Lien pour la cohort :
https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Endometrioid%20Cancer%20(UCEC)

---

gene expression RNAseq

- STAR - Counts (n=585) GDC Hub

More information on the GDC pipeline used to generate this data:
https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/

- STAR - FPKM (n=585) GDC Hub

More information on the GDC pipeline used to generate this data:
https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/

- STAR - FPKM-UQ (n=585) GDC Hub

More information on the GDC pipeline used to generate this data:
https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/

- STAR - TPM (n=585) GDC Hub

More information on the GDC pipeline used to generate this data:
https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/

---

dataset: gene expression RNAseq - STAR - Counts
hub: https://gdc.xenahubs.net
More information on the GDC pipeline used to generate this data:
https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/
cohortGDC TCGA Endometrioid Cancer (UCEC)
dataset IDTCGA-UCEC.star_counts.tsv
downloadhttps://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-UCEC.star_counts.tsv.gz; Full metadata
samples585
version05-10-2024
type of datagene expression RNAseq
unitlog2(count+1)
platformIllumina
ID/Gene Mappinghttps://gdc-hub.s3.us-east-1.amazonaws.com/download/gencode.v36.annotation.gtf.gene.probemap; Full metadata
authorGenomic Data Commons
raw datahttps://docs.gdc.cancer.gov/Data/Release_Notes/Data_Release_Notes/#data-release-400
raw datahttps://api.gdc.cancer.gov/data/
wranglingData from the same sample but from different vials/portions/analytes/aliquotes is averaged; all data is then log2(x+1) transformed.

input data formatROWs (identifiers) x COLUMNs (samples) (i.e. genomicMatrix)
60 661 identifiers X 585 samples All IdentifiersAll Samples

	TCGA-FI-A3PX-01A	TCGA-BG-A221-01A	TCGA-EY-A1GK-01A	TCGA-BG-A2AE-01A	TCGA-AX-A1CE-01A	TCGA-DI-A2QY-01A	TCGA-D1-A1O8-01A	TCGA-EY-A3QX-01A	TCGA-DI-A0WH-01A	TCGA-BK-A4ZD-01A
ENSG00000000003.15	12.67	10.67	10.90	10.54	11.98	13.01	9.246	11.80	10.76	12.75
ENSG00000000005.6	3.000	0.000	1.585	2.322	3.459	1.000	3.459	2.807	2.000	3.459
ENSG00000000419.13	11.69	9.322	9.248	8.807	9.506	10.42	8.849	10.25	8.028	10.80
ENSG00000000457.14	8.945	7.937	9.050	8.200	8.103	8.229	8.061	9.665	8.600	10.65
ENSG00000000460.17	9.338	7.700	7.119	6.820	6.807	7.435	7.044	9.799	7.443	10.36
ENSG00000000938.13	6.340	8.326	7.392	7.547	6.858	5.459	9.134	8.418	5.700	5.833
ENSG00000000971.16	8.907	8.954	7.901	9.236	8.574	9.514	9.887	11.24	10.14	7.919
ENSG00000001036.14	12.08	10.96	10.48	10.18	11.69	11.59	9.855	13.11	10.35	12.54
ENSG00000001084.13	10.60	9.192	9.224	8.087	10.51	10.92	9.006	11.78	8.134	12.19
ENSG00000001167.14	12.15	10.14	9.778	10.26	9.229	11.65	9.308	11.75	9.388	12.63

https://xenabrowser.net/datapages/?dataset=TCGA-UCEC.star_counts.tsv&host=https%3A%2F%2Fgdc.xenahubs.net
RNA-seq gene expression data (STAR pipeline, log2(count+1)) were downloaded from UCSC Xena (TCGA-UCEC cohort, GDC hub).

---

dataset: phenotype - Phenotype
hub: https://gdc.xenahubs.net
cohortGDC TCGA Endometrioid Cancer (UCEC)
dataset IDTCGA-UCEC.clinical.tsv
downloadhttps://gdc-hub.s3.us-east-1.amazonaws.com/download/TCGA-UCEC.clinical.tsv.gz; Full metadata
samples600
version09-07-2024
type of dataphenotype
authorGenomic Data Commons
raw datahttps://docs.gdc.cancer.gov/Data/Release_Notes/Data_Release_Notes/#data-release-410
raw datahttps://api.gdc.cancer.gov/data/
input data formatROWs (samples) x COLUMNs (identifiers) (i.e. clinicalMatrix)
600 samples X 79 identifiers All IdentifiersAll Samples

	age_at_diagnosis.diagnoses	age_at_earliest_diagnosis.diagnoses.xena_derived	age_at_earliest_diagnosis_in_years.diagnoses.xena_derived	age_at_index.demographic	alcohol_history.e
TCGA-AJ-A3NC-01A	2.335e+4	2.335e+4	63.98	63.00	Not Reported
TCGA-AJ-A3NC-11A	2.335e+4	2.335e+4	63.98	63.00	Not Reported
TCGA-KP-A3W0-01A	2.638e+4	2.638e+4	72.27	72.00	Not Reported
TCGA-B5-A11N-01A	2.544e+4	2.544e+4	69.70	69.00	Not Reported
TCGA-BS-A0V7-01A	1.781e+4	1.781e+4	48.79	48.00	Not Reported

