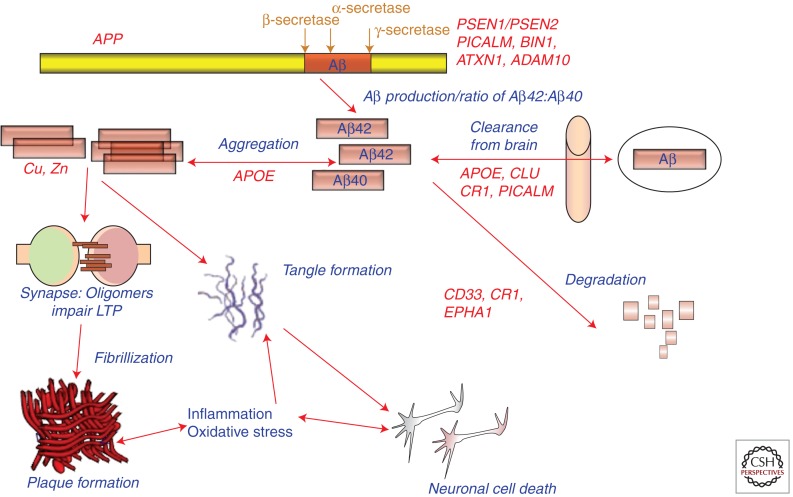
**FINAL: BIOT512**

**NAME: (Seddie McKenzie)**

Family history is the second strongest risk factor for Alzheimer disease (AD) following advanced age. Twin and family studies indicate that genetic factors are estimated to play a significant role in at least 80% of AD cases. Some of this heritability can be explained by common variants with small effect in genes such as *APOE, BIN1, TREM2, CR1, PICALM, CLU, ABCA7, CD2AP, EPHA1*, and the *MS4A* gene cluster. *MS4A* gene cluster has recently emerged as a risk-associated region including *MS4A6A* gene.

*Tanzi 2012*

**Q1: Please find a peer-reviewed reference reporting an association of MS4A6A gene or variant(s) with AD (Boolean operators)!**

**Reference: Alzheimer's disease susceptibility variants in the MS4A6A gene are associated with altered levels of MS4A6A expression in blood**

**Q2: Find a coding (missense) SNP in *MS4A6A* that has been associated with AD!**

**SNP: rs610932**

**Q3: Please find SNP information by Variant Effect Prediction (e.g location, consequence, impact, biotype, the exon number in which the variant is located, amino acid changes from x to y)!**

(hint: use filter to restrict the result to “*show one selected consequence per each variant*” to avoid redundancy)

Please choose transcript MS4A6A-210:ENST00000529054.5 (it is easy to find in Ensembl for consistency).

| **UPLOADED VARIANT** | **LOCATION** | **CONSEQUENCE** | **IMPACT** | **SYMBOL** | **GENE** | **FEATURE TYPE** | **FEATURE** | **Biotype** | **Exon** | **AA** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Rs7232** | [11:60173126-60173126](https://uswest.ensembl.org/Homo_sapiens/Location/View?contigviewbottom=variation_feature_variation%3Dnormal;db=core;r=11:60173076-60173176;tl=ZNk8sGSr7TCnOKOh-3764598) | Missense\_Variant | Moderate | MS4A6A | ENSG00000110077 | Transcript | [ENST00000529054](https://uswest.ensembl.org/Homo_sapiens/Transcript/Summary?db=core;t=ENST00000529054;tl=ZNk8sGSr7TCnOKOh-3764598) | protein\_coding | 7/8 | T/S |

**Q4: chr, location, exon number, direction of transcription, length of the protein encoded by the transcript**

| **CHROMOSOME** | **LOCATION** | **EXON NUMBER** | **DIRECTION** | **LENGTH** |
| --- | --- | --- | --- | --- |
| **11** | [60,172,047-60,184,633](https://uswest.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG00000110077;r=11:60172047-60184633;t=ENST00000529054) | 8 | REVERSE | 276 aa |

