AIMS-HTG-validation.R

t

2024-08-13

```
# HEADER ####
#
# Version: 2024-08-13
#
# Comparison of AIMS results in TCGA-BRCA-RNA-Seq based on
# complete gene list and the subset available in the HTG-Panel
#
#


#
# SETUP ####

Sys.setenv(lang = "en_US")
```

Install required packages if missing ————————————————–

```
# Package names from CRAN
packs <- c("dplyr", "AIMS", "crosstable")

# Install packages not yet installed
installed_packages <- packs %in% rownames(installed.packages())
if (any(installed_packages == FALSE)) {
  install.packages(packs[!installed_packages])
}


# Package names from Bioconductor
bcpacks <- c("cBioPortalData")

# Install bc-packages if not yet installed from Bioconductor
installed_packages <- bcpacks %in% rownames(installed.packages())
if (any(installed_packages == FALSE)) {
  if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
  BiocManager::install(bcpacks[!installed_packages])
}
```

Load required packages ————————————————–

```
invisible(lapply(packs, library, character.only = TRUE))

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: e1071

## Loading required package: Biobase

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:dplyr':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

## Warning: package 'crosstable' was built under R version 4.3.3
```

```
invisible(lapply(bcpacks, library, character.only = TRUE))
```

```
## Loading required package: AnVIL

## Loading required package: MultiAssayExperiment

## Loading required package: SummarizedExperiment

## Loading required package: MatrixGenerics

## Warning: package 'MatrixGenerics' was built under R version 4.3.1

## Loading required package: matrixStats

## Warning: package 'matrixStats' was built under R version 4.3.1

##
## Attaching package: 'matrixStats'

## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians
```

```
## The following object is masked from 'package:dplyr':
##
##     count

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

## The following object is masked from 'package:Biobase':
##
##     rowMedians

## Loading required package: GenomicRanges

## Loading required package: stats4

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:dplyr':
##
##     first, rename

## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

## Warning: package 'IRanges' was built under R version 4.3.1

##
## Attaching package: 'IRanges'
```

```
## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice

## The following object is masked from 'package:grDevices':
##
##     windows

## Loading required package: GenomeInfoDb

## Warning: package 'GenomeInfoDb' was built under R version 4.3.1
```

```r
# FUNCTION Definitions ####
# ************************************************
#
# Function tcgaRseqEntrez  ####
#
#
# The function tcgaRseqEntrez obtaines RNA-Seq data of a provided list
#  of entrezGeneId's from TCGA using the cBioPortal access tools.
#
# THIS IS A SLIGHT CHANGED VERSION OF FUNCTION tcgaRseqGenelist from
#  the file HTG-validation_RS-GGI-SET-Pen.R
#
# We apply the cBioPortalData package to access data from the cBIO Portal
# at www.cbioportal.org
# This will allow to download RNA-Seq data from the TCGA-BRCA cohort.

library(cBioPortalData)
# First we setup some parameters for the cBioportal-access
# Define api
cbio <- cBioPortal()
```

```
## Warning in .service_validate_md5sum(api_reference_url, api_reference_md5sum, : service
version differs from validated version
##     service url: https://www.cbioportal.org/api/v2/api-docs
##     observed md5sum: 7314de5c5e8056e4e07b411b3e5a0cb9
##     expected md5sum: 07ceb76cc5afcf54a9cf2e1a689b18f7
```

```r
# Function definition:
#  (entrezList is a vector of entrezGeneId's)

tcgaRseqEntrez <- function (entrezList) {
  # Download BRCA RNA-Seq data for this list of entrezGeneId's from cBioPortal
  #  as a "MultiAssayExperiemnt" brca_rnaseq
  brca_rnaseq <- cBioPortalData(
    api = cbio,
    studyId = "brca_tcga",
    genes = entrezList, by = "entrezGeneId",
    molecularProfileIds = "brca_tcga_rna_seq_v2_mrna"
    )
  # Extract the RNA-Seq data from the MultiAssayExperiment
  tcgaRseqGenelist <- assay(brca_rnaseq[["brca_tcga_rna_seq_v2_mrna"]])

}
```

