Etapa 03 Projeto Banco de Dados

Tema: Amostras de RNA x Patologias

Pedro Barros Bastos RA:204481
Gabriel Volpato Giliotti RA:197569

Filtragem de dados

Base: https://www.proteinatlas.org/ENSG00000134057.xml

The Human Protein Atlas is a Swedish-based program initiated in 2003 with the aim to map all the human proteins in cells, tissues and organs using integration of various omics technologies, including antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology.

- Conversão dos dados da base em XML para formato .TSV para manipulação utilizando o link: https://xmlconverter.sonra.io/signup
- Conversão do .TSV para .CSV para criação de esquemas SQL no Jupyter utilizando o link: https://onlinetsvtools.com/convert-tsv-to-csv
- Melhor entendimento da base dado o modelo Entidade-Relacionamento gerado pela conversão

Problema: (Focado em análise exploratória)

- **Quais patologias possuem amostras de RNA de tecidos afetados dadas por pessoas com mais de X anos?**


RNASample
- FK_data
- sampleId
- sex
- unitRNA
- expRNA
- age
proteinAtlas_entry_rnaExpression_data
- FK_rnaExpression
- PK_proteinAtlas_entry_rnaExpression_data
- bloodCell
- bloodCell_lineage
- cellLine
...
rnaExpression
- FK_proteinAtlas
- PK_rnaExpression
- rnaDistribution
- rnaDistribution_description
- rnaSpecificity_description
- rnaSpecificity_specificity
- rnaSpecificity_tissue
- rnaSpecificity_tissue_ontologyTerms
...
proteinAtlas_entry_pathologyExpression_data
- survivalAnalysis_dataSource
- survivalAnalysis_isPrognostic
- survivalAnalysis_prognosticType
- survivalAnalysis_pValue
- survivalAnalysis_source
- tissue (a patologia)
- tissue_organ (orgao relacionado)
- FK_proteinAtlas
proteinAtlas
- PK_proteinAtlas
- entry_cellExpression_image_imageUrl
- entry_cellExpression_source
- entry_cellExpression_summary
- entry_cellExpression_technology
- entry_cellExpression_verification
...

SQL

Dadas as patologias, quais destas possuem amostras de RNA dadas por pessoas com mais de 60 anos?

```
select  --RNASample.sampleId,
        --RNASample.age,
        --RNASample.sex,
        distinct
            pathology.tissue

from RNASample RNASample
JOIN proteinAtlas_entry_rnaExpression_data rnaExpressionData ON RNASample.FK_DATA = rnaExpressionData.PK_proteinAtlas_entry_rnaExpression_data
JOIN rnaExpression rnaExpression ON rnaExpression.PK_rnaExpression = rnaExpressionData.FK_rnaExpression
JOIN proteinAtlas pa ON pa.PK_proteinAtlas = rnaExpression.FK_proteinAtlas
JOIN proteinAtlas_entry_pathologyExpression_data pathology ON pathology.tissue_organ = rnaExpressionData.tissue_organ

group by RNASample.age, pathology.tissue
having RNASample.age > 60
;
```

index	TISSUE
0	Ovarian cancer
1	Colorectal cancer
2	Thyroid cancer
3	Testis cancer
4	Breast cancer
5	Cervical cancer
6	Endometrial cancer
7	Head and neck cancer
8	Stomach cancer
9	Liver cancer
10	Renal cancer
11	Prostate cancer
12	Lung cancer
13	Urothelial cancer
14	Glioma