**Q5: Please write a fasta file with the exon sequence associated with the SNP (the SNP is located in that exon)**!

>>MS4A6A\_EXON7

GGAACTCTCTCTCTGATGCTGATTTGCACTCTGCTGGAATTCTGCCTAGCTGTGCTCACT

GCTGTGCTGCGGTGGAAACAGGCTTACTCTGACTTCCCTGGG

**Q6: What cell type in the human brain has the largest gene expression of MS4A6A?**

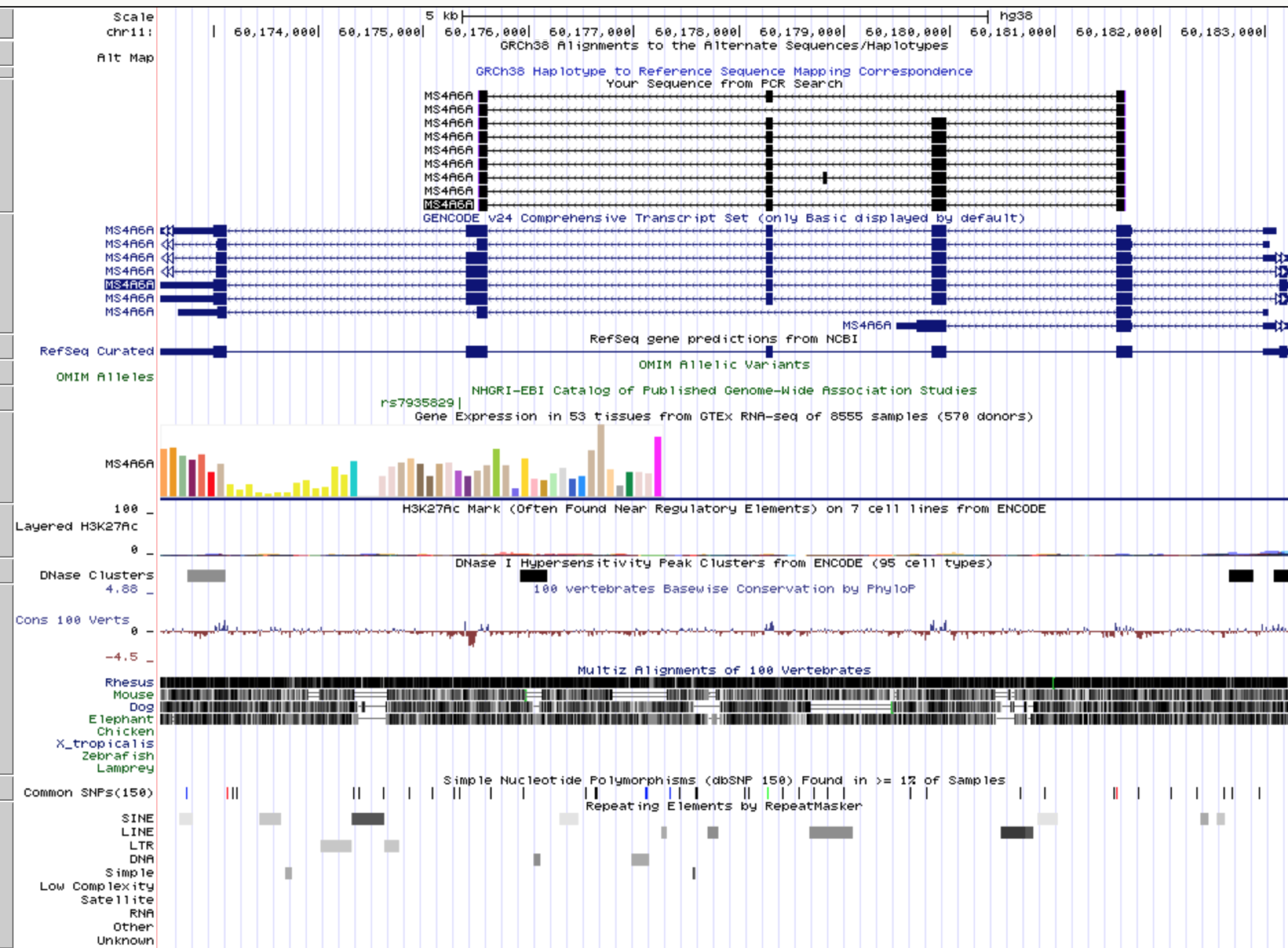
**Brain cell type:**

To understand the contribution of *MS4A6A* gene to the development of AD, researchers generated microglia from human iPS cells to study *MS4A6A* gene biology. *MS4A6A* gene was knocked down by siRNAs (silencing RNAs) to investigate the impact of gene-specific decreased gene expression on cellular functions (loss-of-function phenotype). Before proceeding to RNAseq, the success of MS4A6A gene knockdown was confirmed by qPCR. Accordingly, primers were designed for MS4A6A-210:ENST00000529054.5 transcript using nucleotide sequences from exon2 to exon6 with an expected product length longer than 200bp.

**Q7: Please insert the sequence(s) of one primer-pair and the expected PCR product length below! (hint: change the default output 10 to 1; change the minimum product length accordingly).**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Sequence (5'->3')** | **Template strand** | **Length** | **Start** | **Stop** | **Tm** | **GC%** | **Self complementarity** | **Self 3' complementarity** |
| **Forward primer** | TGGCAATTCACACATAAGGCT | Plus | 21 | 64 | 84 | 58.19 | 42.86 | 4.00 | 3.00 |
| **Reverse primer** | CCAGAGATGATAAACAAAAGCTTGG | Minus | 25 | 562 | 538 | 58.67 | 40.00 | 6.00 | 2.00 |
| **Product length** | 499 | | | | | | | | |

**Q8: Please confirm the outcome of Q7 with “in silico PCR” that those primers will amplify *MS4A6A* transcript! Please insert a figure of the result of “in silico PCR”.**



