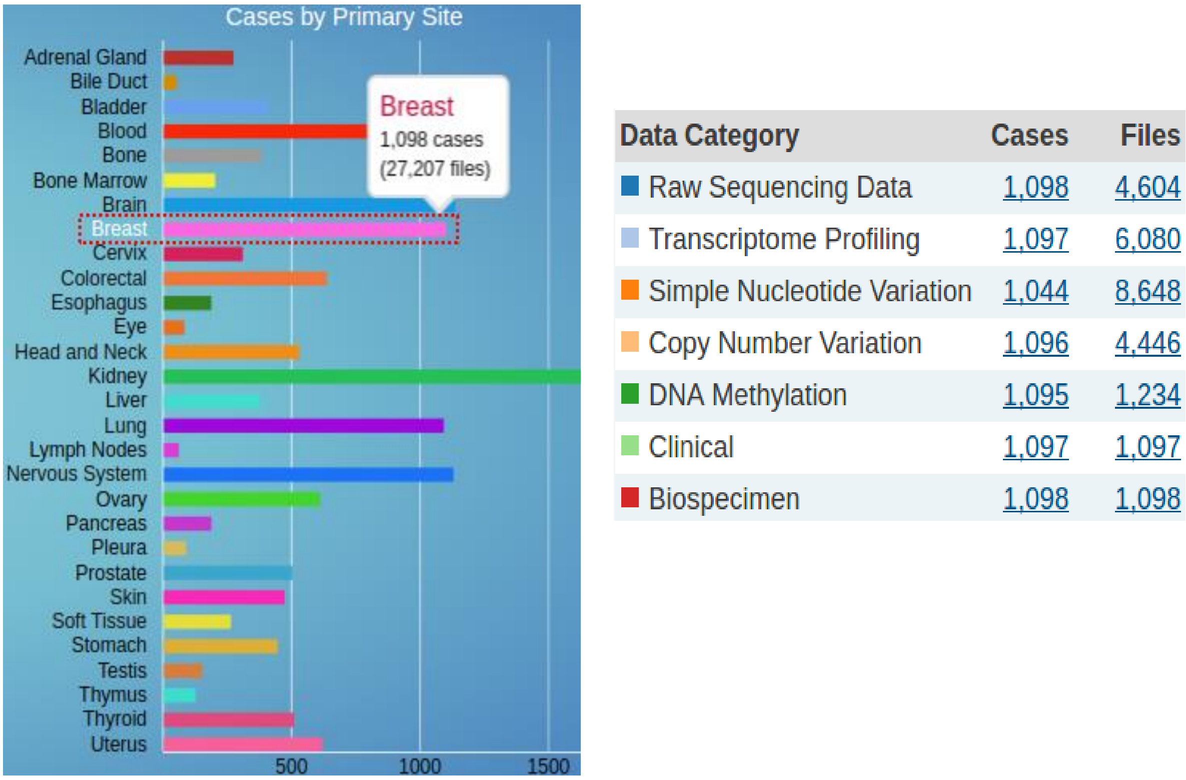
Dataset preparation for deep learning algorithms

Description of the TCGA dataset

The Cancer Genome Atlas (TCGA) is a catalog of genetic mutations responsible for cancer. It is used widely in genome sequencing and bioinformatics. The Cancer Genome Atlas (TCGA) is a US government-funded research project that collects genomic data for cancer analysis. There are different types of genomic data publicly available under this project. Copy number variations (CNV) are one of the genomic data which are publicly available at TCGA website. There are about 33 different cancer types' data available.

TCGA-BRCA is their project that focuses on breast cancer analysis. As shown in the below figure, there exist 27,207 files on the TCGA-BRCA section from 1,098 cases (patients) in the Genomic Data Commons (GDC) data portal. These data can be categorized into 13 categories, which 9 of them can be accessed publicly.



Data Category	Cases	Files
🟦 Raw Sequencing Data	1,098	4,604
🟦 Transcriptome Profiling	1,097	6,080
🟧 Simple Nucleotide Variation	1,044	8,648
🟧 Copy Number Variation	1,096	4,446
🟩 DNA Methylation	1,095	1,234
🟩 Clinical	1,097	1,097
🟥 Biospecimen	1,098	1,098

Fig. 1: *(left) Number of cases and files in TCGA-BRCA project; (right) Number of files per data type and number of cases who had those corresponding files*

