ATACseq Analysis

Step 1 – QC & Alignment

Shell Environment:

The original “.bam” files were first converted into fastq files for fastQC analysis:

* + Converting from bam to fastq files (1\_Convert\_BAMS\_TO\_FASTQS.sh)
  + Running Fastqc (2\_FASTQC\_Script.sh)
  + Trimming using Trimmomatic (3\_Trimmomatic\_Commands\_Generator.sh) by running:
    1. bash 3\_Trimmomatic\_Commands\_Generator.sh > ./Trimmed\_Fastqs/3\_Trimmomatic\_Commands.txt
    2. bash 3\_Trimmomatic\_Commands.txt
  + Fastqc on the trimmed fastq files, P files, not U files (2\_FASTQC\_Script.sh).
  + Fixed pairing using repair.sh command from BBMap (4\_REPAIR\_Fastqs.sh)
  + Genome download (I downloaded release 109):
    1. wget http://ftp.ensembl.org/pub/release-112/fasta/homo\_sapiens/dna/Homo\_sapiens.GRCh38.dna.primary\_assembly.fa.gz
    2. wget <http://ftp.ensembl.org/pub/release-112/gtf/homo_sapiens/Homo_sapiens.GRCh38.109.gtf.gz>
    3. bwa index -p Hg38 Homo\_sapiens.GRCh38.dna.primary\_assembly.fa
  + Mapped the fixed fastqs to hg38 genome (/home/kfirinbal/ATAC-seq\_CovVac/Reference\_Annotations/) using **bwa-mem** tool (5\_BWA\_MEM\_Mapping.sh)
  + After mapping, no duplicates were found, so removing duplicates step was skipped.
  + Removed Mitochondrial Chromosomes using **Samtools** (6\_NoMT\_Subset\_Commands\_Generator.sh) by running:
    1. bash 6\_NoMT\_Subset\_Commands\_Generator.sh > 6\_NoMT\_Subset\_Commands.txt
    2. bash 6\_NoMT\_Subset\_Commands.txt
  + Removed blacklist regions (7\_RemoveBlackListRegions.sh), obtained the blacklist regions from:

obtain hg38-blacklist.v2.bed.gz from <https://github.com/Boyle-Lab/Blacklist/blob/master/lists/hg38-blacklist.v2.bed.gz>

* + Applied shifting using alignment sieve (8\_AlignmentSieve.sh).
  + Calculating genome size:

seqkit fx2tab -nl BSF\_1580\_HTNFCDMXY\_1#G27CD8\_S145935.fastq.gz ==> Read Length = 51

effective genome size: https://deeptools.readthedocs.io/en/latest/content/feature/effectiveGenomeSize.html

Human effective Genome size with read length 51 is **2701495711.**

Step 2 – Peak Calling

Peak calling was performed using macs2 (Macs2\_Generic\_Script.sh)

Also, created a metadata file to split the files into groups of: CD4\_Good, CD4\_Poor, CD8\_Good and CD8\_Poor, based on the general metadata file of the experiment (2024-04-26 Sample info ATAC.xlsx). The metadata file is called Sample\_Sheet\_All.csv, and was prepared for further analysis in R.

Step 3 - TOBIAS

* I created folders for the groups using the following script - 1\_mv\_to\_groups\_folders.sh.
* In the folder /home/kfirinbal/ATAC-seq\_CovVac/Reference\_Annotations/ I ran the following:

1. bgzip -c Homo\_sapiens.GRCh38.dna.primary\_assembly.fa > Homo\_sapiens.GRCh38.dna.primary\_assembly.fa.gz
2. samtools faidx Homo\_sapiens.GRCh38.dna.primary\_assembly.fa.gz
3. bgzip -i Homo\_sapiens.GRCh38.dna.primary\_assembly.fa.gz

* For each group, I merged the .narrowPeak output files of MACS2 into one file that represents the intersected regions between all the samples in the group:  
  cat ${PEAKS\_DIR}/\*.narrowPeak | sort -k1,1 -k2,2n | bedtools merge -i - > ${PEAKS\_DIR}/merged\_peaks.narrowPeak

(MergeNarrowPeaksFilesPerGroup.sh)

Then ran the procedural commands of TOBIAS:

* ATACorrect - nohup bash 2\_TOBIAS\_ATACorrect.sh > TOBIAS\_ATACorrect.log 2>&1 &
* ScoreBigWig - nohup bash 3\_TOBIAS\_ScoreBigwig.sh > TOBIAS\_ScoreBigWig.log 2>&1 &
* BINDetect - nohup bash 4\_TOBIAS\_BINDetect.sh > TOBIAS\_BINDetect.log 2>&1 &

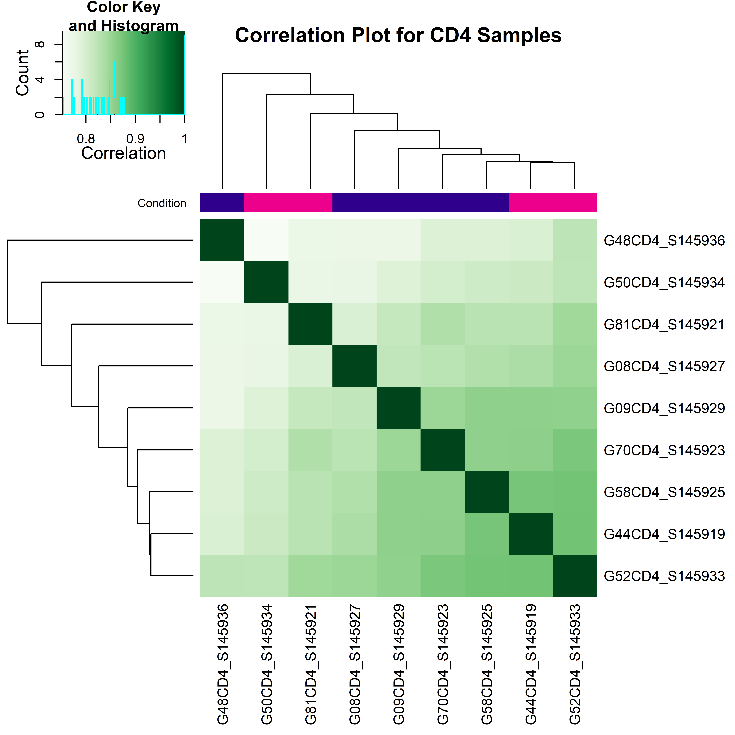
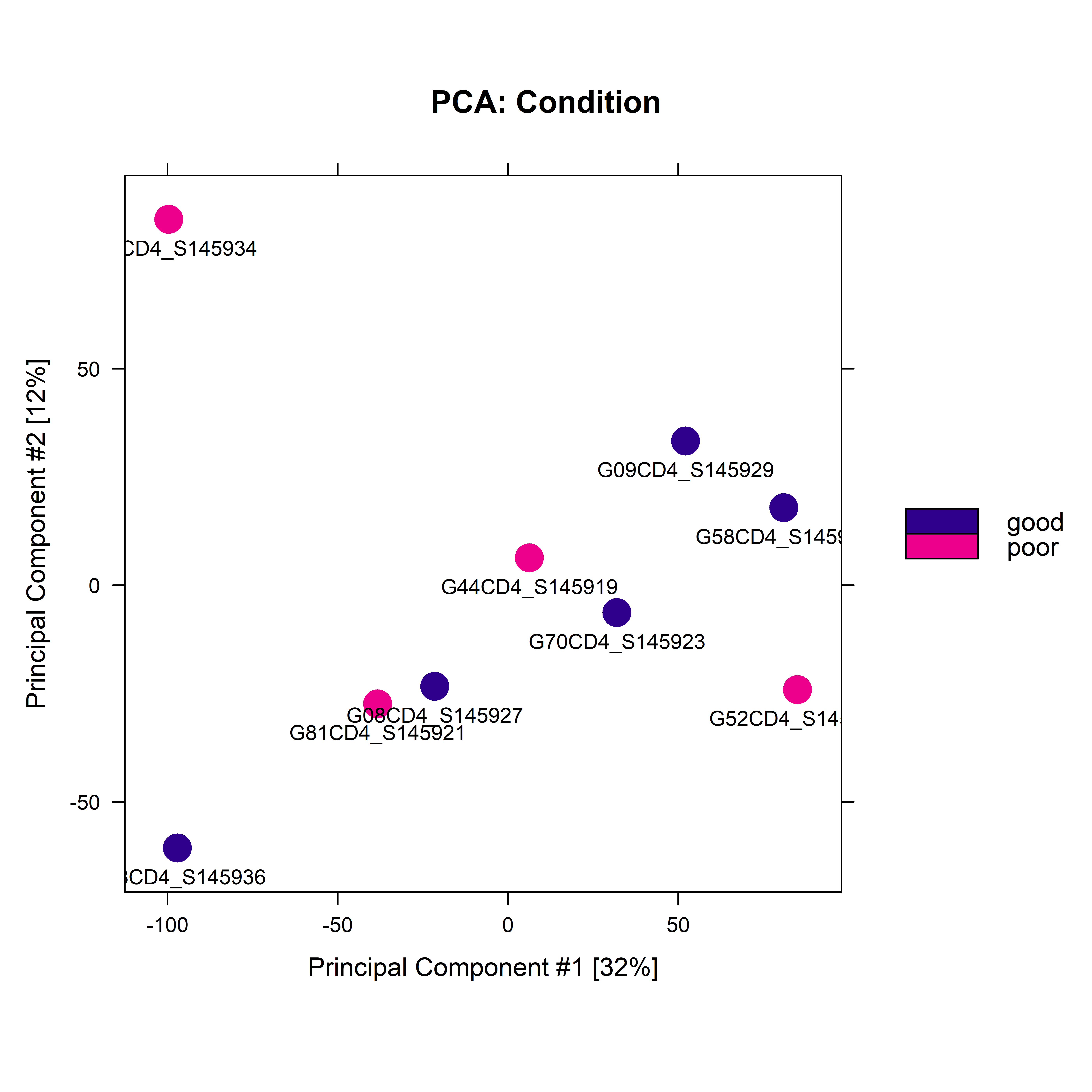
Used the MOTIFS database of JASPAR: JASPAR2024\_CORE\_non-redundant\_pfms\_jaspar.txt (<https://jaspar.elixir.no/download/data/2024/CORE/JASPAR2024_CORE_non-redundant_pfms_jaspar.txt> )