```r
# Function tcgaRseqEntrezERpos  ####
#
## The function tcgaRseqEntrezERpos obtaines RNA-Seq data of a provided list
#  of entrezGeneId's from TCGA using the cBioPortal access tools.
#  and delivers only the data of ERpos BRCA samples.
#
# THIS IS A SLIGHT CHANGED VERSION OF FUNCTION tcgaRseqGenelistERpos from
#  the file HTG-validation_RS-GGI-SET-Pen.R
#
# We apply the cBioPortalData package to access data from the cBIO Portal
# at www.cbioportal.org
# This will allow to download RNA-Seq data from the TCGA-BRCA cohort.

library(cBioPortalData)
# First we setup some parameters for the cBioportal-access
# Define api
cbio <- cBioPortal()

# Function definition:
#  (genelist is a vector of gene symbols)

tcgaRseqEntrezERpos <- function (entrezList) {
  # Download BRCA RNA-Seq data for this genelist from cBioPortal
  #  as a "MultiAssayExperiemnt" brca_rnaseq
  brca_rnaseq <- cBioPortalData(
    api = cbio,
    studyId = "brca_tcga",
    genes = entrezList, by = "entrezGeneId",
    molecularProfileIds = "brca_tcga_rna_seq_v2_mrna"
  )
  # Extract the RNA-Seq data from the MultiAssayExperiment
  tcgaRseqGenelist <- assay(brca_rnaseq[["brca_tcga_rna_seq_v2_mrna"]])

  # Extract the phenotype data for TCGA-samples (by patientId)
  pheno <- colData(brca_rnaseq)

  # Extract the link-information between
  # the patientId ("primary") and  the RNA-seq-colnames ("colname")
  # from the MultiAssayExperiment
  sample_info <- unique(sampleMap(brca_rnaseq)[,2:3])


  # Now we use dplyr functions from tidyR to join
  #    the "ER_STATUS_BY_IHC" from pheno
  #    with the "colname" from sample_info
  # by linking the cases using the patientId == primary

  pdata <- as.data.frame(pheno) %>%
    dplyr::select(patientId, ER_STATUS_BY_IHC) %>%
    left_join(as.data.frame(sample_info), by = join_by(patientId == primary))

  # We can now use pdata to select only ER-positive samples

  pdata.erpos <- pdata %>% filter(ER_STATUS_BY_IHC == "Positive")
```

```
  tcgaRseqGenelistERpos <- tcgaRseqGenelist[, colnames(tcgaRseqGenelist)
                                              %in% pdata.erpos$colname]
}
```

```
# DATA IMPORT ####

library(dplyr)
```

Import list of genes from HTG-panel

```
htgprobes <- pull(read.table("HTG-OncBiomarkerPanel n2559-Genelist.txt",
                        header=FALSE, sep=","))


# Mapping of HTG gene names to entrez IDs, AIMS genes only
#
HTG.AIMS.gene.map <- as.data.frame(matrix(c(
  "ANXA3", "306",
  "APH1B", "83464",
  "AR", "367",
  "ASPM", "259266",
  "BCL2", "596",
  "BIRC5", "332",
  "C1orf106", "55765",
  "CA12", "771",
  "CAV1", "857",
  "CCNB2", "9133",
  "CDC20", "991",
  "CDH3", "1001",
  "CDKN1C", "1028",
  "CDKN3", "1033",
  "CENPF", "1063",
  "CEP55", "55165",
  "CIRBP", "1153",
  "CKS2", "1164",
  "CNIH4", "29097",
  "COL17A1", "1308",
  "CRYAB", "1410",
  "CSTB", "1476",
  "CX3CL1", "6376",
  "DNAJC12", "56521",
  "ERBB2", "2064",
  "ESR1", "2099",
  "FBP1", "2203",
  "FGFR4", "2264",
  "FMO5", "2330",
  "FOXA1", "3169",
  "FOXC1", "2296",
  "GAMT", "2593",
  "GATA3", "2625",
  "GFRA1", "2674",
  "GSN", "2934",
  "GSTP1", "2950",
  "HPN", "3249",
  "HSPA14", "51182",
  "ID4", "3400",
  "IGF1", "3479",
  "IGFBP6", "3489",
  "IRS1", "3667",
  "ITM2A", "9452",
```

```
  "KIF2C", "11004",
  "KIT", "3815",
  "KRT14", "3861",
