**Data cleaning:**

* Remove irrelevant columns from the dataset

**Mutation data:** Deals with change of DNA sequence: deletion, insertion,

**Genomic Data:** A large number of columns appear to represent gene expression levels or other genomic features.

**Gene expression data**: refers to the information that shows which genes are active, and to what extent, in a specific cell or group of cells at a given time. This data is crucial in understanding how genes contribute to the functioning, development, and overall health of an organism. Gene expression is a dynamic process and can change in response to various internal and external factors.

These columns are related to patient information, clinical characteristics, and treatment details.

* **Patient Information:**
  + patient\_id, age\_at\_diagnosis
* **Cancer, Treatment Details, and Clinical Features:**
  + type\_of\_breast\_surgery, cancer\_type, cancer\_type\_detailed, cellularity chemotherapy, hormone\_therapy, radio\_therapy
* **Biological and Clinical Markers:**
  + er\_status\_measured\_by\_ihc, er\_status, pr\_status, her2\_status\_measured\_by\_snp6, her2\_status, 3-gene\_classifier\_subtype
* **Cancer Characteristics:**
  + neoplasm\_histologic\_grade, tumor\_other\_histologic\_subtype, tumor\_size, tumor\_stage primary\_tumor\_laterality, lymph\_nodes\_examined\_positive
* **Prognostic Indices:**
  + nottingham\_prognostic\_index, integrative\_cluster
* **PAM50 plus Claudin-low Subtype:** 
  + A molecular classification of breast cancer
* **Outcome Data:**
  + overall\_survival\_months, overall\_survival, death\_from\_cancer
* **Cohort Information:**
  + cohort, oncotree\_code
* **Other Biomarkers**:
  + er\_status\_measured\_by\_ihc, HER2 status, etc.