Functional Genomic project 2025

Composition

- 10 most populated organs given by

$$\text{top 15 populated}_{\text{RNA}} \cap \text{top 15 populated}_{\text{HISTO}}$$

  (discarding e.g. `Whole Blood`)
- Kept unique `SUBJID` samples in both histology and RNA (to avoid duplicates)
- For a given **organ** :
  1. **RNA expression matrix**

     (`/mnt/iribhm/GTEx/hugo/projet_vincent/output/`**organ**`/RNA_counts_adjusted.csv`)

  2. **Histology expression matrix**

     (`/mnt/iribhm/GTEx/hugo/projet_vincent/output/`**organ**`/HISTO_counts_adjusted.csv`)

  3. **Covariates** (composed of: `SUBJID`, `SEX`, `HARDY_SCALE`, `Pathology Categories`, `Pathology notes`, `AGE_DECADE`, `HISTO_DI`, `RNA_DI`, `RNA_INTEGRITY_NUMBER`, `ISCHEMIC_TIME`)

     (`/mnt/iribhm/GTEx/hugo/projet_vincent/output/`**organ**`/covariates.csv`)

Data

Expression matrices

RNA expression

1. Subset the raw counts matrix using the covariates RNA samples
2. Intersect the samples name (converted to SUBJIDs) with the sample names from histology (converted to SUBJIDs) and kept the associated columns
3. Filter out lowly expressed genes using `edgeR`'s `filterByExpr` function.
4. Expression matrix X becomes $\text{cpm}(X)$ using `edgeR`'s `cpm`
5. For each gene x_i,

$$x_i' = \text{Residuals}\Big(x_i \overset{\text{LOESS}}{\sim} \text{SMRIN} + \text{SMTSISCH}\Big)$$

   giving the adjusted matrix
6. For each gene x_i,

$$x_i'' = \lfloor (x_i' + |\min(x_i')|) \cdot 100 \rceil$$

   giving positive expression counts (suitable as an input for both `DESeq2` and `edgeR`).

Histology expression

- Same as for RNA, except 6. becomes $x_i'' = \lfloor (x_i' + |\min(x_i')|) \cdot 10 \rceil$

Dimorphism indices

- **Did not use neither weighted nor balanced bootstrapped estimates**