second, rather it is what the determines their relationship.

Gene Expression - Drug Relationship

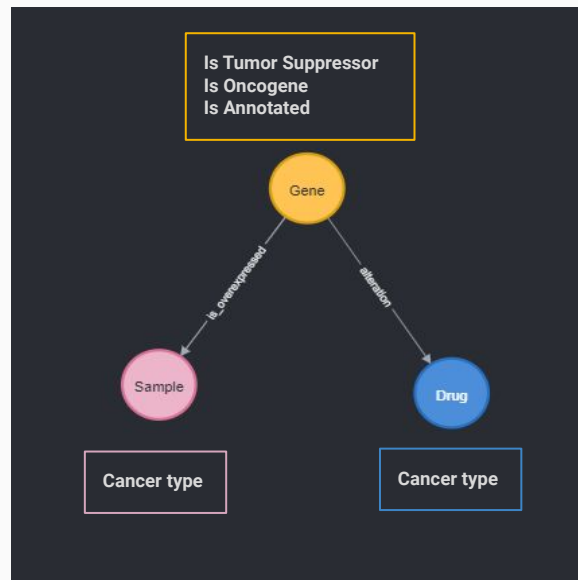
Schema vs. Graph

Martina Betti
Stefania Sferragatta

SQL



NEO4J



Gene Expression - Drug Relationship

Query Representation

Martina Betti
Stefania Sferragatta

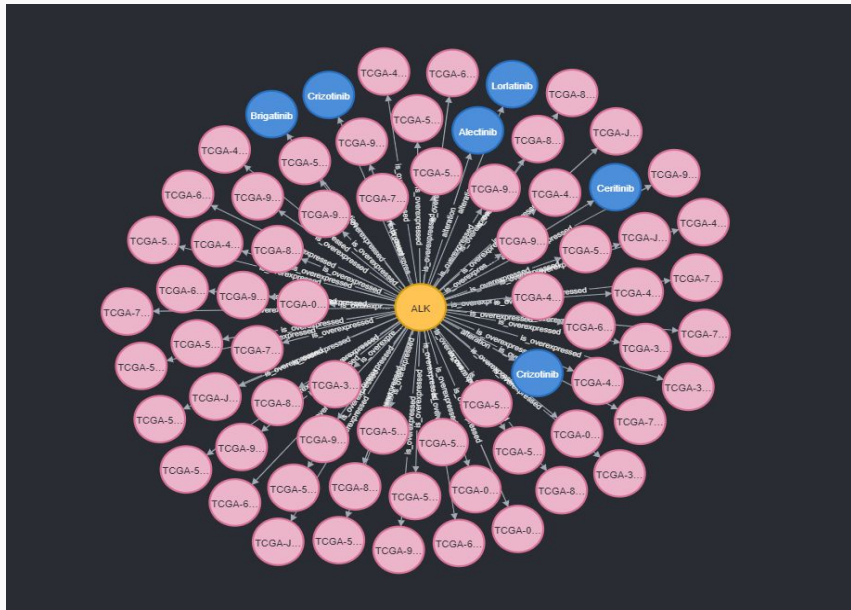
Example of ALK gene graph

Yellow nodes: Genes

Pink nodes: Samples ID

Blue nodes: Drugs

Each edge between the Gene and the Sample ID specifies the relationship “Is Overexpressed”, while each relationship between Gene and Drug specifies the alteration type.



Gene Expression - Drug Relationship

Conclusions

Martina Betti
Stefania Sferragatta

	SQL	NEO4J
PRO	<ul style="list-style-type: none">- Suitable for storing quantitative data and measurements	<ul style="list-style-type: none">- Better describes relationship types- Fast and flexible data loading- Allows for visualization
CONS	<ul style="list-style-type: none">- Time consuming, especially on data loading.- Demanding requirements on input data format.- Obscure relationship types.	<ul style="list-style-type: none">- Hard to use on quantified measurements.