Step 4 – R DiffBind

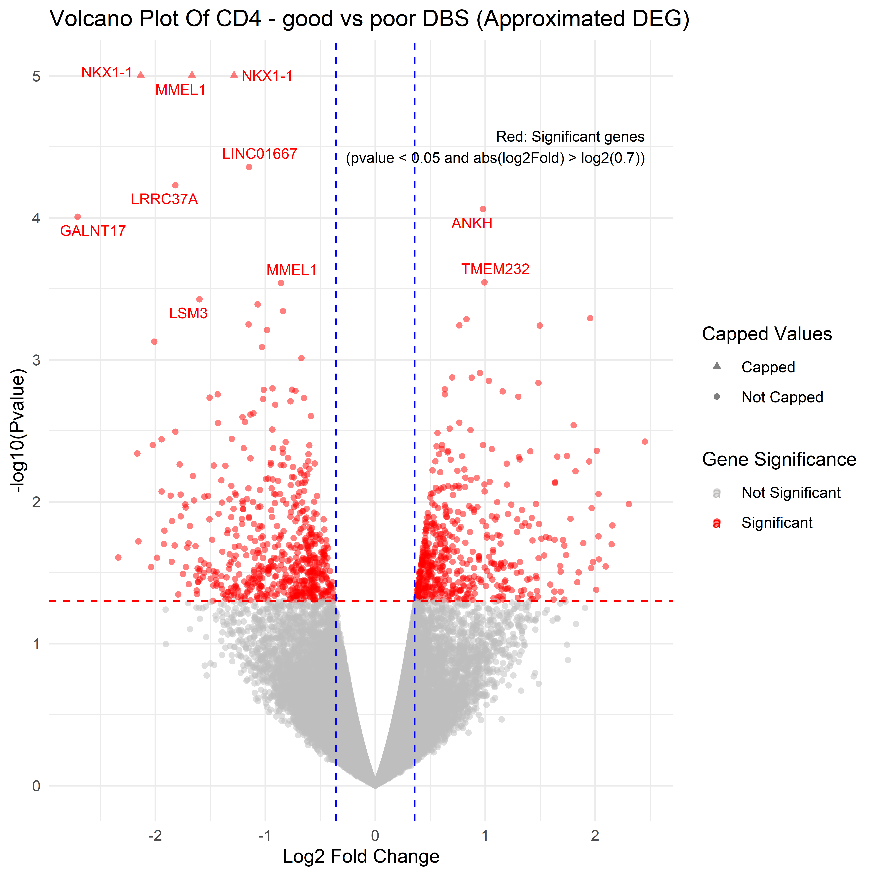
I downloaded the peak files (.xls, .bed, .narrowPeak) and the corresponding NoMT\_subset.blacklist-filtered.shifted.sorted bam files to a folder on my local computer. Then, created a R script to further analyze the data using DiffBind.

