QC for psoriasis data

Elena

November, 2018

Contents


```
## Load R packages and functions:

library(data.table)
library(biomaRt)
library(tidyr)
library(ggplot2)
library(rmarkdown)
library(parallel)
source('/home/ev250/Cincinatti/Functions/various.R')
source('/home/ev250/Bayesian_inf/trecase/Functions/real.data.R')
```

STAR alignment

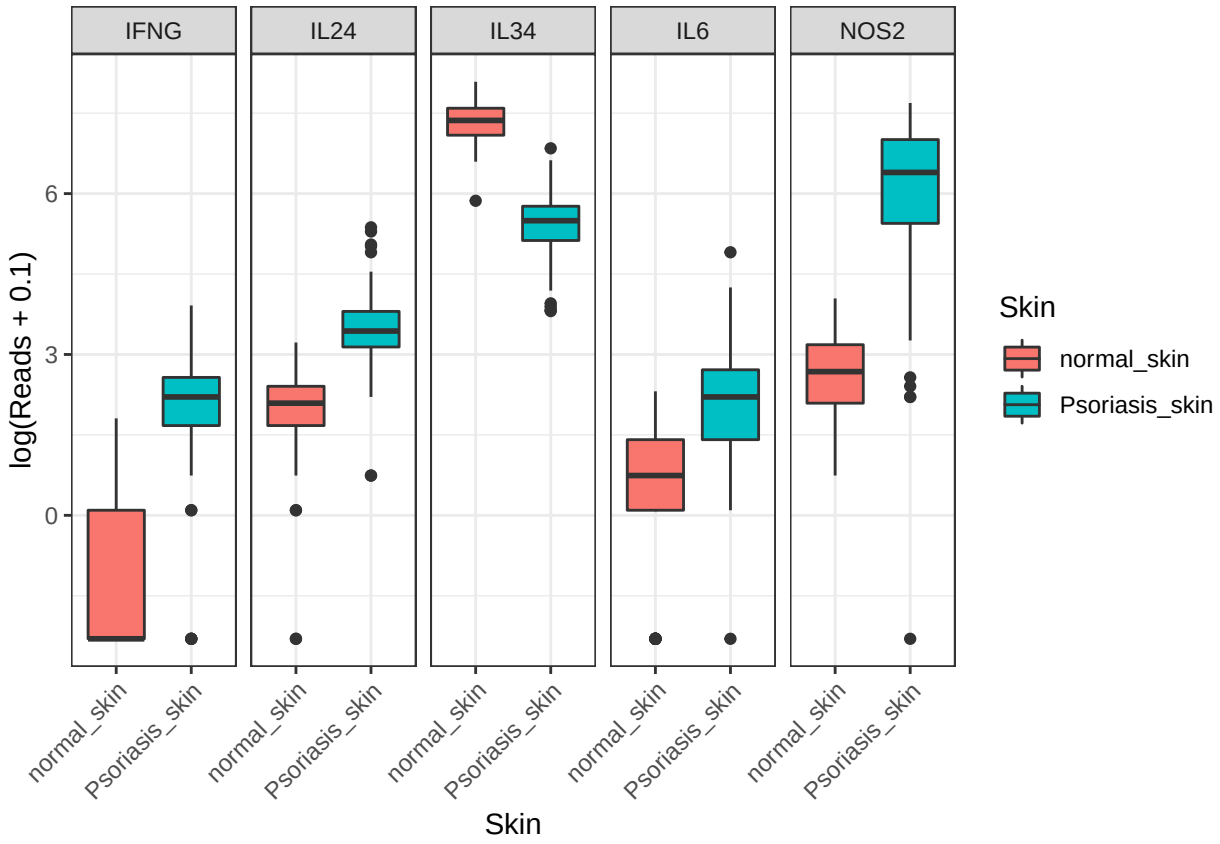
```
dirs <- list.dirs("/mrc-bsu/scratch/ev250/psoriasis/STAR")
files <- unlist(lapply(dirs, list.files, pattern="Log.final.out",
                       full.names=T))
star <- lapply(files, fread, fill=TRUE, sep="|", header=F)
star <- rbindlist(lapply(star, function(i)
    i[c(6,9:10),][, V2:= as.numeric(gsub( "[^0-9.]+", "", V2))]))

starw <- as.data.table(matrix(star$V2,ncol=length(unique(star$V1)),byrow=T))
names(starw) <- star$V1[seq_along(unique(star$V1))]
sum.star <- apply(starw,2,summary)
colnames(sum.star) <- c("Total reads", "Uniq mapped reads", "Uniq map reads (%)")

sum.star
```

```
##         Total reads Uniq mapped reads Uniq map reads (%)
## Min.        9605661           8008930           65.88000
## 1st Qu.    33757620          28012614           78.98000
## Median     39356800          31342660           81.30500
## Mean       37668890          30302756           80.82557
## 3rd Qu.    43321236          34802104           83.86750
## Max.       59247708          42685644           88.16000
```

