Supplementary Material

7. Genomics

In this section we provide a more detailed description for the MSK-IMPACT dataset (Zehir et al., 2017).

7.1. Synopsis: MSK-IMPACT 2017

Data_CNA:

In this file, each column is a sample, and each row is a feature. Each feature is represented by a string (e.g., EGFR) which is the HUGO symbol for the corresponding gene. The feature values are floats, which represent the copy number alteration of the feature's gene. Positive and negative numbers represent duplication or deletion of repeat nucleotides respectively.

Data_Fusion:

This file structure is not as straightforward from an ML perspective. Some general information is easily parsable, but more specific information contained in the `comments` column requires more creative approaches and possibly expert (biochemical background) consultation. The file comprises many rows, each with a gene HUGO symbol and sample ID, among other fields (see below). Each row describes half a fusion in which two genes are mixed together via deletion, inversion, or translocation. Since two genes are involved, two rows describe each half of the fusion effect. The only time this is not true is when the fusion is intergenic. This is when a gene is mixed with DNA material which is located between genes, that is, is non-coding. In this case only the information for the coding DNA material in the fusions is listed in a single row. The remaining fields in each row are as follows:

- Huge_Symbol: unique gene identified for the 410 genes sequenced with some being post-fixed with an Arabic numeral, a Latin letter, or in special cases both to indicate membership in a gene family
- Entrez_Gene_ID: a gene integer ID used by the National Center for Biotechnology Information (NCBI), only present for two samples
- Center: where the sample was taken; set to "MSKCC-DMP", signifying the Memorial Sloan Kettering Cancer Center, for all samples in the file
- Tumor_Sample_Barcode: sample ID
- Fusion: indicating which gene(s) are involved in the fusion
- DNA support: "yes" for all samples, indicating that fusion was detected with DNA sequencing
- RNA support: "unknown" for all samples, indicating that RNA-sequencing was not preformed
- Method: is "NA" for all samples
- Frame: either "unknown", "out-of-frame", or "in-frame" indicating the effect of each fusion
- Comments: short comments written by practitioners giving specifics of the mutation

Data_SV:

structural variation, is a more general classification for mutations than fusion mutations or copy number variations. There are 31 categorical, numerical, and text features describing the type, location, and prevalence of the structural variances.

- Annotation: specific type and location of structural variation
- Breakpoint_Type: the type of breakpoint, that is, the junction between normal and rearranged
- Comments: small notes on some mutations
- Confidence Class: indicates the confidence in the final sequencing, and whether it was automatically or manually determined

The remaining features are not documented in a single source but detail various aspects of the structural variations such as sequencing method, the quality of sequencing, comparison with germline DNA, variation location, and so on.

Data_Mutation_Significance_Contribution:

This file contains 30 numerical features corresponding to known mutation signatures. For each sample, the feature value is the percentage of mutations explained by the corresponding signature

Data_Mutation_Significance_Confidence:

30 numerical features that correspond to the same mutation signatures found abvove, but the feature value is the confidence in the contribution scores instead. The mechanism used to determine confidence in contribution score is described in Huang (2018)[2]

Data_Mutation_mskcc:

This file contains 46 columns describing mutations using a subset of the columns found in the mutation annotation format (MAF) format. All columns and their descriptions can be found in the GDC MAF Format[3], except two:

- Hotspot: zero for all samples; was not configured when the table was made, so the same label was applied to all mutations
- cDNA_change: the nucleotide change described with HGVS expression

Data_Mutation_extended:

This is identical to Data_Mutation_mskcc but is missing the "cDNA_change" column

Data_Gene_Panael_Matrix:

This file records which type of panel was used to extract genomic data from the sample: the 341 gene or 410 gene panel. The 341 genes are a subset of the 410 genes, so this distinction does not manifest in the other records because when creating a record from a sample that used the 341 gene panel; the 69 genes are not tested for were marked as not detected

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5860213/
[3] https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/

Data_cna_hg19:

This file is the output of DNAcopy, which is an open-source software package written in R that "identifies genomic regions with abnormal copy number", that is, copy number alterations. Each row in the file corresponds to one such region or "segment" and has five columns describing it:

- `ID:` the sample ID
- `chrom:` the chromosome the segment is in
- `loc.start:` start location of the segment
- `loc.end:` end location of the segment
- `num.mark:` number of probes bound to the segment
- `seg.mean:` the mean value, across all probes, of the segment; represents the `log2-ratio` of tumor copy number to normal copy number; a positive value indicates that the tumor has a higher copy number and vice versa

Data_clinical_patient:

Each row corresponds to a patient and contains five columns:

- #Patient Identifier: the patient ID
- Sex: the patient's gender
- Patient's Vital Status: whether they are deceased or alive
- Smoking History: whether they previously, currently, or never smoked
- Overall Survival (Months): How long they survived since initial diagnosis; blank if they are currently alive or died after last follow up
- Overall Survival Status: the same as Patient Vital Status with a slightly different format

