

RNA – RNA level expression of genes

1. **GENE ID** : ensemble ID (repeated for each sample ID) - **text**
2. **SAMPLE** : individual code (repeated for different genes) - **text**
3. **CANCER** : code for tumor type (repeated for many sample IDs) - **text**
4. FPKM : transcript per million (normalized count of RNA molecules for a certain gene) - **double**

Threshold on FPKM + cancer type → Gene ID

Converter – Conversion of gene ID

1. NCBI Gene ID : Unique gene ID - **int**
2. **Ensembl gene ID** : Unique gene ID - **text**
3. Approved symbol : Unique gene ID - **text**

Ensembl → Hugo/Approved symbols (one-to-one)

Onco genes – Onco-annotation for all genes

1. **Hugo Symbol** : Hugo symbol/ Approved symbol - **text**
2. Entrez gene ID : Entrez gene ID - **int**
3. OncoKB Annotated : Can be targeted for treatment - **text**
4. Is Oncogene : its level of expression changes tumor growth (+) - **text**
5. Is Tumor Suppressor gene : its level of expression changes tumor growth (-) - **text**

Hugo Symbol → Hugo symbol filtered according to yes/no (is oncogene)

Onco drug – FDA approved molecules

1. **Gene** : Hugo symbol/ Approved symbol - **text**
2. Alterations : Type of alteration of gene - **text**
3. Cancer Types : Tumor type (extensive version) - **text**
4. Drugs : Name of the drug - **text**

Hugo symbol filtered → Drugs

Queries:

- From drug to sample ID (Onco drug → Onco genes → Converter → RNA)
- Drugs associated to cancer type (Onco drug → Onco genes → Converter → RNA)
- Find tumor suppressor genes (Onco drug → Onco genes → Converter → RNA)

Improvements:

- Come stabilire threshold? Media tra tutti gli FPKM di uno stesso gene in uno stesso cancer type?
- Additional table for conversion of extended cancer type to code
- Pre-processing: binary instead of yes and no