Gene expression between normal and psoriatic skin

```
files=list.files("/mrc-bsu/scratch/ev250/psoriasis/Btrecase/inputs/Counts",
                 pattern=".txt", full.names=T)
```

```
names(files) <- sapply(files, function(i) gsub(".txt","",basename(i)))
gexp <- lapply(files, fread)

## transform to long format and merge skin type
gexp <- rbindlist(lapply(gexp, function(i) {
    mat <- as.matrix(i[,2:ncol(i)])
    dt <- data.table(gene_id=rep(i[['gene_id']], ncol(mat)),
                     Reads=as.numeric(mat))
    return(dt)
    }), idcol="Skin")


## DRG according to Li et al, 2014, 134(7):1828-1838.
DRG <- c("IFNG", "NOS2", "IL6", "IL24", "IL34")

## get ENS id
ensembl=useMart("ensembl",dataset="hsapiens_gene_ensembl")
ens  <- getBM(attributes=c("ensembl_gene_id","external_gene_name" ),
              filters="external_gene_name",
              values=DRG,
              mart=ensembl)

## Plot selected genes
gexpSub <- gexp[gene_id %in% ens$ensembl_gene_id,]
gexpSub <- merge(gexpSub, ens, by.x="gene_id", by.y="ensembl_gene_id")
bp <- ggplot(gexpSub, aes(x=Skin, y=log(Reads + 0.1), group=Skin)) +
    geom_boxplot(aes(fill=Skin)) + theme_bw() + theme(axis.text.x = element_text(angle = 45, hjust = 1))

bp + facet_grid(. ~ external_gene_name)
```



Variants called by RNA-seq

- Variants called by RNA with matching POSITION, REF and ALT allele in the reference panel
- Before DP filtering

```
files <- list.files("/mrc-bsu/scratch/ev250/psoriasis/call_vars/RPvar", pattern="chr[0-9]+_Q20_filtRP.t

rna <- lapply(files, name)

## simplify annotations in rna
rna <- rbindlist(mclapply(rna, ann_vcf, mc.cores = parallel::detectCores()))

## get proportion of samples with missing values "./." per snp
cols.gt <- grep("_GT",names(rna))
rna[, miss.p:= apply(rna, 1, function(i)sum(i=="./.")/length(cols.gt))]

## look by annotation
miss.ann <- rna[, summary(miss.p), ANN]
rna.mat <- matrix(miss.ann[['V1']], ncol=6, byrow=T,
                  dimnames=list(unique(miss.ann$ANN),
                                c("Min", "1st Qu", "Median"," Mean","3rd Qu",    "Max")))
rna.mat <- cbind(rna.mat, rna[,.N,ANN][order(match(ANN,rownames(rna.mat))),][['N']])
colnames(rna.mat)[7] <- "N variants"
rna.mat
```

```
##            Min    1st Qu     Median       Mean    3rd Qu       Max
```

```
## proximal     0 0.6590909 0.87500000 0.7562453 0.9545455 0.9943182
## UTR          0 0.0000000 0.02840909 0.1669518 0.2102273 0.9943182
## exonic       0 0.0000000 0.01704545 0.1514043 0.1590909 0.9943182
## intronic     0 0.8295455 0.92613636 0.8606285 0.9659091 0.9943182
## intergenic   0 0.9090909 0.96022727 0.9157291 0.9772727 0.9943182
## intragenic   0 0.8806818 0.94886364 0.8836832 0.9715909 0.9943182
##            N variants
## proximal       278951
## UTR             56659
## exonic          53618
## intronic       265449
## intergenic      35680
## intragenic      17170
```

- After applying DP=10 filter per SNP per samples

```
## apply filter and update
for(i in cols.gt){
    rna[get(names(rna)[i+1]) < 10, names(rna)[i] := "./."]
}
rna[, miss.p:= apply(rna, 1, function(i)sum(i=="./.")/length(cols.gt))]
miss.ann <- rna[, summary(miss.p), ANN]
rna.mat <- matrix(miss.ann[['V1']], ncol=6, byrow=T,
                  dimnames=list(unique(miss.ann$ANN),
                                c("Min", "1st Qu", "Median"," Mean","3rd Qu",    "Max")))
rna.mat <- cbind(rna.mat, rna[miss.p < 1,.N,ANN][order(match(ANN,rownames(rna.mat))),][['N']])
colnames(rna.mat)[7] <- "N variants"
rna.mat
```

```
##            Min     1st Qu    Median      Mean    3rd Qu Max N variants
## proximal     0 1.00000000 1.0000000 0.9716650 1.0000000   1      33851
## UTR          0 0.11931818 0.5909091 0.5498977 0.9886364   1      45727
## exonic       0 0.07954545 0.4204545 0.4827614 0.9375000   1      45717
## intronic     0 1.00000000 1.0000000 0.9961904 1.0000000   1       9330
## intergenic   0 1.00000000 1.0000000 0.9982512 1.0000000   1        827
## intragenic   0 1.00000000 1.0000000 0.9940187 1.0000000   1        703
```