Mining and analysis of genomic and epigenomic data (TCGA) using R

Catharina Olsen & Antonio Colaprico
Academic supervisor: Gianluca Bontempi

Machine Learning Group (MLG)
Interuniversity Institute of Bioinformatics in Brussels (IB)^2

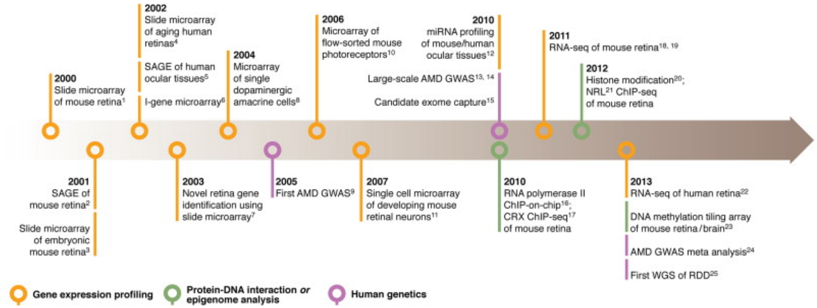
December 6th & 7th 2016

Workshop overview

- day 1
  - introduction R
  - Analyses
    - Differential expression analysis
    - Enrichment analysis
    - Clustering, dendrograms & heatmaps
    - Survival analysis
  - **data in biomedical research: NGS, TCGA, downloading and normalization**
- day 2
  - integrative analysis
  - Command line vs. graphical user interface (introduction to TCGAbiolinksGUI)

Intro to NGS data

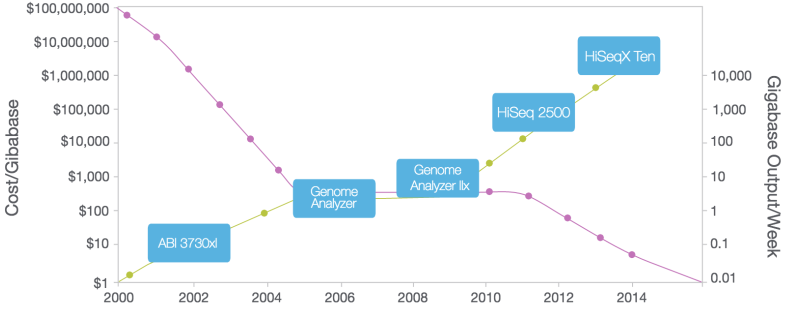
A timeline



2000 Slide microarray of mouse retina[1]
2001 SAGE of mouse retina[2]
Slide microarray of embryonic mouse retina[3]
2002 Slide microarray of aging human retinas[4]
SAGE of human ocular tissues[5]
I-gene microarray[6]
2003 Novel retina gene identification using slide microarray[7]
2004 Microarray of single dopaminergic amacrine cells[8]
2005 First AMD GWAS[9]
2006 Microarray of flow-sorted mouse photoreceptors[10]
2007 Single cell microarray of developing mouse retinal neurons[11]
2010 miRNA profiling of mouse/human ocular tissues[12]
Large-scale AMD GWAS[13, 14]
Candidate exome capture[15]
2010 RNA polymerase II ChIP-on-chip[16]; CRX ChIP-seq[17] of mouse retina
2011 RNA-seq of mouse retina[18, 19]
2012 Histone modification[20]; NRL[21] ChIP-seq of mouse retina
2013 RNA-seq of human retina[22]
DNA methylation tiling array of mouse retina/brain[23]
AMD GWAS meta analysis[24]
First WGS of RDD[25]
Gene expression profiling
Protein-DNA interaction *or* epigenome analysis
Human genetics