After confirming the success of siRNA knockdown by qPCR, researchers isolated total RNAs from different experimental conditions. The experimental setup was the following: cells are transfected with plasmids containing siRNAs for *MS4A6A* and *MS4A4A* transcripts as well as scrambled (SRC) plasmids for controls. In addition, subsets of the cells were treated with LPS (lipopolysaccharides: found in the outer membrane of Gram-negative bacteria) to evaluate inflammatory responses of the microglia phenotypes (how loss-of-function of certain genes perturbs inflammatory response).

amand_MS4.tiff

Please perform differential expression analysis (DEA) on the dataset “rnaseq\_counts\_MS4\_experiment.csv “!

This dataset contains 22 samples under different conditions (6) with biological replicates. Based on the phenotype file ("sample\_info.csv"), you must figure out the condition of each sample for the *RNAseq counts file* to set the condition. At first, the researchers focused on DEGs between MS4A6A knockdown and SRC control population! Please set up the contrast and perform the analysis!

#########################################################

#########################################################

Code Hints:

ms4=read.delim("rnaseq\_counts\_MS4\_experiment.csv",header=T,sep=",", row.names = 1)

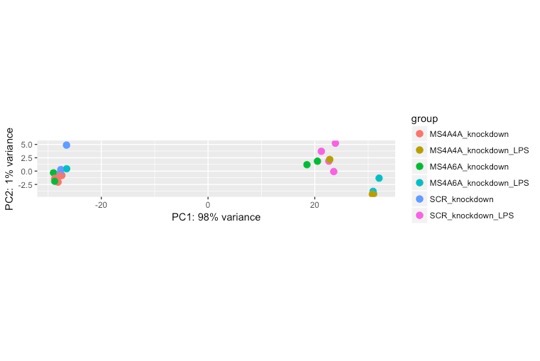
pheno <- read.delim("sample\_info.csv", header=T, sep=",", row.names = 1)

rownames(pheno)==colnames(ms4) #TRUE

Please follow your previous R code!

####################################################################################################################################

**Q9: Please insert a figure of the PCA plot!**

****

**Q10: How much of the variance is explained by the first principal component?**

**Principal Component 1: 98%**

**Q11: What drives the variance in PC1 (hint: gene perturbation or treatment)?**

**The perturbed genes drive the variance**

**Q12: How many significant results are at p < 0.05, 0.01 and 0.001?**

Results:

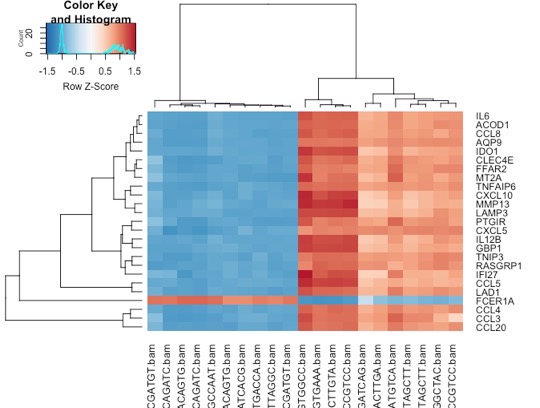
**6355** of gene threshold 0.05

**4725** of gene threshold 0.01

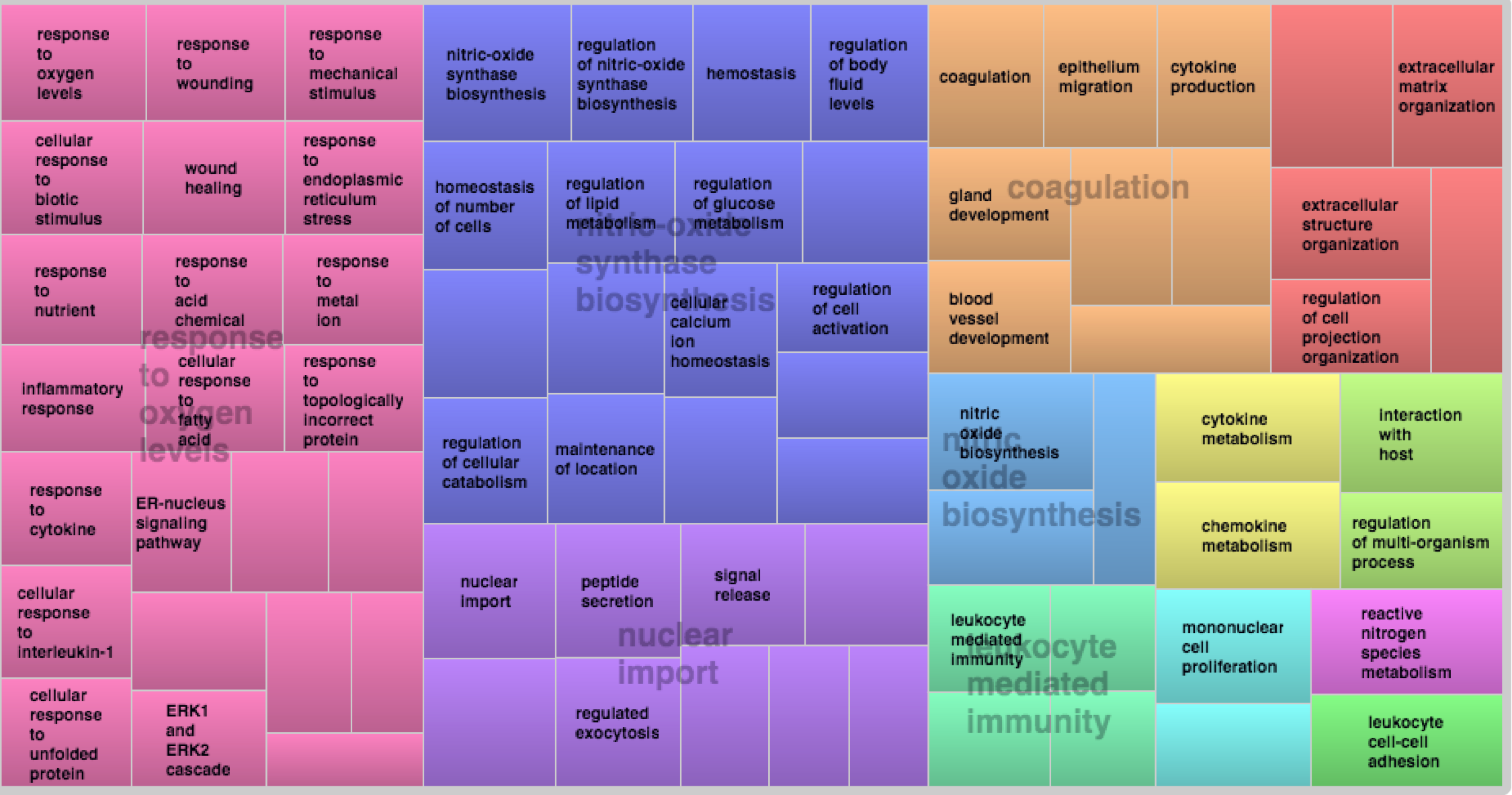
**2777** of gene threshold 0.001

**Q13: Please insert a figure with a heatmap of DEGs!**

The researchers applied a very stringent approach. Genes with adjusted p-value less than 0.001 were considered significant and used for downstream analysis.



**Q14: Please perform functional annotation with GO terms using gProfiler and Revigo! Please insert figure *TreeMap* of the result!**