Using the NoMT\_subset.blacklist-filtered.shifted.sorted bam files and peak files from MACs2, I created a sample sheet metadata for each experiment (Sample\_Sheet\_CD4.csv and Sample\_Sheet\_CD8.csv) as an input for “dba” function to compare between the Good and Poor groups from each experiment. Then ran dba.blacklist, dba.count and dba.normalize. The chosen normalization method was TMM. Performed dba.contrast to compare between Good and Poor for either CD4 or CD8 data. At last, ran dba.analyze using EdgeR method. To generate a correlation matrix of the samples’ counts, I used dba.plotHeatmap. For PCA plot I used dba.plotPCA with the parameters DBA\_CONDITION, label=DBA\_ID. Then I applied dba.report to analyse Differential Binding Sites (DBS), setting to use p values instead of FDRs (Because there were little amount of sites that their FDR was below 0.01) and with the threshold of 0.05. To get all DBS, with no regarding to significance, I used threshold of 1. To that I added gene names with correspondence to the start and end of each site. To generate volcano plots I used the object of all possible DBS, and defined significant genes as ones that have a p value below 0.05 and logFC greater than -log(0.7) or less than log(0.7). Genes with -log10(pval) greater than 5 were considered capped and their position in the plot was placed at the top of y=5. On the volcano plot the names of the top ten most significant genes are shown.  
For pathway enrichment analysis, I used fgsea package in R. The gene list included all the genes from all possible DBS. The chosen pathways database is the C7 immunologic signature gene sets, downloaded directly from MSigDB, c7.all.v2023.2.Hs.symbols.gmt. Important set parameters in the fgsea function: minSize = 15, maxSize = 1000, eps = 0. Then I split the result table into CD4 related pathways and CD8 related pathways, and manually removed some pathways that were not really related to either. Each pathways subset I sorted by NES values column. Then for each pathways subset, I took only the top 30 pathways and defined them as Up regulated or Down regulated based on the sign of their NES value. Lastly I plotted a point plot showing the top 30 pathways based on their NES values, and each point’s size depicting the amount of genes existing in each pathway.