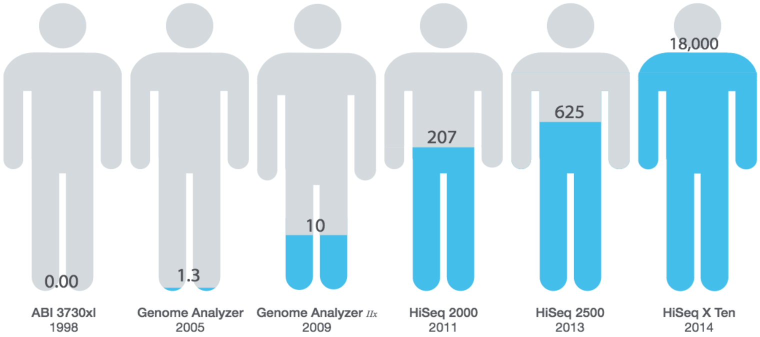
Intro to NGS data

Cost and output over time



Intro to NGS data

Human genomes sequenced anually



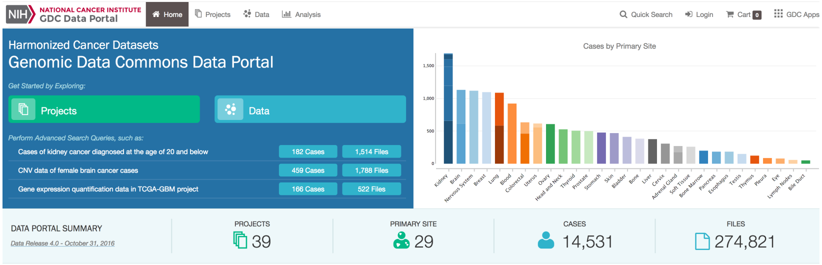
Intro to NGS data

NGS methods

- Genomics
  - whole-genome sequencing
  - exome sequencing
  - de novo sequencing
  - targeted sequencing
- Transcriptomics
  - total RNA and mRNA sequencing
  - targeted RNA sequencing
  - small RNA and noncoding RNA sequencing
- Epigenomics
  - methylation sequencing
  - ChIP sequencing
  - ribosome profiling

The Cancer Genome Atlas (TCGA)

What data is available?



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Data

Perform Advanced Search Queries, such as:

Cases of kidney cancer diagnosed at the age of 20 and below	182 Cases	1,514 Files
CNV data of female brain cancer cases	459 Cases	1,788 Files
Gene expression quantification data in TCGA-GBM project	166 Cases	522 Files

DATA PORTAL SUMMARY

Data Release 4.0 - October 31, 2016

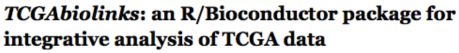
PROJECTS	PRIMARY SITE	CASES	FILES
39	29	14,531	274,821

→ TCGA data are part of this collection

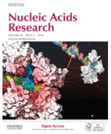
The Cancer Genome Atlas (TCGA)

How can we get this data on our computers?

- Using R



***TCGAbiolinks*: an R/Bioconductor package for integrative analysis of TCGA data**



- TCGAbiolinks was developed to
  - facilitate the TCGA open-access data retrieval,
  - prepare the data using the appropriate pre-processing strategies,
  - provide the means to carry out different standard analyses and
  - allow the user to download a specific version of the data and thus to easily reproduce earlier research results
- load TCGAbiolinks

```r
library(TCGAbiolinks)
```

The Cancer Genome Atlas (TCGA)

TCGAbiolinks – getting started

arguments that the user needs to specify:

- **project**: available cancer types such as TCGA-BRCA or TCGA-LUAD (33 different cancer types available)

- **data.category**: which type of data the user is interested in

```
TCGAbiolinks:::getProjectSummary("TCGA-BRCA")


## $data_categories
##   case_count file_count               data_category
## 1       1095       1234             DNA Methylation
## 2       1097       6080     Transcriptome Profiling
## 3       1098       1098                 Biospecimen
## 4       1044       8648 Simple Nucleotide Variation
## 5       1096       4446       Copy Number Variation
## 6       1098       4604          Raw Sequencing Data
## 7       1097       1097                    Clinical
##
## $case_count
## [1] 1098
##
## $file_count
## [1] 27207
##
## $file_size
## [1] 5.39227e+13
```

The Cancer Genome Atlas (TCGA)

TCGAbiolinks – getting started

arguments that the user needs to specify:

- **project**: available cancer types such as TCGA-BRCA or TCGA-LUAD (33 different cancer types available)
- **data.category**: which type of data the user is interested in
- **data.type**
- **Workflow Type**
- **platform** check the package's vignette for all parameter options/combinations

The Cancer Genome Atlas (TCGA)