  "KRT17", "3872",
  "KRT18", "3875",
  "KRT5", "3852",
  "LAMA3", "3909",
  "LYN", "4067",
  "MAD2L1", "4085",
  "MAP2K4", "6416",
  "MAPT", "4137",
  "MCM2", "4171",
  "MELK", "9833",
  "MLPH", "79083",
  "MMP7", "4316",
  "MNAT1", "4331",
  "NAT1", "9",
  "NDC80", "10403",
  "NEK2", "4751",
  "NQO1", "1728",
  "PARP1", "142",
  "PCNA", "5111",
  "PPAP2B", "8613",
  "PRC1", "9055",
  "PRKX", "5613",
  "PTN", "5764",
  "PTTG1", "9232",
  "RACGAP1", "29127",
  "RBBP8", "5932",
  "RFC4", "5984",
  "RRM2", "6241",
  "S100A8", "6279",
  "SCUBE2", "57758",
  "SERPINA3", "12",
  "SFRP1", "6422",
  "SHC2", "25759",
  "SLC39A6", "25800",
  "SPDEF", "25803",
  "STC2", "8614",
  "TFF3", "7033",
  "TK1", "7083",
  "TNFRSF21", "27242",
  "TOP2A", "7153",
  "TSPAN13", "27075",
  "TSPAN7", "7102",
  "TTK", "7272",
  "TYMS", "7298",
  "UBE2C", "11065"),
  ncol = 2L, byrow = T, dimnames = list(NULL, c("HTGname", "entrezID"))),
stringsAsFactors = F)

stopifnot(!duplicated(HTG.AIMS.gene.map$HTGname))
stopifnot(!duplicated(HTG.AIMS.gene.map$entrezID))
```

```r
library(AIMS)

# internal check: entrez IDs used by AIMS
entrezIDs.AIMS <- unique(unlist(strsplit(AIMSmodel$all.pairs, "<", T)))
stopifnot(length(entrezIDs.AIMS) == 151L)
stopifnot(HTG.AIMS.gene.map$entrezID %in% entrezIDs.AIMS)

entrezIDs.AIMS.HTG <- unique(HTG.AIMS.gene.map$entrezID)
# 91 entrez IDs
```

```r
# ANALYSIS ####


# AIMS using all entrezIDs for TCGA samples####

# Load RNA-Seq data for all entrezIDs from AIMS package

entrezList <- entrezIDs.AIMS

# RNAseq for all TCGA-BRCA (including ERneg)
tcga.AIMS.Rseq <- tcgaRseqEntrez(entrezList)

## harmonizing input:
##   removing 8 colData rownames not in sampleMap 'primary'

# RNAseq for ERpos TCGA-BRCA
tcga.AIMS.Rseq.ERpos <- tcgaRseqEntrezERpos(entrezList)

## harmonizing input:
##   removing 8 colData rownames not in sampleMap 'primary'

# ***************************************************************
#  Calculate AIMS groups including ALL 151 entrezIDs from TCGA  *
# ***************************************************************

# All TCGA samples
tcga.AIMS <- applyAIMS(tcga.AIMS.Rseq, rownames(tcga.AIMS.Rseq))

## Current k = 20

# ERpos subset
tcga.AIMS.ERpos <- applyAIMS(tcga.AIMS.Rseq.ERpos, rownames(tcga.AIMS.Rseq.ERpos))

## Current k = 20

# ***************************************************************
#  Calculate AIMS groups only for entrezIDs with HTG-data       *
# ***************************************************************

# All TCGA samples
tcga.AIMS.Rseq.HTG <- tcga.AIMS.Rseq[entrezIDs.AIMS.HTG,]

tcga.AIMS.HTG <- applyAIMS(tcga.AIMS.Rseq.HTG, rownames(tcga.AIMS.Rseq.HTG))

## You are missing the pair or have more than one 54463<6422 in

## You are missing the pair or have more than one 3872<79624 in

## You are missing the pair or have more than one 2568<55872 in

## You are missing the pair or have more than one 3852<51514 in

## You are missing the pair or have more than one 23650<4751 in

## You are missing the pair or have more than one 2296<8914 in

## You are missing the pair or have more than one 3861<7802 in
```

```
