Project-01-group-01

Chosen cancer type of interest

Brain cancer

Dataset

Loading data set as follows:

```
allDepMapData = readRDS("~/R/bioinfoproject/project-01-group-01/AllDepMapData.RDS")
names(allDepMapData)
```

```
## [1] "expression" "copynumber" "mutation"   "kd.ceres"   "kd.prob"
## [6] "annotation"
```

The list named as `allDepMapData` consists of the following matrices:

- **Gene expression:** consists of gene TPM (transcripts per million) values, which reflect the level of gene expression. Higher values suggest over expression of genes and vice versa. The rows are the gene names and the columns the cancer cell line identifiers.

  ```
  dim(allDepMapData$expression)
  ```

  ```
  ## [1] 49070   544
  ```

- **Gene copy number:** consists of gene copy number (CN) values. These values reflect the copy number level per gene, higher values (CN > 2) might reflect amplification, lower values (CN < 2) might reflect deletion. The rows are the gene names and the columns the cancer cell line identifiers.

  ```
  dim(allDepMapData$copynumber)
  ```

  ```
  ## [1] 23299   544
  ```

- **Gene mutations:** annotation file for the various mutations observed in a sample. The `isDeleterious` flag specifies if the mutations has a functional effect or not.

- **Gene knockdown (CERES):** consists of gene knockdown scores. The score shows how essential the gene is for cell survival. Smaller values reflect higher essentiality. The rows are the gene names and the columns the cancer cell line identifiers.

  ```
  dim(allDepMapData$kd.ceres)
  ```

  ```
  ## [1] 17634   544
  ```

- **Gene knockdown (probability):** probability values for the effect of gene knockdown. A higher probability signifies that knocking down the gene very likely reduces cell proliferation.

  ```
  dim(allDepMapData$kd.prob)
  ```

```
## [1] 17634   544
```

- **Annotation:** gives information regarding the cell lines.

```
dim(allDepMapData$annotation)
```

```
## [1] 544   7
```

```
colnames(allDepMapData$annotation)
```

```
## [1] "DepMap_ID"       "CCLE_Name"       "Aliases"         "Primary.Disease"
## [5] "Subtype.Disease" "Gender"          "Source"
```

```
rownames(allDepMapData$annotation)
```

Things to do

- Identification of 3-5 most prominent mutations or genetic alterations driving cancer growth (non-drugable)
- Identification of appropriate cell line models (differ between cell lines with and without driver mutation)
- Read papers to extract interesting questions we want to address
- Timetable, summary of literature
- Data cleanup, exploration, reduction and modelling

Questions

- Important driver mutations (especially non-drugable) and associated passenger mutations?
- Shall we just focus on glioblastoma? (28 instead of 44 cell lines)

Data cleanup

- Extracting cell lines with brain cancer as primary disease

```
Celllines_BC = allDepMapData$annotation = allDepMapData$annotation[allDepMapData$annota
tion$Primary.Disease == "Brain Cancer", ]
summary(Celllines_BC)
```

```
##   DepMap_ID          CCLE_Name           Aliases
##  Length:44          Length:44          Length:44
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##                   Primary.Disease        Subtype.Disease    Gender
##  Brain Cancer              :44     Glioblastoma   :28           : 6
##  Bladder Cancer            : 0     Medulloblastoma: 6     -1    : 0
##  Bone Cancer               : 0     Glioma         : 4     Female:13
##  Breast Cancer             : 0     Astrocytoma    : 3     Male  :25
##  Colon/Colorectal Cancer   : 0     Meningioma     : 1
##  Endometrial/Uterine Cancer: 0     Neuroglioma    : 1
##  (Other)                   : 0     (Other)        : 1
##      Source
##         :14
##  ATCC   :11
##  HSSRB  : 7
##  DSMZ   : 4
##  HSRRB  : 3
##  KCLB   : 2
##  (Other): 3
```

Literature review on brain cancer

TCGA's Study of Glioblastoma Multiforme (GBM)

- most common, fast-growing, malignent brain tumor in adults with poor prognosis and survivability (less than 15 months) and no effective long-term treatments

Comprehensive genomic characterization defines human glioblastoma genes and core pathways (2008)

