Introduction

The RNASeq have been applied to study human cancer, giving the branch for sub typing the cancer subtypes, through the molecular subtyping tree from the analysis, we can gain the information for the cancer progression and diagnosis, in advance, aim to the subtype of the gene expression clustering, we can design the optimized cancer treatment for specific subtype, and lead to more successful way to treat. it also begun to discover the molecular heterogeneity of cancer, and to be more integrative to analyze cancer transcriptome data combine with other data sources will give more information for digging into the possible way to dissect the epigenetic event for design the treatment. In this project, it will discuss in analytical approaches, including compare the subtype analyze from the RNASeq data compare with microarray subtypes, transcriptional network analysis.

Materials and Methods

Obtained the raw RNASeq data from 1208 breast cancer specimens, then convert the bam file into fasta file for analysis. Splice junction detection and alignment by STAR2, then perform post-alignment QC by RNA-SeQC. The result is the aligned bam file, it need to perform quality filtering by Samtools, and count the fragment by HT-Seq after aligned to the GRCH38 ("human Genome Reference Consortium version 38) with STAR. Then we will generate the "Manifest" file for input to GDC Data transfer tool, download Manifest(cut –f1 gdc_manifest_20161118_234400_htseq_only.txt > uuids_htseq_only.txt), then use grep to capture htseq row include header. Execute GDC Data Transfer Tool command with Manifest file as argument chmod +x <executable>gdc-client download -m gdc_manifest_20161118_234400_htseq_only.txt and download sample meta-data by cut –f1 gdc_manifest_20161118_234400_htseq_only.txt > uuids_htseq_only.txt # and remove header, then perl reformat_ids.pl uuids_htseq_only.txt > Payload.txt. And we need to reformat data to a JSON query request by command curl --request POST --header "Content-Type: application/json" --data @Payload.txt 'https://gdc-api.nci.nih.gov/files' > File_metadata.11.23.16.txt. For convert the gene identifier from ensembl_gene_id to hgnc_symbol and subtype the pam50 gene set, we need to call the biomart package. The command is execute as follows:

```
library(biomaRt)

ensembl <- useEnsembl(biomart="ensembl", dataset="hsapiens_gene_ensembl")
head(listAttributes(ensembl))

crossref.df <- getBM(attributes=c('ensembl_gene_id','hgnc_symbol'),mart=ensembl)
write.table(crossref.pam50,file='ensembl_hgnc.pam50.txt',row.names=F,sep="\t",quote=F)
```

Then we will link those pam50 only ensembl_gene_id to our data set. The subtype define by Microarray is available for 51 of the samples in the set of 119.
We will compare the subtype of microarray with the subtype tree by our RNASeq data. 11 normal part solid tissue from TCGA database is downloaded for relating the clusters in our analysis to the status of the samples as normal and their predicted subtypes. After process the raw data, we need to identified the previous reported cancer subtype: normal, luminal A, luminal B, basal or HER2+ by the normalized heat map drawing. Table the subtype to visualize the correlation of the subtypes and the clustering. To process the htseq files into DESeq dataset, use function DESeqDataSetFromHTSeqCount, and conduct log transformation by rlogTransformation for desirable cluster expression data, then use assay() to get the matrix from DESeq dataset. Subset the data for only pam50 dataset by convert row names from Ensemble ids to hence symbols by the commands as follow:

```
library(data.table)

d <- setDT(d,keep.rownames = T)[]

a<-merge(d,pam50,by.x=‘rn',by.y='ensembl_gene_id')
```

Downloading library gplots for making the heat map by the command as follows:
Create metric x from a for heat map, then make the heat map as follows:

```
library(gplots)

heatmap.2(x, col=redgreen(100),Rowv = TRUE, Colv = T, scale="none",
trace=“none”,margin=c(10,6))
```

(Fig 2)

```
heatmap.2(x-median(x),col=redgreen(100),Rowv = TRUE, Colv = T, scale="none",
trace=“none”,margin=c(10,6))
```

(Fig 3)

We also need to make the histogram of the raw counts per library which is an important guide to check how well are those samples can be expected to separate breast cancer subtypes. The total number of reads mapped to all genes in the GRCH38 reference are sued and show in the figure 1.

The last step is to download the sigclust2 to see the clustering of those cancer subtypes quality by the p-value shows at the node of each separation. The command is perform as follows:

```
library(sigclust2)

shc.obj <- shc(t(x), metric="euclidean", linkage=“ward.D2”)
```

```
plot(shc.obj)
```

the figure is corrected by FWER, so the p-values showed on the nodes are below 0.05 cutoff(Fig 4).

Result and Discussion

To observe the cluster from normalized heat map (Fig 3) and compare with the matching microarray subtypes, it shows LumA is most close to Normal solid tumor, basal is most distal from all other types, which also includes HER2 in the same cluster. After LumA, LumB is the most related subtype to LumA which also include one HER2 and some LumA subtypes nested in the main LumB cluster. This result may lead to potentially different cancer subtype according to the the section of LamA, LamB and HER2 are quite disperse. In the HER2 subtype, this group got highest expression for ERBB2, which are important oncogene in tumorigenesis pathway. For the ER pathway, ERS1 got high expression in most branches, but in HER2 and BASAL subtypes, it expressed neutral. In progesterone receptor gene(PGR), HER2 and basal subtypes got lower expression, while some LumA and LumB subtypes have hight expressions. HER2 subtype p value for branching got 2.43e-36, which is a good quality for sub typing. Also all other subtypes are in a good quality control(p-value) for branching as a subtype.

The result branching to basal and HER2 from the microarray sub typing are not very differnt, the other subtypes could use the most highly expression gene in pam50 gene set to catalog a more organized cancer subtype for correlated oncogene and coordinated to design the treatment for clinical trend.

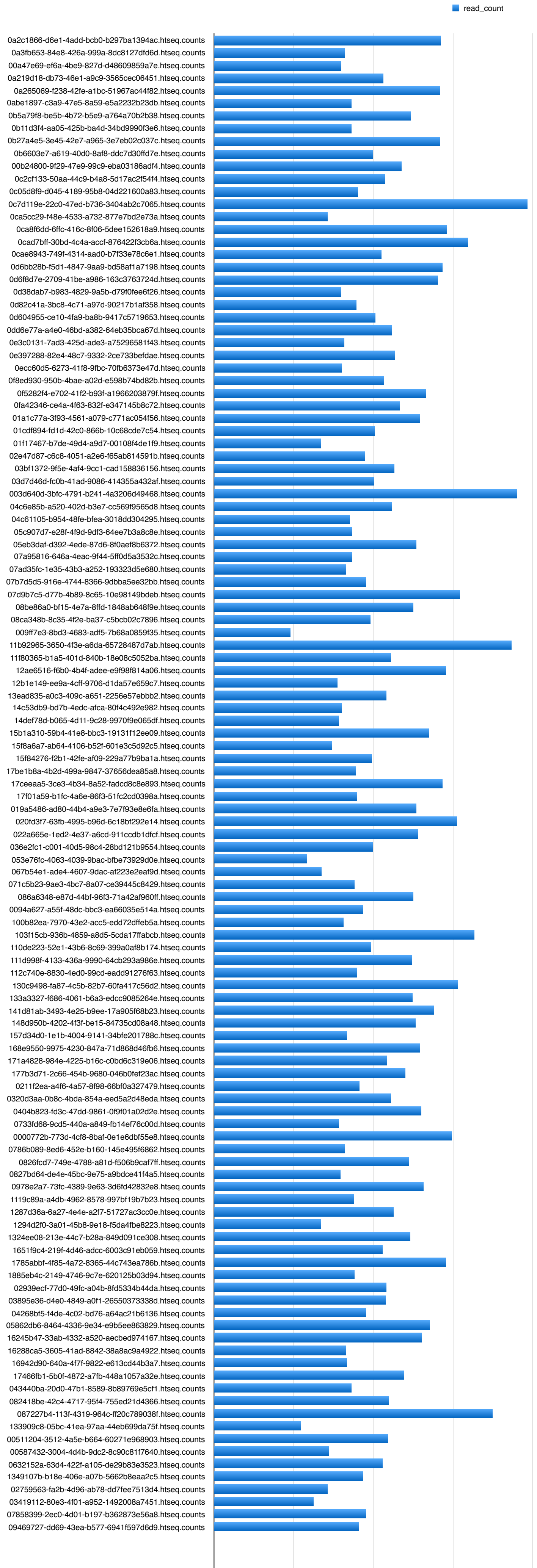


Fig1. The histogram shows the sum of reads mapped to all genes in

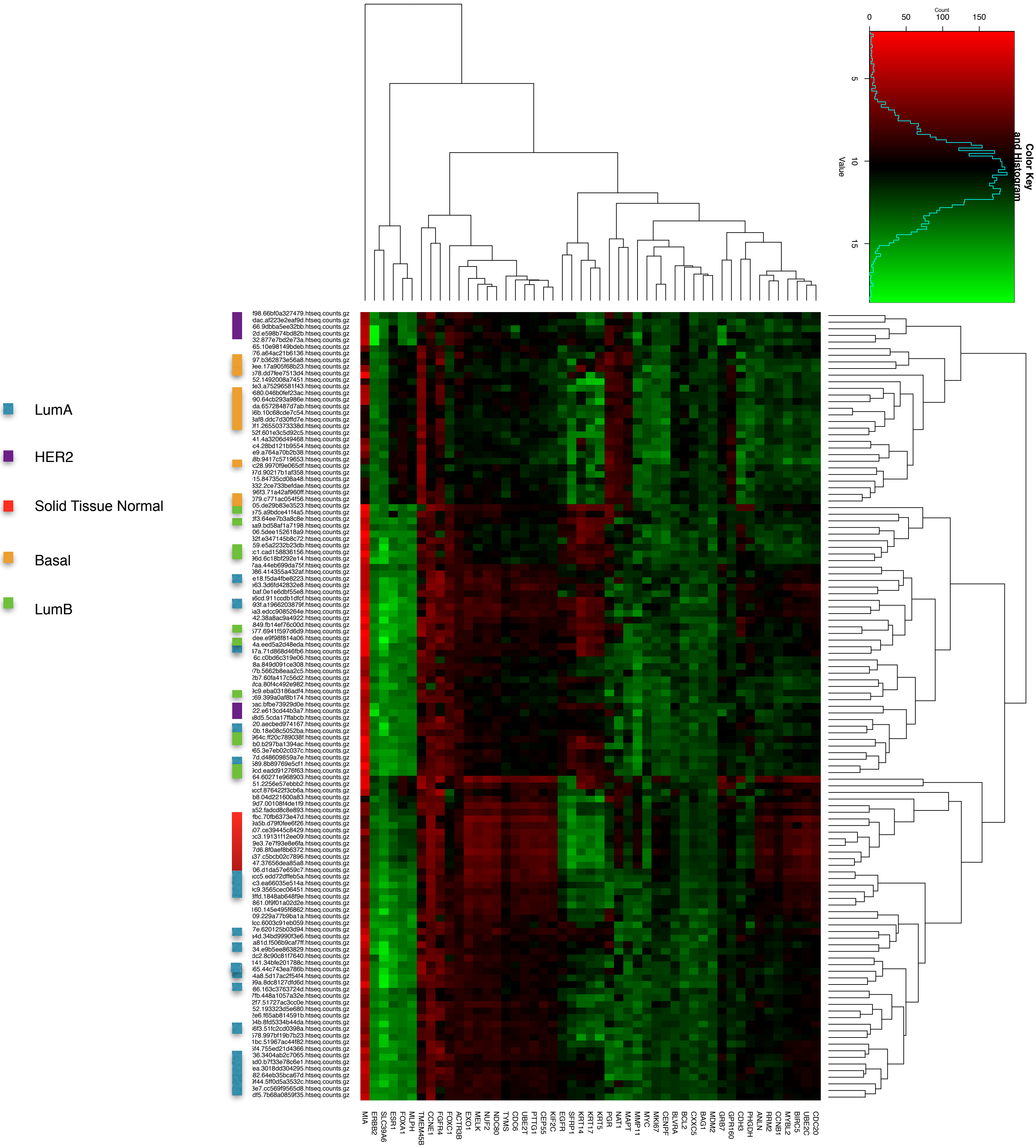


Fig 2. Heatmap embedded with sample and PAM50 gene set dendrograms on axes using by log transformed counts.

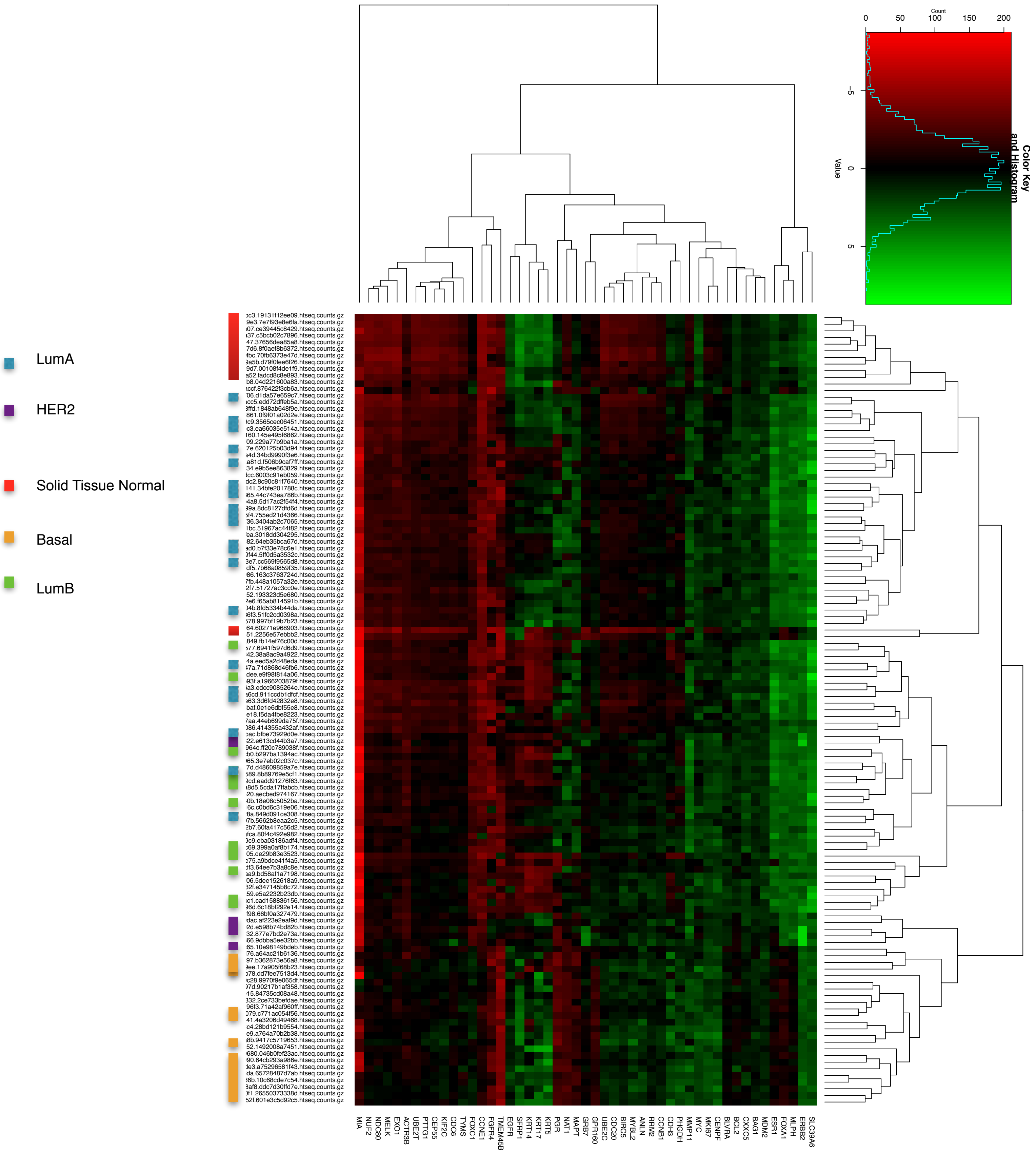
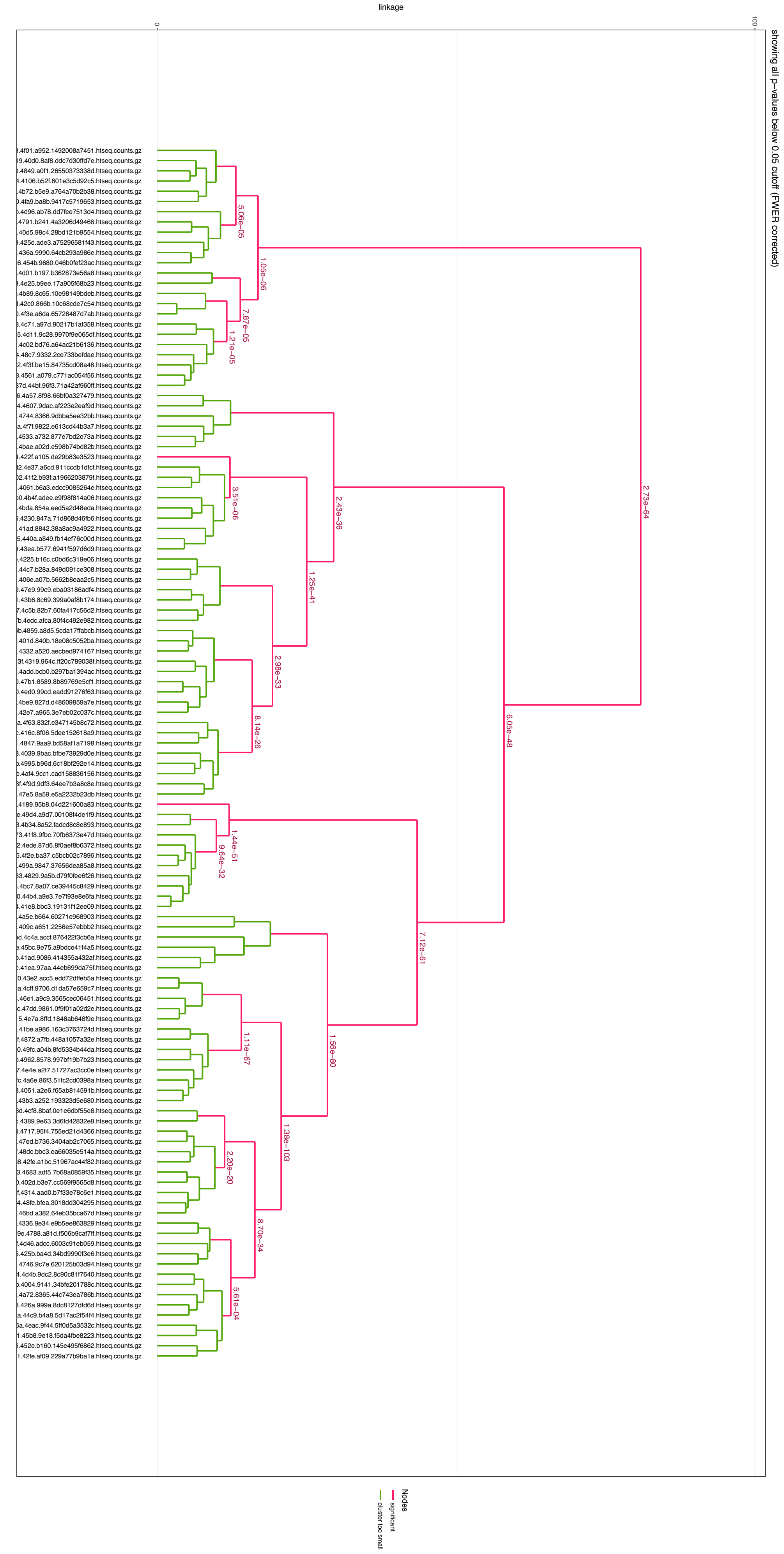


Fig 3. Heatmap in rlog transformed counts after been scaled to the median expression value in the PAM50 gene set labeled with tumor subtype.



linkage
0
100
showing all p−values below 0.05 cutoff (FWER corrected)
.4f01.a952.1492008a7451.htseq.counts.gz
19.40d0.8af8.ddc7d30ffd7e.htseq.counts.gz
.4849.a0f1.26550373338d.htseq.counts.gz
.4106.b52f.601e3c5d92c5.htseq.counts.gz
.4b72.b5e9.a764a70b2b38.htseq.counts.gz
.4fa9.ba8b.9417c5719653.htseq.counts.gz
.4d96.ab78.dd7fee7513d4.htseq.counts.gz
.4791.b241.4a3206d49468.htseq.counts.gz
.40d5.98c4.28bd121b9554.htseq.counts.gz
.425d.ade3.a75296581f43.htseq.counts.gz
.436a.9990.64cb293a986e.htseq.counts.gz
6.454b.9680.046b0fef23ac.htseq.counts.gz
.4d01.b197.b362873e56a8.htseq.counts.gz
.4e25.b9ee.17a905f68b23.htseq.counts.gz
.4b89.8c65.10e98149bdeb.htseq.counts.gz
.42c0.866b.10c68cde7c54.htseq.counts.gz
.4f3e.a6da.65728487d7ab.htseq.counts.gz
.4c71.a97d.90217b1af358.htseq.counts.gz
5.4d11.9c28.9970f9e065df.htseq.counts.gz
.4c02.bd76.a64ac21b6136.htseq.counts.gz
.48c7.9332.2ce733befdae.htseq.counts.gz
2.4f3f.be15.84735cd08a48.htseq.counts.gz
.4561.a079.c771ac054f56.htseq.counts.gz
37d.44bf.96f3.71a42af960ff.htseq.counts.gz
6.4a57.8f98.66bf0a327479.htseq.counts.gz
.4607.9dac.af223e2eaf9d.htseq.counts.gz
.4744.8366.9dbba5ee32bb.htseq.counts.gz
a.4f7f.9822.e613cd44b3a7.htseq.counts.gz
.4533.a732.877e7bd2e73a.htseq.counts.gz
.4bae.a02d.e598b74bd82b.htseq.counts.gz
.422f.a105.de29b83e3523.htseq.counts.gz
2.4e37.a6cd.911ccdb1dfcf.htseq.counts.gz
2.41f2.b93f.a1966203879f.htseq.counts.gz
.4061.b6a3.edcc9085264e.htseq.counts.gz
0.4b4f.adee.e9f98f814a06.htseq.counts.gz
.4bda.854a.eed5a2d48eda.htseq.counts.gz
.4230.847a.71d868d46fb6.htseq.counts.gz
.41ad.8842.38a8ac9a4922.htseq.counts.gz
5.440a.a849.fb14ef76c00d.htseq.counts.gz
.43ea.b577.6941f597d6d9.htseq.counts.gz
.4225.b16c.c0bd6c319e06.htseq.counts.gz
.44c7.b28a.849d091ce308.htseq.counts.gz
.406e.a07b.5662b8eaa2c5.htseq.counts.gz
.47e9.99c9.eba03186adf4.htseq.counts.gz
.43b6.8c69.399a0af8b174.htseq.counts.gz
7.4c5b.82b7.60fa417c56d2.htseq.counts.gz
b.4edc.afca.80f4c492e982.htseq.counts.gz
b.4859.a8d5.5cda17ffabcb.htseq.counts.gz
.401d.840b.18e08c5052ba.htseq.counts.gz
.4332.a520.aecbed974167.htseq.counts.gz
3f.4319.964c.ff20c789038f.htseq.counts.gz
.4add.bcb0.b297ba1394ac.htseq.counts.gz
.47b1.8589.8b89769e5cf1.htseq.counts.gz
.4ed0.99cd.eadd91276f63.htseq.counts.gz
.4be9.827d.d48609859a7e.htseq.counts.gz
.42e7.a965.3e7eb02c037c.htseq.counts.gz
a.4f63.832f.e347145b8c72.htseq.counts.gz
.416c.8f06.5dee152618a9.htseq.counts.gz
.4847.9aa9.bd58af1a7198.htseq.counts.gz
.4039.9bac.bfbe73929d0e.htseq.counts.gz
.4995.b96d.6c18bf292e14.htseq.counts.gz
e.4af4.9cc1.cad158836156.htseq.counts.gz
f.4f9d.9df3.64ee7b3a8c8e.htseq.counts.gz
.47e5.8a59.e5a2232b23db.htseq.counts.gz
.4189.95b8.04d221600a83.htseq.counts.gz
e.49d4.a9d7.00108f4de1f9.htseq.counts.gz
.4b34.8a52.fadcd8c8e893.htseq.counts.gz
73.41f8.9fbc.70fb6373e47d.htseq.counts.gz
2.4ede.87d6.8f0aef8b6372.htseq.counts.gz
5.4f2e.ba37.c5bcb02c7896.htseq.counts.gz
499a.9847.37656dea85a8.htseq.counts.gz
3.4829.9a5b.d79f0fee6f26.htseq.counts.gz
.4bc7.8a07.ce39445c8429.htseq.counts.gz
0.44b4.a9e3.7e7f93e8e6fa.htseq.counts.gz
.41e8.bbc3.19131f12ee09.htseq.counts.gz
.4a5e.b664.60271e968903.htseq.counts.gz
.409c.a651.2256e57ebbb2.htseq.counts.gz
d.4c4a.accf.876422f3cb6a.htseq.counts.gz
e.45bc.9e75.a9bdce41f4a5.htseq.counts.gz
.41ad.9086.414355a432af.htseq.counts.gz
.41ea.97aa.44eb699da75f.htseq.counts.gz
0.43e2.acc5.edd72dffeb5a.htseq.counts.gz
a.4cff.9706.d1da57e659c7.htseq.counts.gz
.46e1.a9c9.3565cec06451.htseq.counts.gz
c.47dd.9861.0f9f01a02d2e.htseq.counts.gz
5.4e7a.8ffd.1848ab648f9e.htseq.counts.gz
.41be.a986.163c3763724d.htseq.counts.gz
f.4872.a7fb.448a1057a32e.htseq.counts.gz
0.49fc.a04b.8fd5334b44da.htseq.counts.gz
.4962.8578.997bf19b7b23.htseq.counts.gz
7.4e4e.a2f7.51727ac3cc0e.htseq.counts.gz
c.4a6e.86f3.51fc2cd0398a.htseq.counts.gz
.4051.a2e6.f65ab814591b.htseq.counts.gz
.43b3.a252.193323d5e680.htseq.counts.gz
8d.4cf8.8baf.0e1e6dbf55e8.htseq.counts.gz
.4389.9e63.3d6fd42832e8.htseq.counts.gz
.4717.95f4.755ed21d4366.htseq.counts.gz
.47ed.b736.3404ab2c7065.htseq.counts.gz
.48dc.bbc3.ea66035e514a.htseq.counts.gz
8.42fe.a1bc.51967ac44f82.htseq.counts.gz
.4683.adf5.7b68a0859f35.htseq.counts.gz
.402d.b3e7.cc569f9565d8.htseq.counts.gz
f.4314.aad0.b7f33e78c6e1.htseq.counts.gz
4.48fe.bfea.3018dd304295.htseq.counts.gz
.46bd.a382.64eb35bca67d.htseq.counts.gz
.4336.9e34.e9b5ee863829.htseq.counts.gz
9e.4788.a81d.f506b9caf7ff.htseq.counts.gz
.4d46.adcc.6003c91eb059.htseq.counts.gz
.425b.ba4d.34bd9990f3e6.htseq.counts.gz
.4746.9c7e.620125b03d94.htseq.counts.gz
.4d4b.9dc2.8c90c81f7640.htseq.counts.gz
.4004.9141.34bfe201788c.htseq.counts.gz
.4a72.8365.44c743ea786b.htseq.counts.gz
.426a.999a.8dc8127dfd6d.htseq.counts.gz
a.44c9.b4a8.5d17ac2f54f4.htseq.counts.gz
3a.4eac.9f44.5ff0d5a3532c.htseq.counts.gz
1.45b8.9e18.f5da4fbe8223.htseq.counts.gz
.452e.b160.145e495f6862.htseq.counts.gz
1.42fe.af09.229a77b9ba1a.htseq.counts.gz
5.06e−05
1.05e−06
7.87e−05
1.21e−05
3.51e−06
2.43e−36
1.25e−41
2.98e−33
8.14e−26
2.73e−64
6.05e−48
1.44e−51
9.64e−32
7.12e−61
1.56e−80
1.11e−67
1.38e−103
2.20e−20
8.70e−34
5.61e−04
Nodes
significant
cluster too small

Fig 4. Sample dendrogram showing the significant clusters in analysis by sigclust2 with p-value on each node.