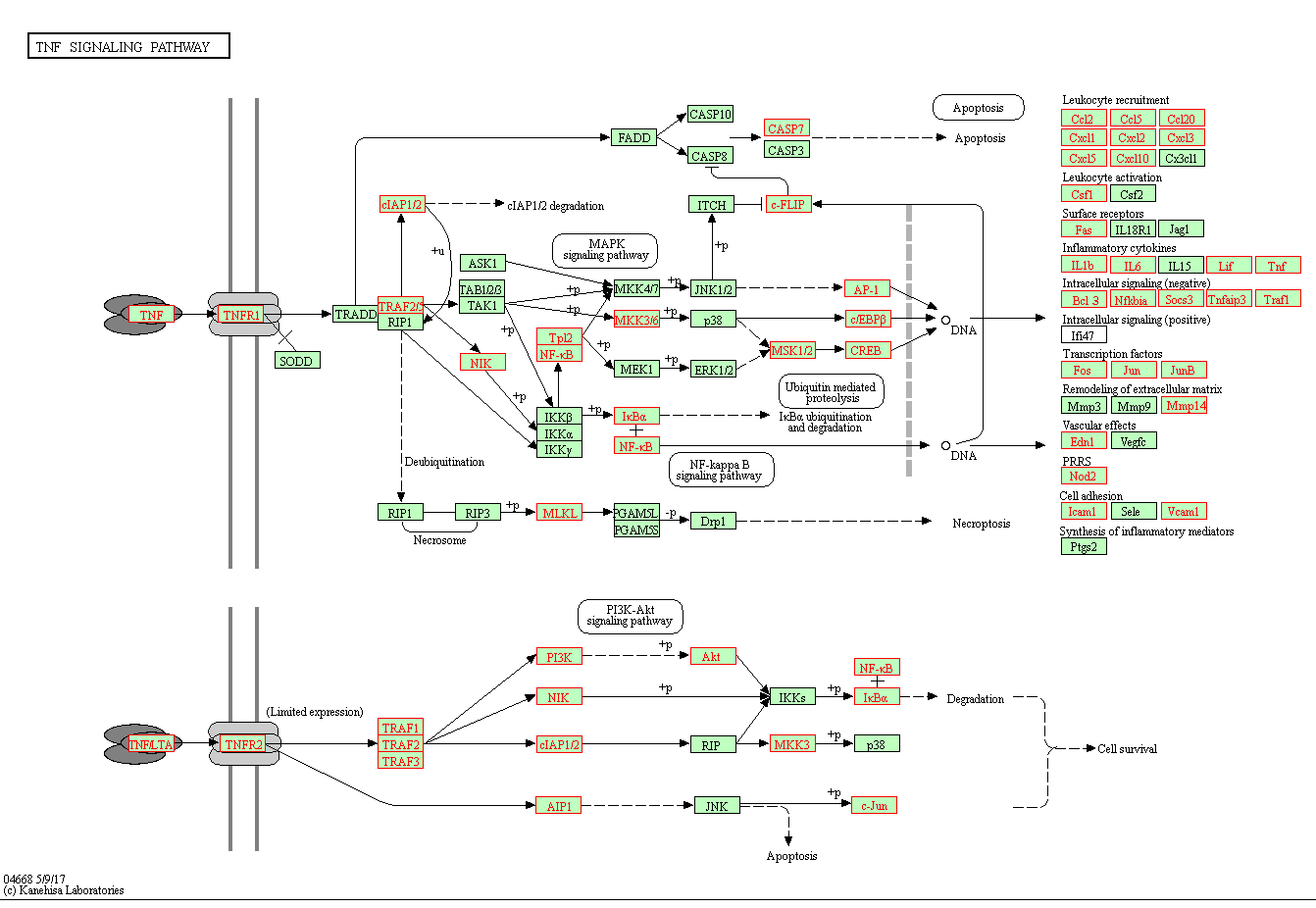
****

**Q15: Perform pathway analysis (functional database-KEGG) with** [webgestalt.org](http://www.webgestalt.org/option.php) **using ORA as the statistical method and genome protein coding as the reference enrichment set! Please insert the first 3 most significant pathways!**

| ID | Name | #Gene | FDR |
| --- | --- | --- | --- |
| Hsa04668 | TNF – signaling pathway – Homo sapiens (human) | 54 | 3.63e-12 |
| Hsa04062 | Chemokine signaling pathway – Homo sapiens (human) | 72 | 3.86e-10 |
| Hsa04064 | NF-kappa B signaling pathways – Homo sapiens (human) | 42 | 7.41e-08 |

#### **Q16: Insert a figure of the most significant KEGG pathway (**hint: click on the KEGG ID on right-hand side (Detailed information of the enriched categories)

#### **Q16: Insert a figure of the most significant KEGG pathway (**hint: click on the KEGG ID on right-hand side (Detailed information of the enriched categories)\



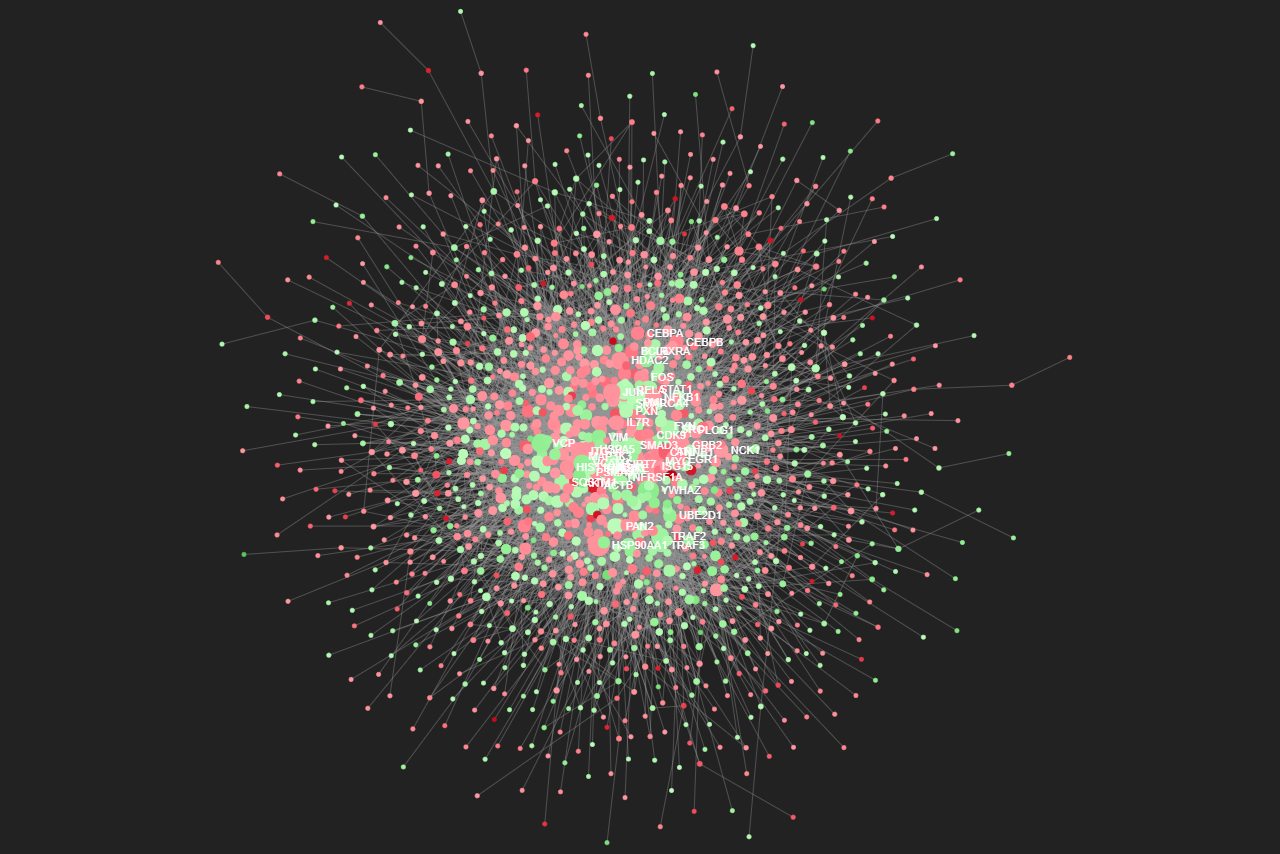
**Q17: What do the genes highlighted in red represent in the picture?**