## You are missing the pair or have more than one 1033<5918 in
## You are missing the pair or have more than one 1466<79682 in
## You are missing the pair or have more than one 3400<6787 in
## You are missing the pair or have more than one 10055<2950 in
## You are missing the pair or have more than one 4781<9070 in
## You are missing the pair or have more than one 2674<57447 in
## You are missing the pair or have more than one 10477<9897 in
## You are missing the pair or have more than one 2886<596 in
## You are missing the pair or have more than one 10948<3861 in
## You are missing the pair or have more than one 6422<8942 in
## You are missing the pair or have more than one 4886<79152 in
## You are missing the pair or have more than one 29842<55765 in
## You are missing the pair or have more than one 4651<57447 in
## You are missing the pair or have more than one 2317<25803 in
## You are missing the pair or have more than one 3909<9914 in
## You are missing the pair or have more than one 54463<55040 in
## You are missing the pair or have more than one 1153<6382 in
## You are missing the pair or have more than one 11065<9338 in
## You are missing the pair or have more than one 54970<55165 in
## You are missing the pair or have more than one 3476<9232 in
## You are missing the pair or have more than one 6241<9620 in
## You are missing the pair or have more than one 2053<259266 in
## You are missing the pair or have more than one 18<79682 in
## You are missing the pair or have more than one 29127<90355 in
## You are missing the pair or have more than one 51295<991 in
## You are missing the pair or have more than one 57496<7298 in
## You are missing the pair or have more than one 1063<23303 in
## You are missing the pair or have more than one 5718<9 in
## You are missing the pair or have more than one 2099<55353 in
## You are missing the pair or have more than one 4137<53335 in
## You are missing the pair or have more than one 2625<3945 in
```

```
## You are missing the pair or have more than one 2617<79083 in

## You are missing the pair or have more than one 55143<90355 in

## You are missing the pair or have more than one 367<3930 in

## You are missing the pair or have more than one 2203<9933 in

## You are missing the pair or have more than one 36<5613 in

## You are missing the pair or have more than one 2568<57758 in

## You are missing the pair or have more than one 7272<7802 in

## You are missing the pair or have more than one 4067<54843 in

## You are missing the pair or have more than one 1058<79624 in

## You are missing the pair or have more than one 10403<18 in

## You are missing the pair or have more than one 3202<9833 in

## You are missing the pair or have more than one 1063<1264 in

## You are missing the pair or have more than one 3861<51571 in

## You are missing the pair or have more than one 10403<4222 in

## You are missing the pair or have more than one 1058<55084 in

## You are missing the pair or have more than one 4239<9055 in

## You are missing the pair or have more than one 1036<4085 in

## You are missing the pair or have more than one 142<7450 in

## You are missing the pair or have more than one 128553<51182 in

## You are missing the pair or have more than one 7083<72 in

## Current k = 20
```

```
# ERpos subset
tcga.AIMS.ERpos.Rseq.HTG <- tcga.AIMS.Rseq.ERpos[entrezIDs.AIMS.HTG,]

tcga.AIMS.ERpos.HTG <- applyAIMS(tcga.AIMS.ERpos.Rseq.HTG,
                                 rownames(tcga.AIMS.ERpos.Rseq.HTG))
