Sample annotation check (DCIS RNAseq data)

Acharya C.R.

May 28, 2018

The following is a documented preprocessing of RNAseq raw count data obtained from Cedars-Sinai group.

Our first step includes reading the raw count data, 'CountTable_withoutDups.csv'.

```
> dat = read.csv("/Users/ca31/Research/DCIS/Duke/RNAseq_dat/CountTable_withoutDups.csv",
     header = T, na.strings = c("", NA))
> dat$GeneID = gsub("\\..*", "", dat$GeneID)
> rownames(dat) = dat[, 1]
> dat = dat[, -1]
> dat[1:5, 1:15]
```

	E29	E30	E33	E34	E39	E40	E1	E15	E16	E2	E17	E3	E18	E4	E19
ENSG00000223972	0	0	0	1	0	0	0	1	1	1	1	0	1	0	0
ENSG00000227232	0	1	0	0	0	0	0	0	0	0	0	1	2	0	2
ENSG00000243485	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000237613	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ENSG00000268020	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0

In the above table, genes are in rows and samples in column.

We now read the phenotype data, 'SamplesTable.csv'.

```
> ## Read the phenotype data
> pheno = read.csv("/Users/ca31/Research/DCIS/Duke/RNAseq_dat/SamplesTable.csv",header=T,na.stri
> dim(pheno)
```

```
[1] 985  26

> pheno$Replicates = as.character(pheno$Replicates)
> pheno$Tissue_Type = as.character(pheno$Tissue_Type)
> pheno$ClinicalDiagnosis = as.character(pheno$ClinicalDiagnosis)
> head(pheno)

  Patient_ID2 Sample_ID Tissue_Type DNA_Sigs DNA_Sig3  DNA_Sig2 DNA_Sig1
1          P1       E29        DCIS     n.a.     n.a. NOT_DCIS2     n.a.
2          P1       E30        DCIS     n.a.     n.a.      n.a.     n.a.
3          P1       E33        DCIS     n.a.     n.a. NOT_DCIS2     n.a.
4          P1       E34        DCIS     n.a.     n.a. NOT_DCIS2     n.a.
5          P1       E39 Stroma away     n.a.     n.a.      n.a.     n.a.
6          P1       E40 Stroma away     n.a.     n.a.      n.a.     n.a.
  ER_status PR_status HER2_status Person Kit Sample_Sectioned_Date
1         +         +           -   Jian  V3              14/07/07
2         +         +           -   Jian  V3              14/07/07
3         +         +           -   Jian  V3              14/07/07
4         +         +           -   Jian  V3              14/07/07
5         +         +           -   Jian  V3              14/07/07
6         +         +           -   Jian  V3              14/07/07
  Laser_Dissected_Date Sequencing_Completed_Date Project Replicates Patient_ID
1             14/07/17                  15/04/26   10794         2a     DCIS11
2             14/07/17                  15/04/26   10794         2d     DCIS11
3             14/07/17                  15/04/26   10794         2h     DCIS11
4             14/07/17                  15/04/26   10794         2i     DCIS11
5             14/07/17                  15/04/26   10794         7a     DCIS11
6             14/07/17                  15/04/26   10794         7e     DCIS11
     Sample_ID2 SampleIDFromDuke DateOfShipment ResearchCorePathologicDiagnosis
1 Prj10794_SE29           DCIS11      8/28/2013                            DCIS
2 Prj10794_SE30           DCIS11      8/28/2013                            DCIS
3 Prj10794_SE33           DCIS11      8/28/2013                            DCIS
4 Prj10794_SE34           DCIS11      8/28/2013                            DCIS
5 Prj10794_SE39           DCIS11      8/28/2013                            DCIS
```

```
6 Prj10794_SE40           DCIS11      8/28/2013                          DCIS
  ClinicalDiagnosis                    recurrenceStatus_
1        IDC + DCIS death with breast cancer recurrence
2        IDC + DCIS death with breast cancer recurrence
3        IDC + DCIS death with breast cancer recurrence
4        IDC + DCIS death with breast cancer recurrence
5        IDC + DCIS death with breast cancer recurrence
6        IDC + DCIS death with breast cancer recurrence
  TimeToLastFollow_up_inDays_FromDateOfDiagnosis
1                                           2796
2                                           2796
3                                           2796
4                                           2796
5                                           2796
6                                           2796
  RecurrenceFreeSurvival_inDays_FromDateOfDiagnosisToDateOfRecurr
1                                                            2766
2                                                            2766
3                                                            2766
4                                                            2766
5                                                            2766
6                                                            2766
```

All the tissue types or regions are labeled in column "Tissue_Type", and the foci within each tissue type or region are labeled "Replicates".

```
> table(pheno$Tissue_Type)

      Athypical epithelium          Benign epithelium
                        61                         44
                      DCIS           DCIS (papillary)
                       445                          8
              DCIS (solid)                Hyperplasia
                        13                          1
```

```
                      IDC         Inflammatory focus
                      124                         16
       Normal epithelium Normal epithelium (lobule)
                      100                          3
 Stroma adjacent to DCIS     Stroma adjacent to IDC
                       61                          6
             Stroma away
                      103
```