The red highlighted genes represent genes included in the biological processes

**Q18: Please perform network analysis with** [www.networkanalyst.ca](http://www.networkanalyst.ca) **tool! Choose the “*list of genes or protein module*”! Upload the significant genes (FDR< 0.001) with logFC values.**

**Select**: PP- interaction, IMEx Interactome , zero-order network and GO:MF in Functional explorer (all nodes). View=expression (the nodes should be red and green)

**Please insert the network figure!**

****

**Q19: Please perform functional analysis! Please list the 3 most significant MF terms!**

**"Pathway","Total","Expected","Hits","P.Value","FDR"**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**Q20: Definition of a hub gene!**

An open question is whether or when **hub gene** selection leads to more meaningful **gene** lists than a standard statistical analysis based on significance testing when analyzing genomic data sets like **gene** expression or DNA methylation data

**Q21: Describe “betweeness centrality” of a network!**

Betweenness centrality is an indicator of a node's [centrality](https://ipfs.io/ipfs/QmXoypizjW3WknFiJnKLwHCnL72vedxjQkDDP1mXWo6uco/wiki/Centrality.html) in a [network](https://ipfs.io/ipfs/QmXoypizjW3WknFiJnKLwHCnL72vedxjQkDDP1mXWo6uco/wiki/Graph_(discrete_mathematics).html).

**Q22: Pick a hub gene with the highest betweeness centrality measurement!**

***Hub Gene***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Name** | **Degree** | **Betweeness** | **Expr** |
| **P07900** | **HSP90AA1** | **143** | **119870** | **1.1149** |

**Q23: Find the following info for the hub gene:**

**chr, location, number of exons, transcription direction (Ensembl)**

| Chromosome | location | Exons | Transcription dir. |
| --- | --- | --- | --- |
| 14 | [102,081,049-102,139,686](http://uswest.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG00000080824;r=14:102081049-102139686;t=ENST00000334701) | 12 | Reverse strand |

To confirm the gene expression of the hub gene, please make a bar graph with package “ggplot2”! Please study the code below and tailor to your specific needs!

install.packages("dplyr")

install.packages("ggplot2")

library(dplyr)

library(ggplot2)

library(reshape)

norm\_data <- counts(dds, normalized=TRUE)

norm\_data <- as.data.frame(norm\_data)

norm\_data$GENE <- rownames(norm\_data)

head(norm\_data)

gene <-filter(norm\_data, GENE=="your\_hubgene")

gene <- gene[,-23]

tgene <- t(gene)

tgene <- as.data.frame(tgene)

tgene$GENE <- "your\_hubgene"