```

```
## You are missing the pair or have more than one 54463<6422 in

## You are missing the pair or have more than one 3872<79624 in

## You are missing the pair or have more than one 2568<55872 in

## You are missing the pair or have more than one 3852<51514 in

## You are missing the pair or have more than one 23650<4751 in

## You are missing the pair or have more than one 2296<8914 in

## You are missing the pair or have more than one 3861<7802 in
```

```
## You are missing the pair or have more than one 1033<5918 in
## You are missing the pair or have more than one 1466<79682 in
## You are missing the pair or have more than one 3400<6787 in
## You are missing the pair or have more than one 10055<2950 in
## You are missing the pair or have more than one 4781<9070 in
## You are missing the pair or have more than one 2674<57447 in
## You are missing the pair or have more than one 10477<9897 in
## You are missing the pair or have more than one 2886<596 in
## You are missing the pair or have more than one 10948<3861 in
## You are missing the pair or have more than one 6422<8942 in
## You are missing the pair or have more than one 4886<79152 in
## You are missing the pair or have more than one 29842<55765 in
## You are missing the pair or have more than one 4651<57447 in
## You are missing the pair or have more than one 2317<25803 in
## You are missing the pair or have more than one 3909<9914 in
## You are missing the pair or have more than one 54463<55040 in
## You are missing the pair or have more than one 1153<6382 in
## You are missing the pair or have more than one 11065<9338 in
## You are missing the pair or have more than one 54970<55165 in
## You are missing the pair or have more than one 3476<9232 in
## You are missing the pair or have more than one 6241<9620 in
## You are missing the pair or have more than one 2053<259266 in
## You are missing the pair or have more than one 18<79682 in
## You are missing the pair or have more than one 29127<90355 in
## You are missing the pair or have more than one 51295<991 in
## You are missing the pair or have more than one 57496<7298 in
## You are missing the pair or have more than one 1063<23303 in
## You are missing the pair or have more than one 5718<9 in
## You are missing the pair or have more than one 2099<55353 in
## You are missing the pair or have more than one 4137<53335 in
## You are missing the pair or have more than one 2625<3945 in
```

```
## You are missing the pair or have more than one 2617<79083 in
## You are missing the pair or have more than one 55143<90355 in
## You are missing the pair or have more than one 367<3930 in
## You are missing the pair or have more than one 2203<9933 in
## You are missing the pair or have more than one 36<5613 in
## You are missing the pair or have more than one 2568<57758 in
## You are missing the pair or have more than one 7272<7802 in
## You are missing the pair or have more than one 4067<54843 in
## You are missing the pair or have more than one 1058<79624 in
## You are missing the pair or have more than one 10403<18 in
## You are missing the pair or have more than one 3202<9833 in
## You are missing the pair or have more than one 1063<1264 in
## You are missing the pair or have more than one 3861<51571 in
## You are missing the pair or