TCGAbiolinks – Step 1

- query database to obtain list of samples to download
  → downloads information for the available samples

```
query.exp <- GDCquery(project = "TCGA-BRCA",
legacy = TRUE,
data.category = "Gene expression",
data.type = "Gene expression quantification",
platform = "Illumina HiSeq", file.type = "results",
experimental.strategy = "RNA-Seq",
sample.type = c("Primary solid Tumor","Solid Tissue Normal"))

## Accessing GDC. This might take a while...
```

- available information

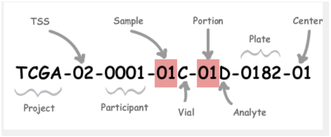
```
colnames(query.exp[[1]][[1]])

##  [1] "center"                "data_type"
##  [3] "updated_datetime"      "file_name"
##  [5] "md5sum"                "data_format"
##  [7] "acl"                   "access"
##  [9] "platform"              "state"
## [11] "state_comment"         "file_id"
## [13] "data_category"         "file_size"
## [15] "cases"                 "submitter_id"
## [17] "type"                  "tags"
## [19] "experimental_strategy" "tissue.definition"
```

The Cancer Genome Atlas (TCGA)

Sample identification

- each TCGA sample has a unique identifier: its **file_id** (A Universally Unique Identifier (UUID) is a randomly-generated, 32-digit hexadecimal value)
- historically samples were identified by their barcode



- issues with the barcode system:
  - TCGA has become more complex and the barcode structure cannot hold the data required
  - There are not enough barcode permutations to capture all the representations required.
  - Barcodes are coupled with the metadata that forms them, which becomes an issue when the metadata changes

The Cancer Genome Atlas (TCGA)

Exercise: data download

- select only 10 samples for download (downloading the full data set will take too much time $>$ 1GB)

```
samplesDown <- query.exp$results[[1]]$cases
dataSmTP <- TCGAquery_SampleTypes(barcode = samplesDown, typesample = "TP")
dataSmNT <- TCGAquery_SampleTypes(barcode = samplesDown, typesample = "NT")
mybarcode <- c(sample(dataSmTP,5), sample(dataSmNT, 5))

print(mybarcode)

##  [1] "TCGA-D8-A1XU-01A-11R-A14M-07" "TCGA-B6-A0IG-01A-11R-A034-07"
##  [3] "TCGA-A8-A06T-01A-11R-A00Z-07" "TCGA-A2-A04U-01A-11R-A115-07"
##  [5] "TCGA-AO-A12H-01A-11R-A115-07" "TCGA-BH-A209-11A-42R-A157-07"
##  [7] "TCGA-BH-A208-11A-51R-A157-07" "TCGA-BH-A0DO-11A-22R-A12D-07"
##  [9] "TCGA-BH-A0E0-11A-13R-A089-07" "TCGA-BH-A18N-11A-43R-A12D-07"
```

- regenerate the query information for the selected samples

```
queryDown <- GDCquery(project = "TCGA-BRCA",
 data.category = "Gene expression",
 data.type = "Gene expression quantification",
 platform = "Illumina HiSeq",
 file.type = "results",
 experimental.strategy = "RNA-Seq",
 sample.type = c("Primary solid Tumor","Solid Tissue Normal"),
 barcode = mybarcode,
 legacy = TRUE)

## Accessing GDC. This might take a while...
```

The Cancer Genome Atlas (TCGA)

Exercise: data download

- use `GDCdownload` function

```
  GDCdownload(queryDown, directory = "./TCGAexample")

## Of the 10 files for download 10 already exist.
## All samples have been already downloaded
```

The Cancer Genome Atlas (TCGA)