- 91 samples (72 untreated and 19 treated cases)
- High level focal amplifications for *EGFR* (41/91), *CDK4*, *PDGFRA*, *MDM2* and *MDM4*
- Significant mutated somatic genes: *TP53*, *PTEN*, *NF1*, *EGFR*, *ERBB2*, *RB1*
- Frequent genetic alterations in three critical signalling pathways:
    - RB pathway (most common: homozygotic deletion of *CDKN2A/CDKN2B* (55%/53%) or amplification of *CDK4* (14%) locus)
    - Inactivation of p53 pathway by *ARF* (55%) deletion (followed by *TP53* mutation), *MDM2* (11%) or *MDM4* (4%) amplificaton, additionally to *p53* mutation
    - RTK/RAS/PI3k pathway: deletions or mutations of *PTEN* and at least of one RTK (*EGFR*, *ERBB2*, *PDGFRA* 13% , *MET* 4%)
- *MGMT* methylation status (promotor of DNA repair enzyme) predicts sensitivity to alkylating agents (treatment method which leads to cell death). But cells develop resistancy and a hypermutator phenotype (MMR-defective) … **selective strategy to target mismatch-repair-deficient cells combinated with alkylating therapy to prevent resistance** (didnt quite understand the molecular background, but this outlook was part of the discussion)

The Somatic genomic landscape of glioblastoma (2013)

- GBM growth is driven by a signaling network with functional redundancy that permits adaptation in response to targeted molecular treatments
- GBM subtypes with molecular and epigenetic differences, which may affect clincical outcome and sensitivity of individual tumors to therapy
    - Subtypes: proneural, nerual, classical and mesenchymal
    - Proneural with G-CIMP and non-G-CIMP phenotype
    - survival advantage of proneural subtype with G-cIMP **(-> target genes affected by G-CIMP phenotype likely contribute to improved diagnosis; possible with our data to analyze?)**
    - least survivability with non-G-CIMP proneural GBMs and not mesenchymal GBM
- *MGMT* DNA Methylation as a predictive biomarker, but just in GBM classical subtype
- 40% of tumors harbor at least one nonsynonymous mutation among the chromatin-modifier genes (chromatin organization may play role in GBM pathology), additionally to alterations in siganture oncogenes like *EGFR*
- high frequency of complex *EGFR* fusion and deletion variants (also *MDM2* and *CDK4*) -> different *EGFR* alterations might exist concurrently in a tumor and yield differential biological activities/responses to any given targeted inhibitor **(find common second-site targets. Possible?)**
- Confirms results of 1st paper (251 samples):
    - at least one RTK was altered in 67% of GBM overall (mostly *EGFR* 57% and *PDGFRA* 13%)
    - 25% of GBM show PI3K-kinase mutation and 35% *PTEN* mutation/deletion (mutually exclusive)
    - 10% of GBM had *NF1* mutated or deleted
    - 90% of GBM had at least one alteration in PI3K pathway, 40% had more than one
    - p53 pathway was dysregulated in 85% through mutation/deletion of *TP53* (28%), amplification of *MDM1/2/4* (15%) oder deletion of *CDKN2A* (58%); mutually exclusivity between *TP53* and *MDM/CDKN2A*
    - 80% of tumors had alteration of Rb function: *CDKN2A* 56%, *CDK4/6* 16% and *RB1* 8%
- subset of individuals survives more than 3 years
    - **Question: factors for long-term survival (guess not realizable with our data)**

Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality

- Combination of two mutated genes is lethal for the cell, while the mutation of just on gene is viable
- Find synthetic lethal (SL) partners of non-drugable mutatated genes and use them as treatment approach
- DAISY = data mining SL identification pipeline
    - Approach for statistically inferring SL-interactions (SLI) from cancer genomic data of both cell lines and clinical samples
    - Genomic survival of the fittest (SoF): Cells with SL coinactivation are eliminated from the population. Identification of SLI by analyzing somatic copy number alterations (SCNA) and somatic mutation data
    - Detection of SL-partners of gene (underexpression or low copy number show essentiallity)
    - Pairwise gene coexpression SL pairs tend to participate in closely related biological processes/likely to be coexpressed)
    - Gene pair screening to find pairs, which fulfill all three conditions

    - Synthetic dosage lethality: overexpression of gene A renders gene B essential (Examination whether gene B has higher SCNA with gene A verexpressed) -> correlation?
    - DAISY returns p values denoting the significance of SLI/SDLI between two genes according to each inference strategy and all strategies together (SoF and coexpression most important)
- **Wilcoxon rank sum test: evidence of significantly higher SCNA level of gene B when gene A is deleted + graphs with correlated expression**
- **Paired t-test to compare Gene A restored and -deficient cells regarding their sensitivity towards knockdown of gene B (SL-pair) and unsensitivity towards knockdown of random genes, respectively.**
- Construction of genome-wide networks
    - genomic location: SL pairs reside on different chromosomes or at large distance on the same, wherease SDL pairs reside close to each other
- gene essentiallity is cellline-specific
- **Computation of Kaplan-Meier score to determine if counderexpression of two SL pair genes increases tumor vulnerability (KM score higher of SL pairs than randomly selected pairs)**
    - **p-test to show that the high KM score is a result of SLI and not gene essentiallity**
    - **KM-plots to show that survivability of the patient is positively correlated with the number of SL pairs counderexpressed**

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.