have more than one 10403<4222 in
## You are missing the pair or have more than one 1058<55084 in
## You are missing the pair or have more than one 4239<9055 in
## You are missing the pair or have more than one 1036<4085 in
## You are missing the pair or have more than one 142<7450 in
## You are missing the pair or have more than one 128553<51182 in
## You are missing the pair or have more than one 7083<72 in
## Current k = 20
# ************************************************************************
```

```r
# COMPARE OBTAINED RESULTS FOR AIMS ####

# Finally we compare the results obtained from the entrezID set rules


# All TCGA samples

# All rules
table(tcga.AIMS$cl)
```

```
##
##  Basal   Her2   LumA   LumB Normal
##    191    100    409    327     73
```

```r
# HTG gene rules only
table(tcga.AIMS.HTG$cl)
```

```
##
##  Basal   Her2   LumA   LumB Normal
##    187    101    441    265    106
```

```r
# cross table
table(tcga.AIMS$cl, tcga.AIMS.HTG$cl)
```

```
##
##          Basal Her2 LumA LumB Normal
##   Basal    184    4    0    0      3
##   Her2       3   82    3    6      6
##   LumA       0    2  374    2     31
##   LumB       0   13   55  257      2
##   Normal     0    0    9    0     64
```

```r
# ERpos samples

# All rules
table(tcga.AIMS.ERpos$cl)
```

```
##
##  Basal   Her2   LumA   LumB Normal
##     21     44    387    304     56
```

```r
# HTG gene rules only
table(tcga.AIMS.ERpos.HTG$cl)
```

```
##
##  Basal   Her2   LumA   LumB Normal
##     19     46    412    250     85
```

```r
# cross table
table(tcga.AIMS.ERpos$cl, tcga.AIMS.ERpos.HTG$cl)
```

```
##
##          Basal Her2 LumA LumB Normal
##   Basal     19    0    0    0      2
##   Her2       0   33    3    5      3
##   LumA       0    2  353    2     30
```

```
##   LumB       0   11   48  243      2
##   Normal     0    0    8    0     48

# ************************************************************************
```

```r
# CROSSTAB with nice formatting and percentages

ct.labels = read.table(header=TRUE, text="
  name        label
  aims.all    'Subtype all genes'
  aims.all.pr 'Prob all genes'
  aims.htg    'Subtype HTG'
  aims.htg.pr 'Prob HTG'
  ")



# For all TCGA samples:
#
# Define dataframe with colnames
tcga.AIMS.ct <- as.data.frame(cbind(tcga.AIMS$cl, tcga.AIMS$prob,
                                    tcga.AIMS.HTG$cl, tcga.AIMS.HTG$prob))
colnames(tcga.AIMS.ct) = c("aims.all", "aims.all.pr", "aims.htg", "aims.htg.pr")
# Import labels in tibble
tcga.AIMS.ct <- tcga.AIMS.ct %>%
  import_labels(ct.labels, name_from="name", label_from="label") %>%
  as_tibble()
```

Cross table with column-percentages (subtype all genes) for all TCGA samples

```
crosstable(tcga.AIMS.ct, aims.all, by=aims.htg, total="both", percent_digits=1,
           percent_pattern="{n}\n{p_col}")  %>%
  as_flextable(compact = TRUE)
```

	Subtype HTG					Total
	Basal	Her2	LumA	LumB	Normal	
Subtype all genes						
Basal	184 98.4%	4 4.0%	0 0%	0 0%	3 2.8%	191 (17.4%)
Her2	3 1.6%	82 81.2%	3 0.7%	6 2.3%	6 5.7%	100 (9.1%)
LumA	0 0%	2 2.0%	374 84.8%	2 0.8%	31 29.2%	409 (37.2%)
LumB	0 0%	13 12.9%	55 12.5%	257 97.0%	2 1.9%	327 (29.7%)
Normal	0 0%	0 0%	9 2.0%	0 0%	64 60.4%	73 (6.6%)
Total	187 (17.0%)	101 (9.2%)	441 (40.1%)	265 (24.1%)	106 (9.6%)	1100 (100.0%)

Cross table with row-percentages (subtype htg) for all TCGA samples

```
crosstable(tcga.AIMS.ct, aims.all, by=aims.htg,
           total="both",
           percent_digits=1,
           percent_pattern="{n}\n{p_row}")  %>%
  as_flextable(compact = TRUE)
```

	Subtype HTG					Total
	Basal	Her2	LumA	LumB	Normal	
Subtype all genes						
Basal	184 96.3%	4 2.1%	0 0%	0 0%	3 1.6%	191 (17.4%)
Her2	3 3.0%	82 82.0%	3 3.0%	6 6.0%	6 6.0%	100 (9.1%)
LumA	0 0%	2 0.5%	374 91.4%	2 0.5%	31 7.6%	409 (37.2%)
LumB	0 0%	13 4.0%	55 16.8%	257 78.6%	2 0.6%	327 (29.7%)
Normal	0 0%	0 0%	9 12.3%	0 0%	64 87.7%	73 (6.6%)
Total	187 (17.0%)	101 (9.2%)	441 (40.1%)	265 (24.1%)	106 (9.6%)	1100 (100.0%)

```
# ***************************************************************************
# For ERpos subset of samples:

# Define dataframe with colnames
tcga.AIMS.ERpos.ct <- as.data.frame(cbind(tcga.AIMS.ERpos$cl, tcga.AIMS.ERpos$prob,
                                          tcga.AIMS.ERpos.HTG$cl, tcga.AIMS.ERpos.HTG$prob))
colnames(tcga.AIMS.ERpos.ct) = c("aims.all","aims.all.pr","aims.htg","aims.htg.pr")
# Import labels in tibble
tcga.AIMS.ERpos.ct <- tcga.AIMS.ERpos.ct %>%
  import_labels(ct.labels, name_from="name", label_from="label") %>%
  as_tibble()
```

Cross table with column-percentages (subtype all genes) for ERpos TCGA subset:

```
crosstable(tcga.AIMS.ERpos.ct, aims.all, by=aims.htg, total="both",
           percent_digits=1, percent_pattern="{n}\n{p_col}")  %>%
  as_flextable(compact = TRUE)
```

	Subtype HTG					Total
	Basal	Her2	LumA	LumB	Normal	
Subtype all genes						
Basal	19 100.0%	0 0%	0 0%	0 0%	2 2.4%	21 (2.6%)
Her2	0 0%	33 71.7%	3 0.7%	5 2.0%	3 3.5%	44 (5.4%)
LumA	0 0%	2 4.3%	353 85.7%	2 0.8%	30 35.3%	387 (47.7%)
LumB	0 0%	11 23.9%	48 11.7%	243 97.2%	2 2.4%	304 (37.4%)
Normal	0 0%	0 0%	8 1.9%	0 0%	48 56.5%	56 (6.9%)
Total	19 (2.3%)	46 (5.7%)	412 (50.7%)	250 (30.8%)	85 (10.5%)	812 (100.0%)

Cross table with row-percentages (subtype htg) for ERpos TCGA subset:

```
crosstable(tcga.AIMS.ERpos.ct, aims.all, by=aims.htg, total="both",
           percent_digits=1, percent_pattern="{n}\n{p_row}")  %>%
  as_flextable(compact = TRUE)
```

	Subtype HTG					Total
	Basal	Her2	LumA	LumB	Normal	
Subtype all genes						
Basal	19 90.5%	0 0%	0 0%	0 0%	2 9.5%	21 (2.6%)
Her2	0 0%	33 75.0%	3 6.8%	5 11.4%	3 6.8%	44 (5.4%)
LumA	0 0%	2 0.5%	353 91.2%	2 0.5%	30 7.8%	387 (47.7%)
LumB	0 0%	11 3.6%	48 15.8%	243 79.9%	2 0.7%	304 (37.4%)
Normal	0 0%	0 0%	8 14.3%	0 0%	48 85.7%	56 (6.9%)
Total	19 (2.3%)	46 (5.7%)	412 (50.7%)	250 (30.8%)	85 (10.5%)	812 (100.0%)

```
# ******************************************************************************
# SESSION INFO ####
sessionInfo()

## R version 4.3.0 (2023-04-21 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 11 x64 (build 22631)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=German_Germany.utf8  LC_CTYPE=German_Germany.utf8
## [3] LC_MONETARY=German_Germany.utf8 LC_NUMERIC=C
## [5] LC_TIME=German_Germany.utf8
##
## time zone: Europe/Berlin
## tzcode source: internal
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] cBioPortalData_2.12.0       MultiAssayExperiment_1.26.0