tgene$group <- pheno$condition

rownames(tgene) <- NULL

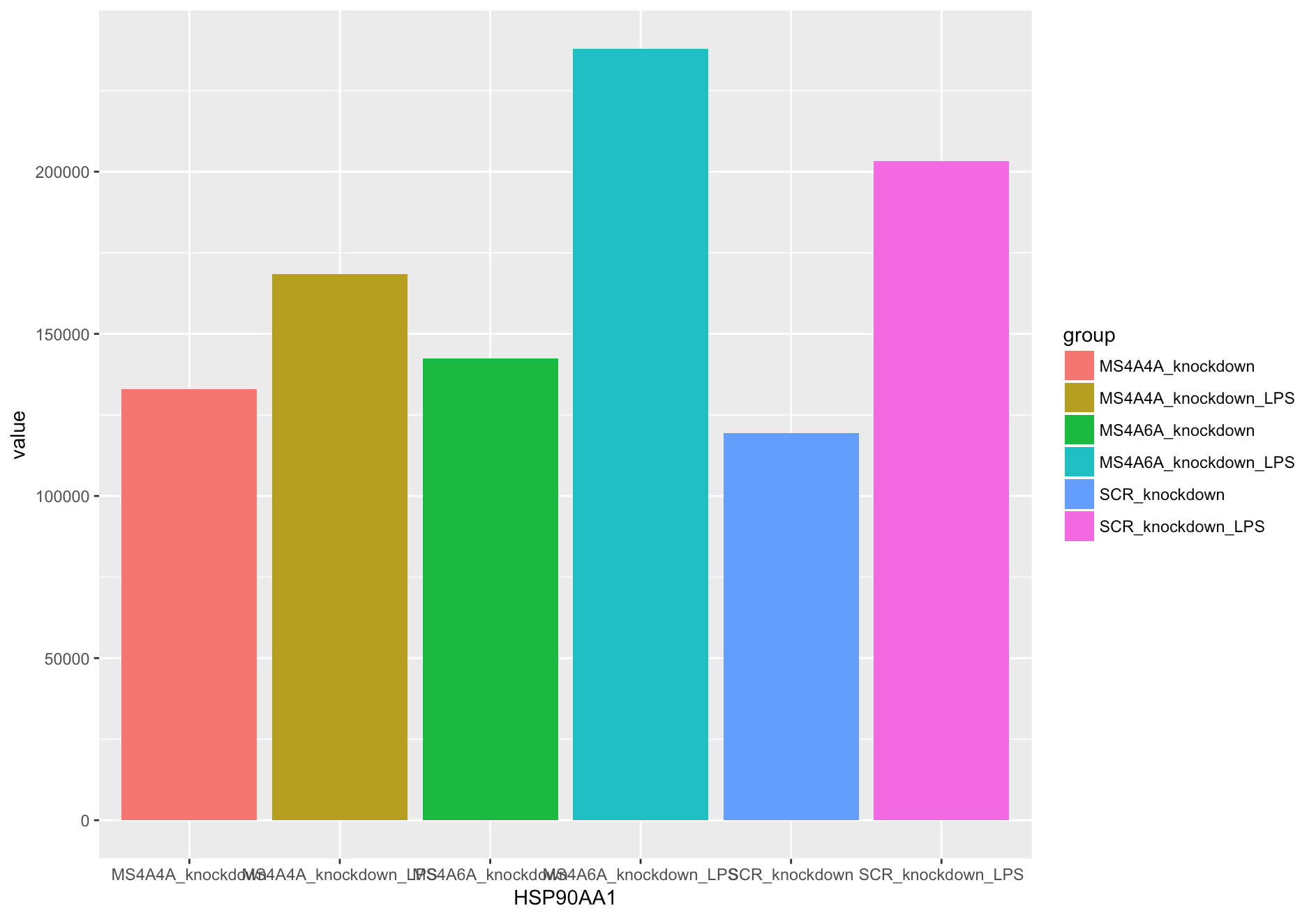
var <- melt(tgene)

p <- ggplot(data = var, aes(x =group, y =value, fill = group))

p <- p + geom\_bar(stat = "identity") + labs(x="your\_hubgene")

p

**Q24: Please insert a bar graph with the hub gene expression under experimental conditions!**

****

**Q25: Insert the human protein sequence encoded by the hub gene in fasta format (how many aa)?**

**>>\_**HSP90AA1

**Translation (854 aa):**

MPPCSGGDGSTPPGPSLRDRDCPAQSAEYPRDRLDPRPGSPSEASSPPFLRSRAPVNWYQEKAQVFLWHLMVSGSTTLLCLWKQPFHVSAFPVTASLAFRQSQGAGQHLYKDLQPFILLRLLMPEETQTQDQPMEEEEVETFAFQAEIAQLMSLIINTFYSNKEIFLRELISNSSDALDKIRYESLTDPSKLDSGKELHINLIPNKQDRTLTIVDTGIGMTKADLINNLGTIAKSGTKAFMEALQAGADISMIGQFGVGFYSAYLVAEKVTVITKHNDDEQYAWESSAGGSFTVRTDTGEPMGRGTKVILHLKEDQTEYLEERRIKEIVKKHSQFIGYPITLFVEKERDKEVSDDEAEEKEDKEEEKEKEEKESEDKPEIEDVGSDEEEEKKDGDKKKKKKIKEKYIDQEELNKTKPIWTRNPDDITNEEYGEFYKSLTNDWEDHLAVKHFSVEGQLEFRALLFVPRRAPFDLFENRKKKNNIKLYVRVFIMDNCEELIPEYLNFIRGVVDSEDLPLNISREMLQQSKILKVIRKNLVKKCLELFTELAEDKENYKKFYEQFSKNIKLGIHEDSQNRKKLSELLRYYTSASGDEMVSLKDYCTRMKENQKHIYYITGETKDQVANSAFVERLRKHGLEVIYMIEPIDEYCVQQLKEFEGKTLVSVTKEGLELPEDEEEKKKQEEKKTKFENLCKIMKDILEKKVEKVVVSNRLVTSPCCIVTSTYGWTANMERIMKAQALRDNSTMGYMAAKKHLEINPDHSIIETLRQKAEADKNDKSVKDLVILLYETALLSSGFSLEDPQTHANRIYRMIKLGLGIDEDDPTADDTSAAVTEEMPPLEGDDDTSRMEEVD