The following changes were made to the phenotype annotation data –

1. Relabel all "IDC" to "IBC".
2. HER2 status of some samples were changed from "++" to "+".
3. All the upper case replicate values were transformed to a lower case.
4. Spelling error in "Athypical" epithelium changed to "Atypical" epithelium.

```
> pheno$ClinicalDiagnosis[grep("IDC$", pheno$ClinicalDiagnosis)] <- "IBC"
> pheno$ClinicalDiagnosis[which(pheno$ClinicalDiagnosis == "IDC + DCIS")] <- "IBC + DCIS"
> rownames(pheno) = pheno$Sample_ID
> pheno$ER_status = as.character(pheno$ER_status)
> pheno$PR_status = as.character(pheno$PR_status)
> pheno$HER2_status = as.character(pheno$HER2_status)
> pheno$HER2_status[pheno$HER2_status == "++"] <- "+"
> pheno$Replicates = as.character(pheno$Replicates)
> pheno$Replicates = tolower(pheno$Replicates)
```

A new sample annotation file labeled "rectified_1.csv" was used to correct some mislabeled samples in the original annotation file.

NOTES –

1. DCIS (papillary) samples in the original sample annotation were assigned numbers that do not match Joe Geradts' sample annotation (ranges from digits 1 - 9 followed by letters indicating foci).

2. The same samples were assigned Joe's annotation code value '9', which refers to category 'Other'. However, Joe indicated that these samples should be labeled a '2'.

3. Replicate label values were also changed for patients 'P49' and 'P51' to values in the rectified_1.csv text file.

4. Replicate label of patient sample 'C07' was changed to '3b'.

5. There is only one "Hyperplasia" sample. Joe suggested to change this classification to "Benign epithelium".

6. All samples labeled "DCIS (papillary)" and "DCIS (solid)" are consolidated to one type, "DCIS".

7. All samples labeled "Normal epithelium (lobule)" are re-labeled as "Normal epithelium".

Two other columns were created from the "Replicate" column – 1) a column with Joe's region code value, and 2) Foci.

```
> pheno_rect = read.csv("/Users/ca31/Research/DCIS/Duke/RNAseq_dat/rectified_1.csv",
    header = T, na.strings = c("", "NA"))
> pheno_rect$Claire_annotation = as.character(pheno_rect$Claire_annotation)
> pheno_rect = pheno_rect[!is.na(pheno_rect$Patient_ID2), ]
> pheno[grep(" \\(papillary\\)*", pheno$Tissue_Type), ]$Replicates = pheno_rect[grep(" \\(papill
    pheno_rect$Tissue_Type), ]$Claire_annotation
> pheno[grep(" \\(papillary\\)*", pheno$Tissue_Type), ]$Replicates = gsub("9",
    "2", pheno[grep(" \\(papillary\\)*", pheno$Tissue_Type),
        ]$Replicates)
> pheno[pheno$Patient_ID2 == "P49", ]$Replicates = pheno_rect[pheno_rect$Patient_ID2 ==
    "P49", ]$Claire_annotation
> pheno[pheno$Patient_ID2 == "P51", ]$Replicates[-c(1:8)] = pheno_rect[pheno_rect$Patient_ID2 ==
    "P51", ]$Claire_annotation[-c(1:8)]
> pheno[grep("C07", pheno$Sample_ID), ]$Replicates <- "3b"
> pheno$Tissue_Type = gsub("Athypical epithelium", "Atypical epithelium",
    pheno$Tissue_Type)
> pheno$Tissue_Type = gsub("Hyperplasia", "Benign epithelium",
    pheno$Tissue_Type)
```

```
> pheno$Tissue_Type = gsub(" \\(papillary\\)*", "", pheno$Tissue_Type)
> pheno$Tissue_Type = gsub(" \\(solid\\)*", "", pheno$Tissue_Type)
> pheno$Tissue_Type = gsub(" \\(lobule\\)*", "", pheno$Tissue_Type)
> pheno$Tissue_Type = as.factor(as.character(pheno$Tissue_Type))
> pheno$Foci = tolower(substring(pheno$Replicates, 2, 3))
> pheno$Region = substring(pheno$Replicates, 1, 1)
```

Confirm Joe's code value assignments to different regions or tissue types.

	1	2	3	4	5	6	7	8	9
Atypical epithelium	0	0	0	0	0	0	0	0	61
Benign epithelium	0	0	45	0	0	0	0	0	0
DCIS	0	466	0	0	0	0	0	0	0
IDC	124	0	0	0	0	0	0	0	0
Inflammatory focus	0	0	0	0	0	0	0	16	0
Normal epithelium	0	0	0	103	0	0	0	0	0
Stroma adjacent to DCIS	0	0	0	0	0	61	0	0	0
Stroma adjacent to IDC	0	0	0	0	6	0	0	0	0
Stroma away	0	0	0	0	0	0	103	0	0