CD4 Experiment – Good vs Poor

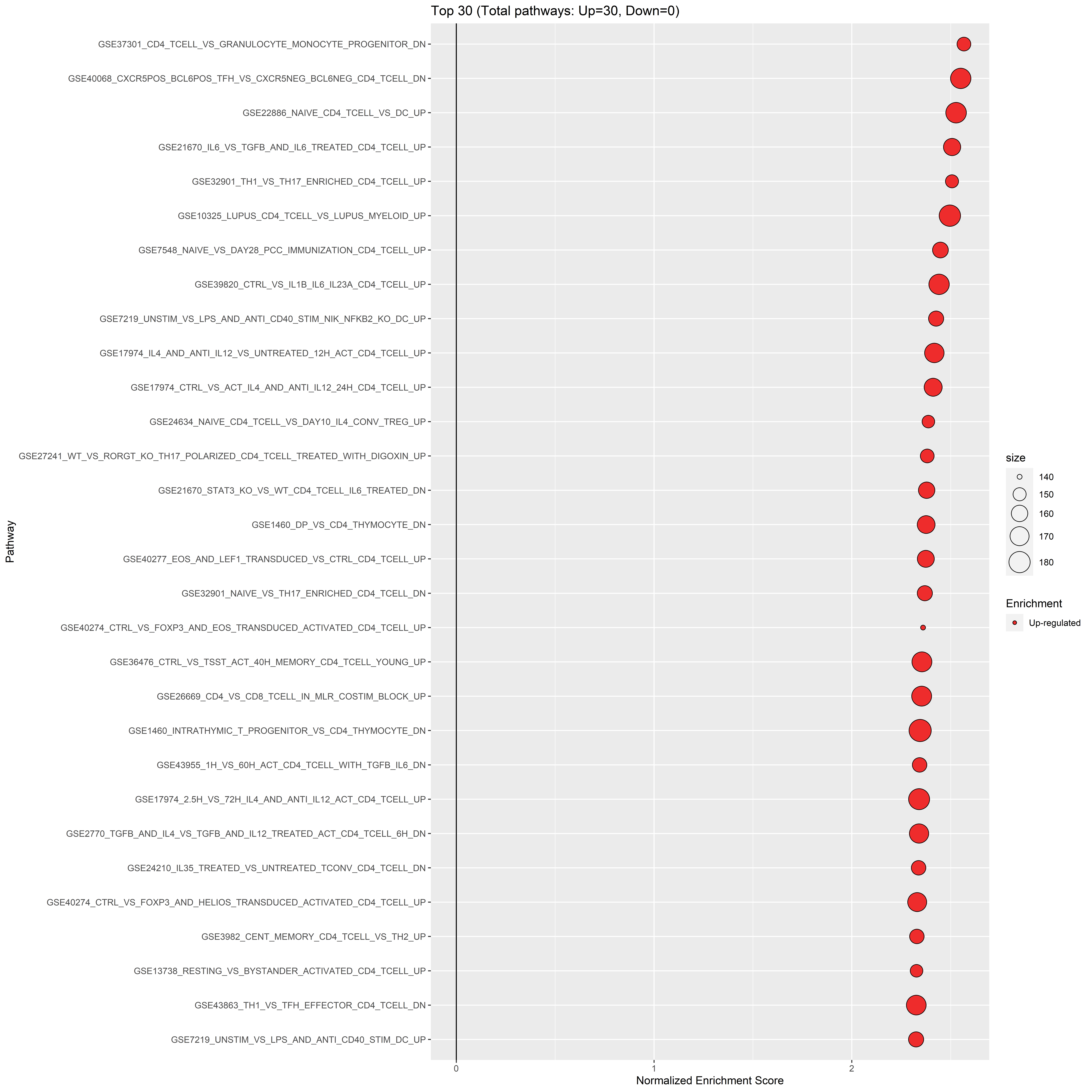
Correlation heatmap plot and PCA plot:

Volcano plot:

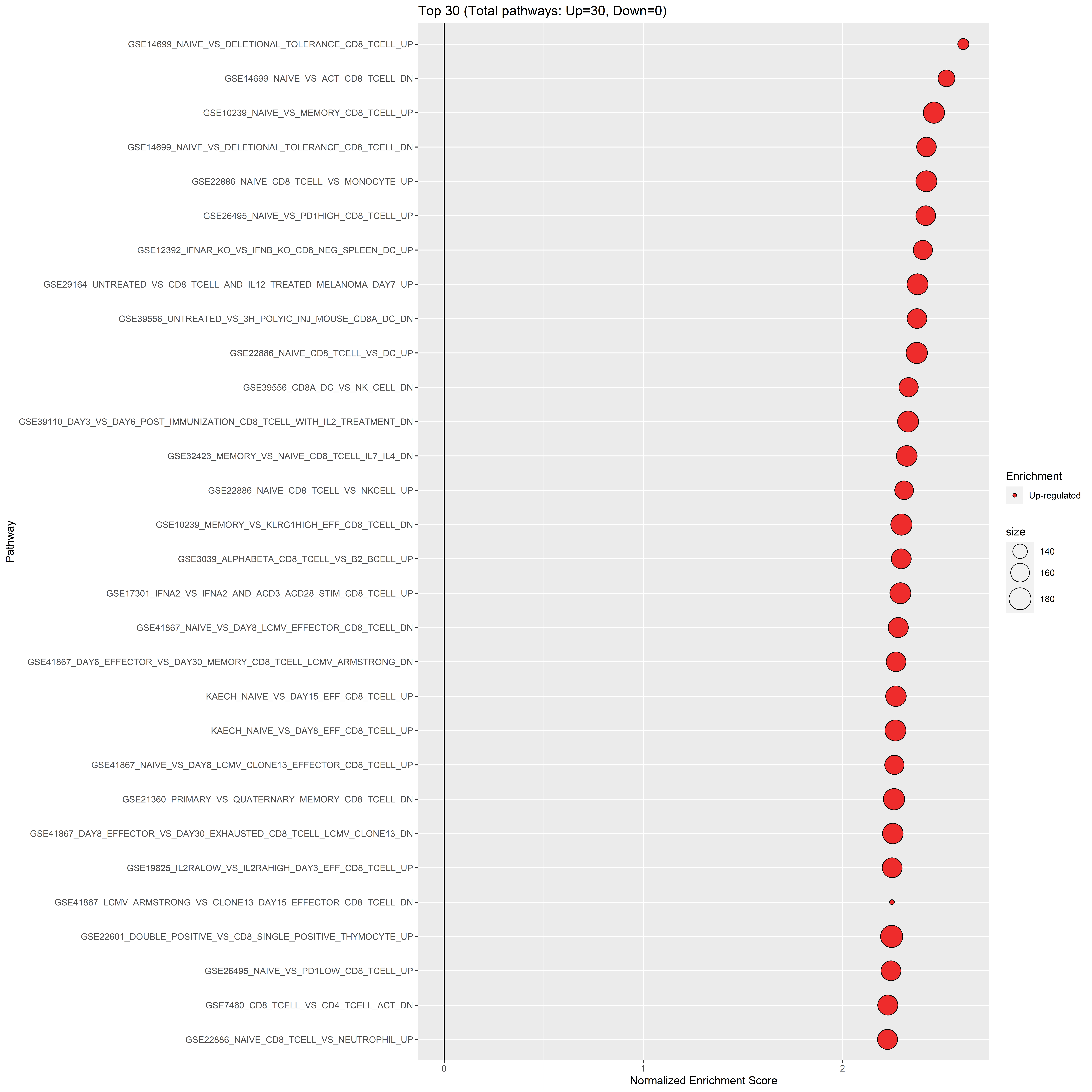


GSEA plots:

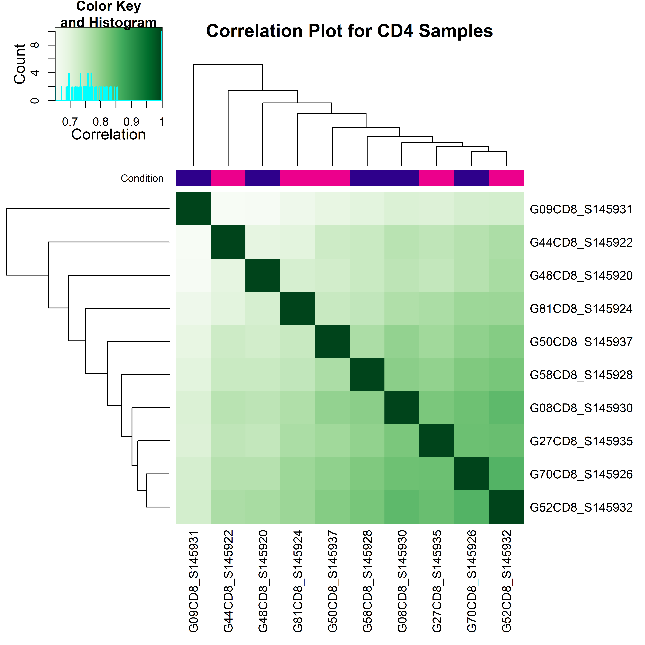
* CD4 related pathways:

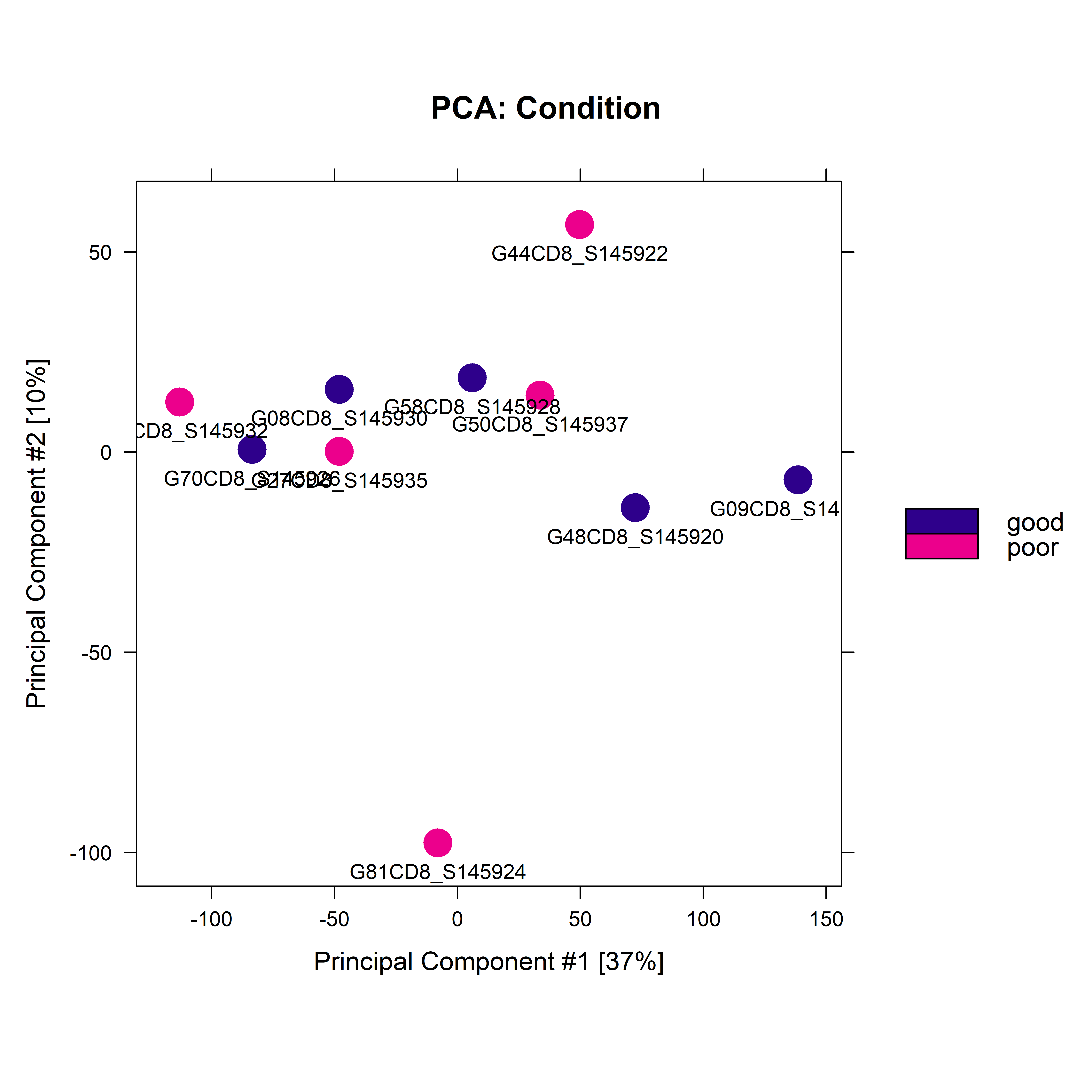


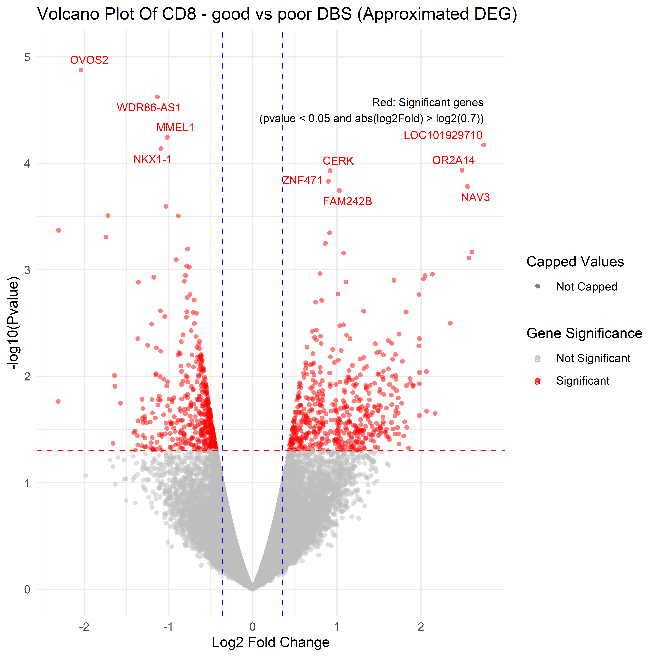
* CD8 related pathways:



CD8 Experiment – Good vs Poor

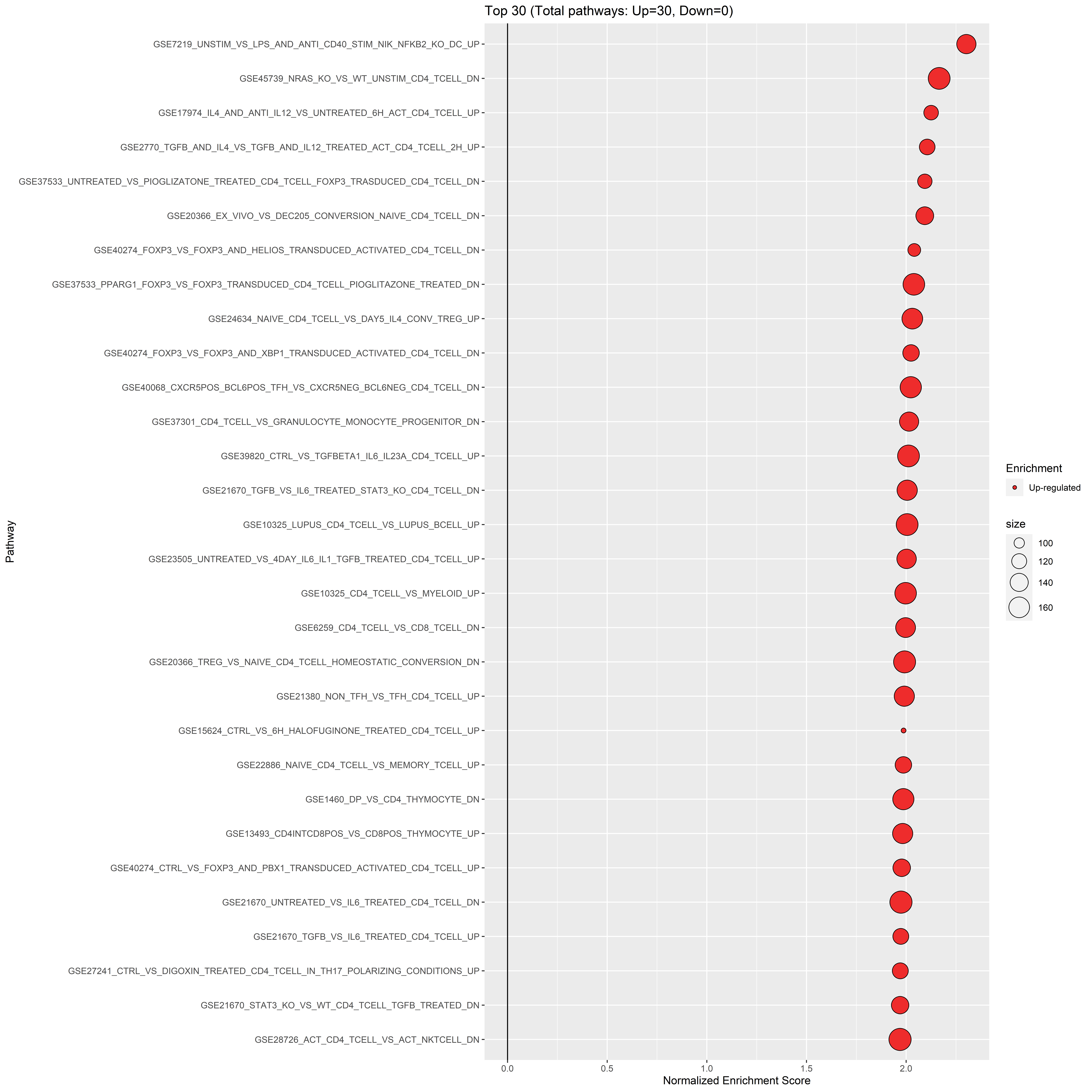
Correlation heatmap plot and PCA plot:



Volcano plot:

GSEA plots:

* CD4 related pathways:



* CD8 related pathways:

