Differential gene expression in males vs females for head and neck squamous cancers (HNSC)

Czuee Morey

May 27, 2016

The gene expression data for HNSC was taken from the Cancer genome atlas study [1]. The data contained 566 samples for RNA-seq from males and females. In some cases, normal samples from the same individual were also present. The data was treated suitably for analysis. In part 1, the normal tissues were removed and only tumor samples were considered. In part 2, the samples for which both tumor and normal tissue are present were considered. The log of gene expression(+1) was considered for all cases, and the expression was centered around the mean in some cases.

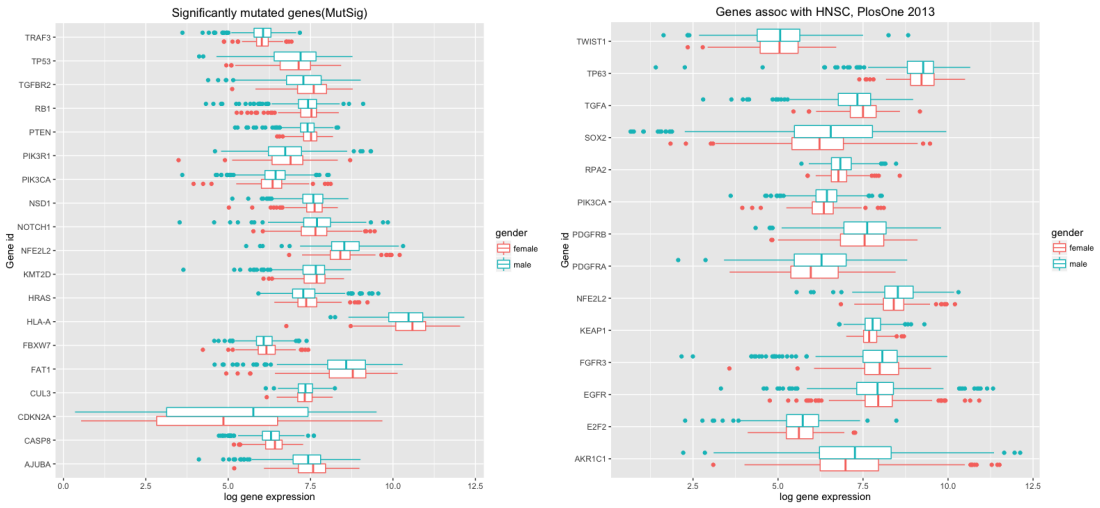


Figure 1: Significantly mutated genes [1] and genes associated with HNSC [2] do not show differential expression between males and females.

1 Gene expression in tumors

1.1 Selecting genes with highest variability

The genes were filtered to select only genes that would have the highest variance in their gene expression and are more likely to have differential expression between males and females. The genes were sorted by the standard deviation (SD) of gene expression over all samples (figure 2). The top 2,000 genes were used for further analysis.

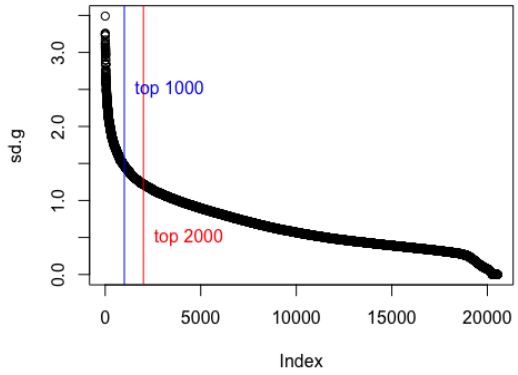


Figure 2: Plot for Standard deviation of gene expression ordered from highest to lowest. The vertical lines indicate the position for first 1000 and 2000 genes that were selected for further study.

1.2 Genes with significant difference in means between males and females

in ordert o find the genes that are different in males and females, the mean of the distrubution in the two sexes were compared. There were 21 genes which had a significant difference in their means (figure 3).

Table 1 lists the description and location of these genes. All the genes identified were X or Y linked genes, which obviously have differential expression in males and females. Hence, in the next step, I considered the gene expression difference between tumor and normal samples from the same individual to avoid identifying obvious differences between males and females.

2 Gene expression difference in tumor minus normal

Out of the 566 samples, only few samples had tumor and normal tissue data from the same patient and were extracted. The tumor and normal tissue data was present for 14 females and 29 males. In this part, the number fo genes to be considered for analysis

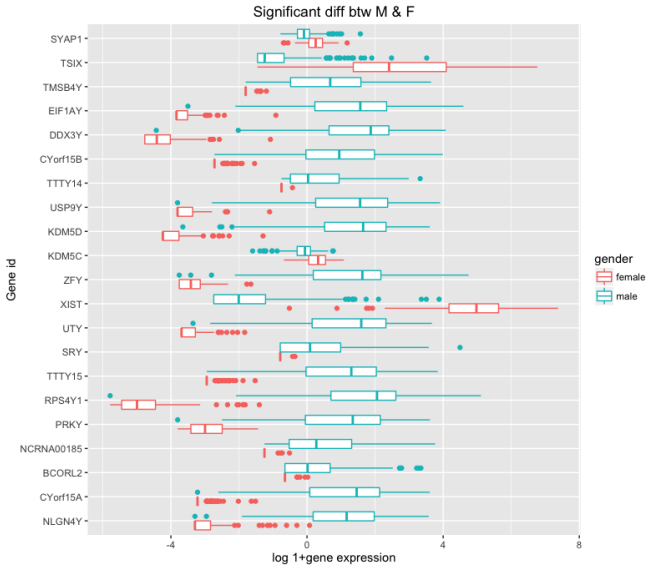


Figure 3: The genes that have a difference between their means in males and females greater than the SD of the entire distribution were selected. This analysis resulted in 21 genes.

was identified as 10,000 genes with highest variance on the basis of a QQ-plot of the SD values.

The data was processed so that the $(tumor - normal)$ values were calculated for each patient. This gave a data frame of (T-N) values for 10,000 gene columns and patient rows.

The t-test statistic was calculated as a function of gender for each gene. The p-values calculated for each gene were adjusted using the q-value package and p.adjust packages. The Wilcoxon test was also performed and the p-values were used to identify the significant genes. The q-values did not give any significant genes eevn at $q < 0.1$ with this test. The different methods picked up similar genes. Figure 4 shows the genes identified through the t-test and Wilcoxin test. The significant genes identified through the t-test are listed in table 2.

References

[1] Network, T. C. G. A. (2015) *Comprehensive genomic characterization of head and neck squamous cell carcinomas.* Nature 517, 576–582.

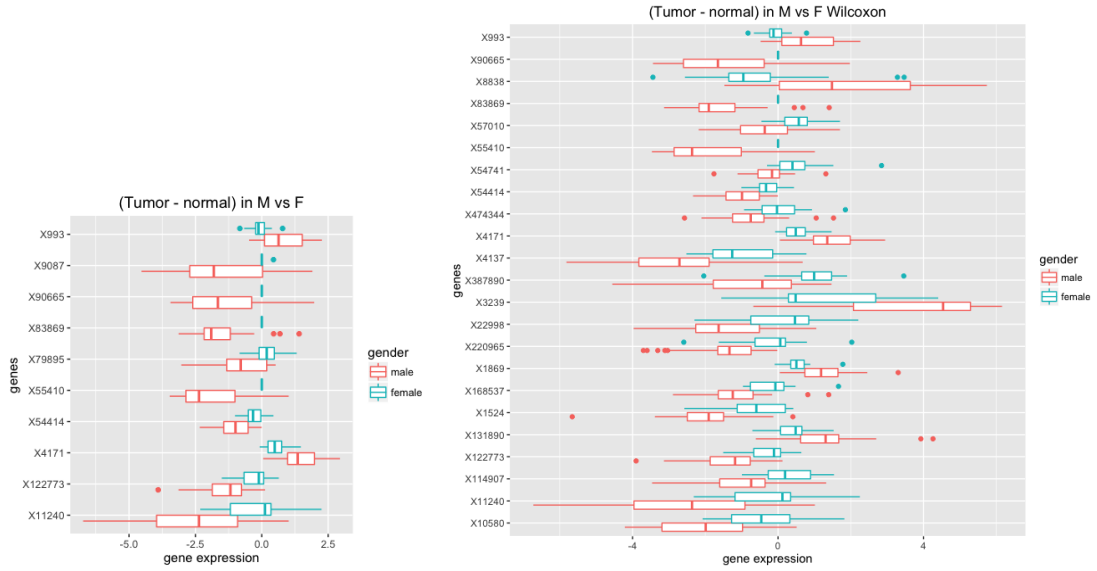
[2] Vonn Walter et al (2013) *Molecular Subtypes in Head and Neck Cancer Exhibit Distinct Patterns of Chromosomal Gain and Loss of Canonical Cancer Genes* Plos One 8(2), e56823.

Table 1: The genes that have a difference between their means in males and females greater than the SD of the entire distribution were selected. 19 descriptions - NCRNA00185 and TTTY14 are the same gene. Also CYorf15A and B have the same description (TXLNGY). All the genes are X and Y linked only.

Approved Symbol	Approved Name	Chromosome
DDX3Y	DEAD-box helicase 3, Y-linked	Yq11
EIF1AY	eukaryotic translation initiation factor 1A, Y-linked	Yq11.223
KDM5C	lysine demethylase 5C	Xp11.22-p11.21
KDM5D	lysine demethylase 5D	Yq11
NLGN4Y	neuroligin 4, Y-linked	Yq11.221
PRKY	protein kinase, Y-linked, pseudogene	Yp11.2
RPS4Y1	ribosomal protein S4, Y-linked 1	Yp11.3
SRY	sex determining region Y	Yp11.3
SYAP1	synapse associated protein 1	Xp22.31
TMSB4Y	thymosin beta 4, Y-linked	Yq11.221
TSIX	TSIX transcript, XIST antisense RNA	Xq13.2
TTTY14	testis-specific transcript, Y-linked 14 (non-protein coding)	Yq11.222
TTTY15	testis-specific transcript, Y-linked 15 (non-protein coding)	Yq11.1
USP9Y	ubiquitin specific peptidase 9, Y-linked	Yq11.2
UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	Yq11.221
XIST	X inactive specific transcript (non-protein coding)	Xq13.2
ZFY	zinc finger protein, Y-linked	Yp11.3
BCORP1	BCL6 corepressor pseudogene 1	Yq11.222
TXLNGY	taxilin gamma pseudogene, Y-linked	Yq11.222

Table 2: Significant genes identified through the t-test with q-value < 0.1. The Y-linked genes can be considered as false positives.

Approved Symbol	Approved Name	Chromosome
ATP8B4	ATPase phospholipid transporting 8B4 (putative)	15q21.2
CDC25A	cell division cycle 25A	3p21
KLHDC1	kelch domain containing 1	14q21.3
MCM2	minichromosome maintenance complex component 2	3q21
PADI2	peptidyl arginine deiminase 2	1p35.2-p35.1
SIAE	sialic acid acetylesterase	11q24
TBL1Y	transducin (beta)-like 1, Y-linked	Yp11.2
TMSB4Y	thymosin beta 4, Y-linked	Yq11.221
TTTY14	testis-specific transcript, Y-linked 14 (non-protein coding)	Yq11.222



(a) t-test

(b) Wilcoxon test

Figure 4: Significant genes identified through statistical tests (a) t-test with q-value $<$ 0.1 and (b) Wilcoxin test with P-value $<$ 0.001.