Table 1: data selection criteria from this large collection of data set

Data Type	Explanation	Used/Not-used
Somatic Mutation	Description of Single-Nucleotide Polymorphism (SNP) per patient.	**Not-used** because data is too large (3x10^9 base pair) yet too sparse (very little SNP is known)
Copy Number Segment	Amount of Copy Number Variation (CNV) per patient.	**Used as input for cancer type detection** because the data is based on arbitrary CNV position per patient. So the size of data per patient is different.
DNA Methylation	Amount of methylated DNA per CpG probe identifiers.	**Used as input**
Gene Expression	Amount of gene expression per patient based on gene id.	**Used as input**

miRNA Expression	Amount of miRNA expression per patient based on miRNA identifiers.	**Used as input**
Isoform Expression	Expansion of miRNA expression data.	**Not-used** because the data is just the expansion of miRNA expression data. Also, data size per patient is different.
Clinical	Patient id, diagnosis, treatment type and result, and etc.	**Used as label**

Data preprocessing

Download all used file (DNA methylation, gene expression, miRNA expression, clinical). All file is named after file_id, without any trace of case_id (patient_id). We then request meta files from GDC API. These include:

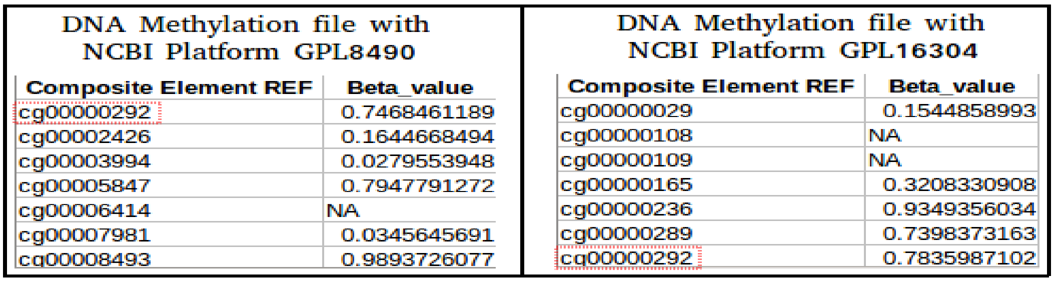
- list of the patient (in case_id)
- list of files (file_id) owned by each patient (in case_id)
- creating a list of file amount per data type for each patient (in case_id) based on aforementioned metafiles

Then we preprocessing clinical data from XML files per patients to JSON and CSV format per categories. These categories include:

- list of id (patient id, drug id, radiation id, follow-up id),
- general pathology (histological type, site, biopsy method, and etc),
- receptor status (ER status, PGR status, HER2 status), etc.

Preprocessing DNA methylation data:

- From all DNA methylation files, some of them came from normal sample instead of the tumor sample. All files from normal sample are thrown away because our prediction concern for different types of cancer, instead of differentiating normal patients from cancer patients
- The DNA methylation files on the TCGA-BRCA project came from two different types of sequencing platform. From all 1,095 DNA methylation files (each corresponding to 1 distinct patient) that were used, 342 of them used the NCBI Platform GPL8490, while the remaining 892 used the NCBI Platform GPL16304.



DNA Methylation file with NCBI Platform GPL8490

Composite Element REF	Beta_value
cg00000292	0.7468461189
cg00002426	0.1644668494
cg00003994	0.0279553948
cg00005847	0.7947791272
cg00006414	NA
cg00007981	0.0345645691
cg00008493	0.9893726077

DNA Methylation file with NCBI Platform GPL16304

Composite Element REF	Beta_value
cg00000029	0.1544858993
cg00000108	NA
cg00000109	NA
cg00000165	0.3208330908
cg00000236	0.9349356034
cg00000289	0.7398373163
cg00000292	0.7835987102

Fig. 2: *First 7 rows of both types of DNA Methylation files. Files with Platform GPL8490 have total of 27,578 rows and files with Platform GPL8490 have total of 485,577 rows*

The tables above show the difference between Platform GPL8490 files and Platform GPL16304 files. Platform GPL8490 only consist of 27,578 CpG sites, while Platform GPL16304 consist of 485,577 CpG sites. However, 25,978 CpG sites of those are intersected between that two platform (e.g. cg00000292 as was shown in tables above).

So, this preprocessing step will filter in those 25,978 CpG sites from both types of file and filter out the remaining CpG sites. The **Beta_values** are the one that we used as the Neural Network input.

Preprocessing gene expression data:

- From all gene expression files, some of them came from normal sample instead of the tumor sample. All files from the normal sample are thrown away because our prediction concern for different types of cancer, instead of differentiating normal patients from cancer patients.

Preprocessing miRNA expression data:

- From all miRNA expression files, some of them came from normal sample instead of the tumor sample. All files from the normal sample are thrown away because our prediction concern for different types of cancer, instead of differentiating normal patients from cancer patients.

Preprocessing of CNV data:

1. Data Collection

We used 14 different most common cancer types data where each cancer type has at least 400 samples. From each cancer patient, two separate tissue samples are collected. One sample is from cancer tissue and another is from normal tissue(usually from blood samples). Copy number variations are collected from both cancer samples and normal samples. In this study, we want to predict whether a sample is from normal tissue or cancer tissue sample. We also want to predict the type of cancer of a cancerous sample.

There are several ways to download the data from TCGA website. To download bulk data, the easiest would be to use TCGA data transfer tool. It can be downloaded from here
https://gdc.cancer.gov/access-data/gdc-data-transfer-tool
To download CNV data of cancer, first, go to https://portal.gdc.cancer.gov/.

Then **repository->cases->project->select a project(TCGA-GBM for example)**. At the top of this webpage, there are options to select different file categories such as data category, access level etc. At top of the same page, under "**Advanced Search**" option, it is also possible to use different queries to filter different types of files in a group. A sample query we used to download GBM cancer samples of male patients:

```
cases.project.project_id in [TCGA - GBM] and files.access in ["open"] and
files.data_type in ["Copy Number Segment"] and cases.demographic.gender in [male]
and cases.samples.sample_type in ["Primary Tumor"] and files.platform in
["Affymetrix SNP 6.0"]
```

In above, we select GBM project, files that are publicly available and we only select copy number variation files. For each cancer project, we customize above search query according to the diagram below:

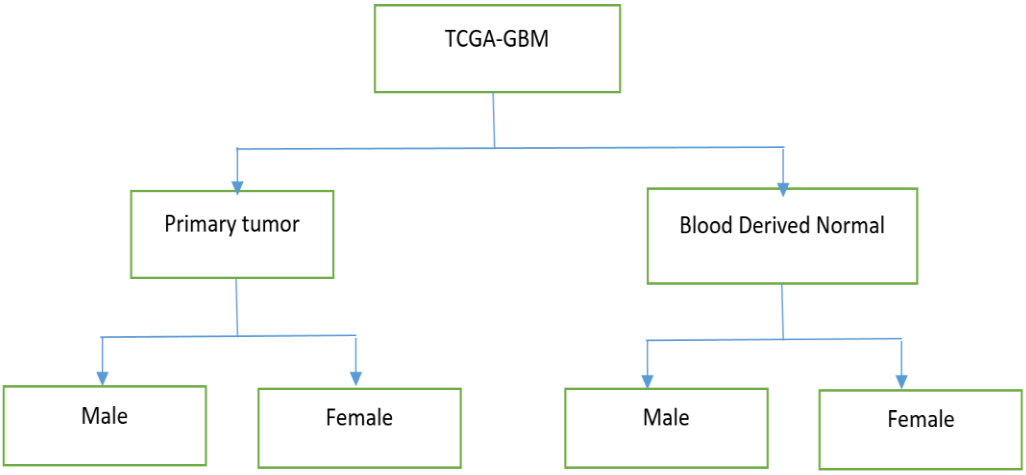


Fig. 3: Download each cancer type samples in four different groups

TCGA CNV data do not contain the information whether a sample is from normal tissue or cancer tissue, also gender and type of cancer are not specified as well. But these are important information that could help our algorithm train better. For example, breast cancer is common in female but not in the male. Similarly, Prostate Adenocarcinoma(PRAD) is found only in the male. Since this information are not specified in the CNV data, we will download the files in separate groups so that we can add this information later in the data.

By changing the above query, we can group different files together, then select "Add all files to the Cart" in TCGA. A number of files selected will be shown in the Cart on the top right corner. Then download the manifest from Cart->Download->Manifest.

Now sample files could be downloaded by running the command from the terminal:

```
$ ./path_to_gdc_tool download -m path_to_manifest
```

It will look something like this:

```
$ ./gdc-client download -m gdc_manifest_20171121_221707.txt
```

We should download the files in organized groups so that our next steps become easier.

2. Data Preprocessing

Data preprocessing has two steps. First, we need to merge all cancer sample data and the normal sample data. Then we remove noisy segmentation values between a certain range.

2.1 Merge files

In our previous download step, we got a normal sample file and a cancer sample file from each patient. We need to merge the data of all these files into one file. Merging of files will be done in following steps. For each cancer type, we:

- Merge all-male sample data
- Add a new column "ctype" which means cancer type and give it a unique number which will represent the type of the cancer
- Add another new column "gender" and give 1 for male
- Now, merge all-female sample data
- Add a new column, "ctype" meaning cancer type and give it same number as given for male samples
- Add another new column "gender" and give 0 for female
- Repeat the same process with for other cancer samples.

2.2 Preprocessing

Most of the machine learning problems deal with numeric data. In our data, column "chr" represents chromosome number have string data "X" and "Y" which represents X chromosome and Y chromosome. We replace "X" value of "chr" column with 23 and "Y" with 24. We add another new column "cnv_length" which represents the length of a CNV and its value is the difference between "end" and "start" positions of a CNV.

Another important feature in CNV data is segmentation meas. Segmentation means represents the number of CNVs at a DNA location. The higher the segmentation mean, the higher the number of CNV at that location. And negative segmentation mean represents copy number loss and a positive value means an amplification of copy numbers. We remove copy numbers from our calculation whose segmentation value is between -0.2 and +0.2. Segmentation values between -0.2 and +0.2 are usually considered noise.

3. Download gene data and preprocess

A list of mutated genes for each cancer is available under each project in TCGA website. We will use this data to prepare our final datasets for deep learning algorithms.

3.1 Download Gene data

For each cancer, TCGA list all the affected genes found in different tissue samples. It can be downloaded from https://portal.gdc.cancer.gov/ repository->cases->project->select the project name from the table->Most Frequently Mutated Genes->select JSON. This will download a JSON file containing all the mutated genes for this cancer.

3.2 Extract Gene Details

In the above-downloaded gene files, genes details are given in JSON format. To make our calculation easier, we will convert this data to a CSV file. Most of the files contain an almost same number of genes except few exceptions.

Hence, we merge the gene details from all files and remove duplicate genes. In TCGA CNV data, we have coordinates as start and end points of DNA but gene coordinates are given as cytoband in our downloaded gene files. We need to get gene coordinates as start and end points for our later calculation.

3.3 Get Gene Coordinates

To get the gene coordinates, we used an R-script. Using, “ensembl” library of “biomaRt” package of R. We can send a request to ensembl genome browser to get the details of a gene. It acts as API for ensembl genome browser. Example of a sample query:

```
getBM(attributes=c('chromosome_name', 'start_position', 'end_position', 'strand'),
    filters=c('hgnc_symbol'), values='RPL8',     mart=ensembl)
```

The above query should generate output as follows:

```
chromosome_name start_position end_position strand
1               8      144789765    144792587     -1
```

In the above query, we are sending a request to get different properties such as 'chromosome name', 'start position', 'end position' and 'strand' of the gene ‘RPL8’. Filter 'hgnc_symbol' ensures that the requested gene is from homoserine groups. For each gene symbols, For each gene, we do the same.

3.4 Preprocess Genes

Like before, we need to change chromosome name “X” to 23 and “Y” to 24. In ensembl database, there are few genes whose details are missing.

Hence we remove those genes with missing information. Also, some gene names appear multiple times. We remove those duplicate genes as well. At the end of our processing, we got a list of 20329 unique genes that were found in 14 different cancer patients.

4. Prepare data for the deep learning algorithms

To prepare data suitable for machine learning problems, first, we need to understand the data.

4.1 CNV and gene data details

A glimpse of CNV sample data looks as follows:

```
Sample    Chromosome        Start          End         Num_Probes       Segment_Mean
1           1               61735          855630         59            0.243
1           1               857100         3291811        693           -0.3489
1           2               3296382        5963169        2063          0.0184
```

In above each row means, a sample with ID 1 has copy number variation in chromosome 1 and the variation extends from location 61735 to 855630. The positive segmentation mean refers that, there is copy number gain at this location. The Same explanation is applicable for the second row as well, except its’ negative segmentation mean represents there is copy number loss at that given location. In our preprocessing step we added few more new information such as gender, cnv length which is the difference between the two locations end and start and the type of cancer.

After processing gene data, it looks like this

```
   gene_id          name        chr    start         end     strand    is_cancer_gene_census
ENSG00000234585    CCT6P3        7     65038354     65074713    1
ENSG00000166822    TMEM170A      16    75443054     75465497   -1
ENSG00000108946    PRKAR1A       17    68511780     68551319    1             True
```

Here the first row tells us that, gene `CCT6P3` is located on chromosome 7 from location 65038354 to 65074713. The last column “is_cancer_gene_census” is true if the gene is an oncogene. Genes which play important role in cancer growth are called oncogenes. Till now, Till now, there are 576 genes were identified as oncogenes by TCGA. Due to lack of sufficient data, we removed 8 oncogenes from our calculation. In our experiment, we will use total 20329 genes.

Considering gender and genes as features, we prepare two datasets to work with, one dataset with all genes as features and another dataset with only oncogenes as a feature. For each gene, we check whether a CNV overlaps with a gene coordinate. If it overlaps, we take segmentation value at that coordinate meaning the change of copy number of a gene at that location.

A gene is considered to overlap with a CNV if its starting point or ending point falls between the starting or ending points of a CNV or vice versa. What we mainly check here is that, due to cancer which genes are changed in terms of copy number variation. Usually, each person has different copy numbers but cancer affected genes will have more changes in copy numbers. We are using different deep learning algorithms to find patterns in gene copy number change due to cancer and determine whether a person has cancer or not. We also want to determine different cancer types based on the copy number change pattern of genes.