##  [3] SummarizedExperiment_1.30.2 GenomicRanges_1.52.0
##  [5] GenomeInfoDb_1.36.1         IRanges_2.34.1
##  [7] S4Vectors_0.38.1            MatrixGenerics_1.12.3
##  [9] matrixStats_1.0.0           AnVIL_1.12.3
## [11] crosstable_0.7.0            AIMS_1.32.0
## [13] Biobase_2.60.0              BiocGenerics_0.46.0
## [15] e1071_1.7-13                dplyr_1.1.2
##
## loaded via a namespace (and not attached):
##   [1] rstudioapi_0.15.0         jsonlite_1.8.5
##   [3] magrittr_2.0.3            GenomicFeatures_1.52.1
##   [5] rmarkdown_2.22            BiocIO_1.10.0
##   [7] zlibbioc_1.46.0           ragg_1.2.5
##   [9] vctrs_0.6.2               Rsamtools_2.16.0
##  [11] memoise_2.0.1             RCurl_1.98-1.12
##  [13] askpass_1.1               htmltools_0.5.5
##  [15] S4Arrays_1.0.5            forcats_1.0.0
##  [17] progress_1.2.2            lambda.r_1.2.4
##  [19] curl_5.0.2                htmlwidgets_1.6.2
##  [21] futile.options_1.0.1      cachem_1.0.8
##  [23] uuid_1.1-0                GenomicAlignments_1.36.0
##  [25] mime_0.12                 lifecycle_1.0.3
##  [27] pkgconfig_2.0.3           Matrix_1.6-1
##  [29] R6_2.5.1                  fastmap_1.1.1
##  [31] GenomeInfoDbData_1.2.10   shiny_1.7.4.1
##  [33] digest_0.6.31             RaggedExperiment_1.24.0
##  [35] AnnotationDbi_1.62.2      textshaping_0.3.6
##  [37] RSQLite_2.3.1             filelock_1.0.2
##  [39] RTCGAToolbox_2.30.0       fansi_1.0.4
##  [41] RJSONIO_1.3-1.8           httr_1.4.6
##  [43] abind_1.4-5               compiler_4.3.0
```

```
##  [45] proxy_0.4-27              withr_2.5.0
##  [47] bit64_4.0.5               fontquiver_0.2.1
##  [49] backports_1.4.1           BiocParallel_1.34.2
##  [51] DBI_1.1.3                 biomaRt_2.56.1
##  [53] openssl_2.0.6             rappdirs_0.3.3
##  [55] DelayedArray_0.26.7       rjson_0.2.21
##  [57] gfonts_0.2.0              tools_4.3.0
##  [59] zip_2.3.0                 httpuv_1.6.11
##  [61] glue_1.6.2                restfulr_0.0.15
##  [63] promises_1.2.0.1          grid_4.3.0
##  [65] checkmate_2.3.0           generics_0.1.3
##  [67] tzdb_0.4.0                class_7.3-21
##  [69] tidyr_1.3.0               data.table_1.14.8
##  [71] hms_1.1.3                 xml2_1.3.5
##  [73] utf8_1.2.3                XVector_0.40.0
##  [75] pillar_1.9.0              stringr_1.5.0
##  [77] RCircos_1.2.2             later_1.3.1
##  [79] BiocFileCache_2.8.0       lattice_0.21-8
##  [81] rtracklayer_1.60.0        bit_4.0.5
##  [83] tidyselect_1.2.0          fontLiberation_0.1.0
##  [85] Biostrings_2.68.1         miniUI_0.1.1.1
##  [87] knitr_1.43                fontBitstreamVera_0.1.1
##  [89] crul_1.5.0                futile.logger_1.4.3
##  [91] xfun_0.39                 DT_0.29
##  [93] stringi_1.7.12            yaml_2.3.7
##  [95] codetools_0.2-19          evaluate_0.21
##  [97] httpcode_0.3.0            officer_0.6.6
##  [99] gdtools_0.3.7             tibble_3.2.1
## [101] cli_3.6.1                 xtable_1.8-4
## [103] systemfonts_1.0.4         Rcpp_1.0.10
## [105] GenomicDataCommons_1.24.2 dbplyr_2.3.3
## [107] png_0.1-8                 XML_3.99-0.14
## [109] rapiclient_0.1.3          parallel_4.3.0
## [111] TCGAutils_1.20.2          ellipsis_0.3.2
## [113] readr_2.1.4               blob_1.2.4
## [115] prettyunits_1.1.1         bitops_1.0-7
## [117] purrr_1.0.1               crayon_1.5.2
## [119] flextable_0.9.6           rlang_1.1.1
## [121] rvest_1.0.3               KEGGREST_1.40.0
## [123] formatR_1.14
```