Preprocessing

- load downloaded data
- use SummarizedExperiment object for downstream analysis

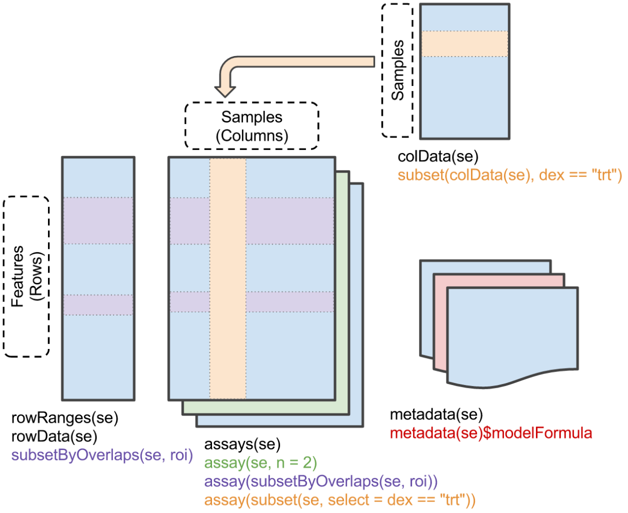
```r
library(SummarizedExperiment)

brca.exp <- GDCprepare(query = queryDown, save = TRUE,
  save.filename = "brcaExp.rda", directory =  "./TCGAexample")
```

```
##
  |
  |                                                                  |   0%
  |
  |======                                                            |  10%
  |
  |=============                                                     |  20%
  |
  |====================                                              |  30%
  |
  |==========================                                        |  40%
  |
  |================================                                  |  50%
  |
  |=======================================                           |  60%
  |
  |==============================================                    |  70%
  |
  |====================================================              |  80%
  |
  |==========================================================        |  90%
```

The Cancer Genome Atlas (TCGA)

Summarized experiment



colData(se)
subset(colData(se), dex == "trt")

rowRanges(se)
rowData(se)
subsetByOverlaps(se, roi)

assays(se)
assay(se, n = 2)
assay(subsetByOverlaps(se, roi))
assay(subset(se, select = dex == "trt"))

metadata(se)
metadata(se)$modelFormula

The Cancer Genome Atlas (TCGA)

Summarized experiment

- let's look at the data for a moment

```
  brca.exp

## class: RangedSummarizedExperiment
## dim: 20330 100
## metadata(0):
## assays(2): raw_count scaled_estimate
## rownames(20330): A1BG|1 A1CF|29974 ... ZZEF1|23140 ZZZ3|26009
## rowData names(3): gene_id entrezgene
##   transcript_id.transcript_id_TCGA-BH-A0BQ-11A-33R-A115-07
## colnames(100): TCGA-BH-A0BQ-11A-33R-A115-07
##   TCGA-AC-A5XU-01A-11R-A28M-07 ... TCGA-BH-A0H5-11A-62R-A115-07
##   TCGA-E9-A3X8-01A-31R-A22U-07
## colData names(85): sample patient ...
##   subtype_Integrated.Clusters..no.exp.
##   subtype_Integrated.Clusters..unsup.exp.
```

- retrieve the experiment data use `assays` accessor; each of the assay data sets can be accessed using the `$` operator

```
  assays(brca.exp)$raw_count[1:3,1:3]

##              TCGA-BH-A0BQ-11A-33R-A115-07 TCGA-AC-A5XU-01A-11R-A28M-07
## A1BG|1                             155.77                      1530.03
## A1CF|29974                           0.00                         0.00
## A2M|2                             90816.58                     33675.67
##              TCGA-E9-A1ND-11A-43R-A144-07
## A1BG|1                             241.00
## A1CF|29974                           1.00
## A2M|2                             80950.84
```

The Cancer Genome Atlas (TCGA)

Preprocessing

- search for possible outliers
- perform an Array Array Intensity correlation (AAIC)

```
dataPrep <- TCGAanalyze_Preprocessing(object = brca.exp, cor.cut = 0.6)
```

The Cancer Genome Atlas (TCGA)

Normalization

- Within-lane normalization procedures to adjust for GC-content effect

```
dataNorm <- TCGAanalyze_Normalization(tabDF = dataPrep,
                                      geneInfo = geneInfo,
                                      method = "gcContent")


## Warning in geneNames[, 1] == names(tmp[which(tmp > 1)]):  longer object length is
not a multiple of shorter object length
## I Need about  2.5 seconds for this Complete Normalization Upper Quantile
[Processing 80k elements /s]
## Step 1 of 4:  newSeqExpressionSet ...
## Step 2 of 4:  withinLaneNormalization ...
## Step 3 of 4:  betweenLaneNormalization ...
## Step 4 of 4:  .quantileNormalization ...

dataFilt <- TCGAanalyze_Filtering(tabDF = dataNorm,
                                  method = "quantile",
                                  qnt.cut =  0.25)
```

Questions?