Data_clinical_sample:

Each row corresponds to a sample and contains 16 columns:

- #Patient Identifier: the patient ID
- Sample Identifier: the sample ID
- Sample Collection Source: whether the sample was collected in house at MSKCC or by another party
- Specimen Preservation Type: method used to preserve the sample
- Specimen Type: how the specimen was collected
- DNA Input: amount of DNA in the sample in nanograms
- Sample Coverage: number of unique reads that included a given nucleotide during sequencing
- Tumor purity: percentage of cancer cells in sample
- Matched Status: whether the patient was matched with a gene therapy
- Sample Type: whether the sample came from the primary or metastatic tumor
- Primary site: the location of the primary tumor
- Metastatic site: where the metastasis occurred
- Sample Class: type of cancer tissue (is tumor for all samples)
- OncoTree Code: unique code for specific tumor type
- Cancer Type: type of cancer that caused the tumor
- Cancer Type: Detailed: more specific sub-type of the cancer

7.2. Glossary

breakpoint the beginning or end point of a structural variation

chromosomal inversion mutation in which a segment of a chromosome is reversed

chromosomal translocation two types; Robertsonian translocation, which occurs when two non-homologous chromosomes get attached, and reciprocal translocation, which occurs when parts are exchanged between two non-homologous chromosomes

chromosome a molecule that contains part of the genome in a condensed, manageable package

copy number alteration a change in copy number that has arisen in any cell of the body after conception

fusion when two previously independent genes are combined or fused together that results from chromosomal translocation, interstitial deletion, and chromosomal inversion

gene family a set of several similar genes formed by duplication, generally with similar function

genome genetic material of an organism

homologous chromosomes two chromosomes that carry the same genes, one from each parental source

HUGO symbol unique gene identifier derived from the HUGO gene nomenclature guidelines

in-frame a mutation, where the translation into protein is not completely disrupted, creating still-functional proteins

interstitial deletion a mutation in which part of the DNA (not including the terminal portion of a chromosome) is left out during DNA replication

metastasis secondary malignant growth distant from the primary site of cancer

mutation alteration of the nucleotide sequence of the genome

mutation signatures characteristic combinations of mutations arising from specific mutation processes

non-homologous chromosomes two chromosomes that do not carry the same genes in contrast to homologous chromosomes

out-of-frame a mutation, where the translation into a protein is completely disrupted, creating non-functional proteins.

primary site the location on the body where the first tumor progression begins

somatic mutation genetic alteration acquired by cells that are the progeny of cancerous cells

structural variation any kind of structural variation to the genome

8. Homomorphic Encryption

Below is an example presented by Sathya et al. (2018) to introduce the high-level concept of Homomorphic Encryption (HE).

1. Let m be the plain text message.
2. Let a shared public key be a random odd integer p.
3. Choose a random large q, small r, $|r| \leq p \div 2$.
4. Ciphertext $c = pq + 2r + m$ (ciphertext c is close to multiple of p).
5. Perform homomorphic addition/multiplication as required.
6. Decrypt: $m = (c \mod p) \mod 2$.

Homomorphic addition can be illustrated as follows

$$c_1 = q_1 \times p + 2 \times r_1 + m_1 \tag{3a}$$
$$c_2 = q_2 \times p + 2 \times r_2 + m_2 \tag{3b}$$
$$c_1 + c_2 = (q_1 + q_2) \times p + 2 \times (r_1 + r_2) + (m_1 + m_2), \tag{3c}$$

and Homomorphic multiplication as follows

$$c1 = q1 \times p + 2 \times r1 + m1 \tag{4a}$$
$$c_2 = q_2 \times p + 2 \times r_2 + m_2 \tag{4b}$$
$$c_1 \times c_2 = ((c_1 \times q_2) + q_1 \times c_2 \times q_1 \times q_2) \times p + 2(2 \times r_1 \times r_2 + r_1 \times m_2 + m_1 \times r_2) + m_1 \times m_2. \tag{4c}$$

Although homomorphic encryption holds massive potential in theory, it suffers from notable shortcomings in practice. In many cases, it is limited to only addition and multiplication meaning many functions must be approximated with high degree polynomials which incur a large computational overhead. Even when using only this subset of operations, homomorphic operations are orders of magnitude slower than conventional operations on plaintext data. Homomorphic encryption also leads to substantial ciphertext expansion of a magnitude proportional to the targeted security strength. Further, homomorphic encryption schemes do not allow unlimited operations without first decrypting and re-encrypting or running an expensive denoising operation. These and other considerations make homomorphic encryption not easily adaptable to practical applications without substantial foresight and planning. For a more in depth review of the limitations and practical considerations of homomorphic encryption (Aslett et al., 2015b).