TCGA-D1-A17Q-01A	2.002e+4	2.002e+4	54.84	54.00	Not Reported
TCGA-BG-A0VT-01A	2.060e+4	2.060e+4	56.44	56.00	Not Reported
TCGA-AP-A0LG-01A	2.005e+4	2.005e+4	54.92	54.00	Not Reported
TCGA-AP-A0LL-01A	2.070e+4	2.070e+4	56.71	56.00	Not Reported
TCGA-AX-A060-01A	2.847e+4	2.847e+4	78.01	77.00	Not Reported

https://xenabrowser.net/datapages/?dataset=TCGA-UCEC.clinical.tsv&host=https%3A%2F%2Fgdc.xenahubs.net&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

FICHIER 1 — Expression génique RNA-seq

Identité du fichier

- **Dataset ID** : TCGA-UCEC.star_counts.tsv

- **Type de données** : expression génique RNA-seq
- **Nombre d'échantillons** : **585**
- **Format** : matrice gènes × échantillons
  - lignes = gènes (IDs Ensembl)
  - colonnes = échantillons TCGA
- **Source** : GDC (via UCSC Xena)
- **Plateforme** : Illumina
- **Pipeline** : STAR (GDC RNA-seq pipeline)

🔢 **Nature exacte des valeurs (POINT TRÈS IMPORTANT)**

- Les valeurs **ne sont PAS des counts bruts**
- Les valeurs sont :

log2(count + 1)

Cela est explicitement indiqué par :

- unit : log2(count+1)
- wrangling : all data is then log2(x+1) transformed

🧠 **Interprétation méthodologique**

✔️ **Ce qui a déjà été fait par le pipeline** :

- alignement RNA-seq (STAR)
- comptage par gène
- agrégation des aliquots d'un même échantillon
- transformation **log2(count + 1)**

❌ **Ce qui n'est PAS fait** :

- pas de z-score
- pas de scaling par échantillon pour le ML
- pas de sélection de gènes
- pas de QC biologique (outliers, gènes peu exprimés)

👉 **Donc** :

- ce fichier est **semi-normalisé**
- tu **ne peux plus revenir aux counts bruts**
- mais tu **dois encore faire** :
  - QC gènes
  - QC échantillons
  - standardisation (z-score) avant NN

🎯 **Usage dans ton projet**

- **Entrée principale** du MLP supervisé
- **Entrée** de l'autoencodeur
- Très bien adapté aux réseaux de neurones

👉 À l'oral, tu pourras dire :

"Nous avons utilisé les données RNA-seq fournies par le pipeline GDC (STAR), transformées en log2(count+1), puis appliqué un QC et une standardisation adaptés aux réseaux de neurones."

🏷️ **FICHIER 2 — Données cliniques / phénotypes**

📄 **Identité du fichier**

- **Dataset ID** : TCGA-UCEC.clinical.tsv
- **Type de données** : phenotype (données cliniques)
- **Nombre d'échantillons** : **600**
- **Format** : table échantillons × variables
- **Source** : GDC (via UCSC Xena)

🧬 **Contenu principal**

Ce fichier contient notamment :

- âge au diagnostic
- histologie
- grade tumoral
- informations de suivi
- **annotations moléculaires TCGA**

👉 **C'est dans ce fichier que se trouvent les sous-types TCGA** :

- POLE
- MSI
- Copy-Number Low
- Copy-Number High

(Certains noms de colonnes sont longs / peu lisibles, on les identifiera ensemble dans Python.)

🧠 **Interprétation méthodologique**

✔ Données **déclaratives / cliniques**
✔ Aucune normalisation nécessaire
✔ Sert à :
- définir les **labels** pour la classification
- faire du QC (cohérence échantillons)
- enrichir l'interprétation biologique

❌ **Ce n'est PAS** une entrée du réseau de neurones (dans un premier temps)

🎯 **Usage dans ton projet**
- Définir la variable cible du MLP (classification multi-classe)
- Vérifier la cohérence des échantillons RNA-seq
- Faire des analyses descriptives (répartition des sous-types)

📜 **RÉCAPITULATIF SYNTHÉTIQUE**

Fichier	Rôle	Normalisation
TCGA-UCEC.star_counts.tsv.gz	Features (X)	log2(count+1) **déjà fait**, z-score **à faire**
TCGA-UCEC.clinical.tsv.gz	Labels (y) + clinique	aucune normalisation