Next, researchers were curious about the conservation between human and mouse proteins and they performed sequence analysis.

**Q26: Please perform pair-wise global sequence alignment with Needleman-Wunsch Global Align Protein Sequences tool!**

|  |  |
| --- | --- |
| Score | 3587 |
| Identity | 726/855 85% |
| Positives (similarity + identity) | 731/855. 85% |
| Gaps | 123/855. 14% |
| **Homologous** | Y |

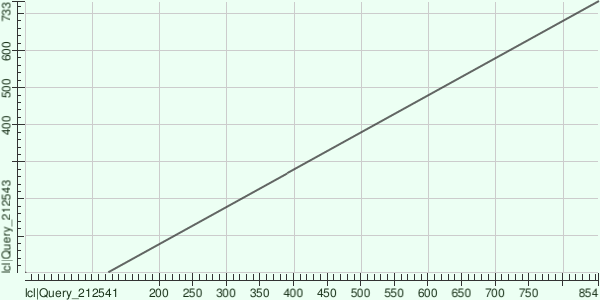
Hint: search for the protein sequences by indicating the species

symbol human

symbol mouse

without indicating the species, the default is human protein

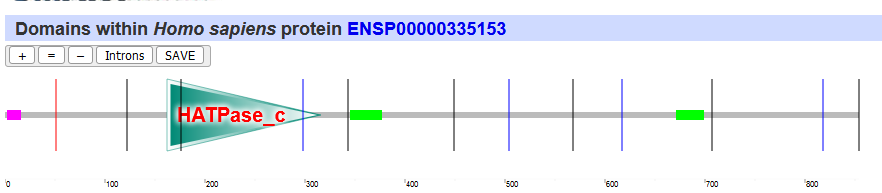
**Q27: Insert a dot plot to support your last claim in the previous Q26!**



To outline the potential functional role of this hub gene in cellular processes researchers performed an initial domain/motif analysis (SMART).

**Q28: What is the predicted motif in the protein encoded by the hub gene?**

**Please insert the figure!**

****

**Q29: What is the potential** **posttranslational modifications of this protein?**

Phosphorylation (44), Acetylation (28),

Methylation (1), Malonylation (1)

Ubiquitination (45), Nitrosylation (1)

**Q30: What is the potential function of this protein?**

heat shock protein 90kDa alpha (cytosolic), class A member 1; Molecular chaperone that promotes the maturation, structural maintenance and proper regulation of specific target proteins involved for instance in cell cycle control and signal transduction. Undergoes a functional cycle that is linked to its ATPase activity. This cycle probably induces conformational changes in the client proteins, thereby causing their activation. Interacts dynamically with various co-chaperones that modulate its substrate recognition, ATPase cycle and chaperone function

**Q31: Could this protein be a potential dug target? Determine with STITCH** [http://stitch.embl.de](http://stitch.embl.de/) **if there is any drug in the database influencing the hub protein (only one drug)**

**Name the drug! Tanespimycin**

**Please insert the figure!**

****

**Q32: Does this protein have FDA approved inhibitors** <http://dgidb.org/> **(drug-gene interaction)**

**List the first one if there is any!**

Drug Name: Rifabutin

|  |  |
| --- | --- |
| Year of Approval - 1992 |  |
| Drug Class. ANTI-BIOTIC |  |

**Total points: 100**

|  |  |  |
| --- | --- | --- |
| Question | Max point |  |
| 1. Ref paper | 1 |  |
| 2. rsSNP | 1 |  |
| 3. SNP info | 3 |  |
| 4. Transcript info | 2.5 |  |
| 5. Fasta format | 3 |  |
| 6. Brain Expression | 2.5 |  |
| 7. Primers | 5 |  |
| 8.in silico PCR | 5 |  |
| 9. PCA plot | 5 |  |
| 10. PCA1 | 3 |  |
| 11. PCA driven | 3 |  |
| 12. Significant DEGs | 3 |  |
| 13. Heatmap | 5 |  |
| 14. gProfiler, Revigo | 5 |  |
| 15. ORA | 5 |  |
| 16. ORA figure | 5 |  |
| 17. Figure Inteprot | 3 |  |
| 18. Network | 5 |  |
| 19. MF functional | 3 |  |
| 20. Def hub gene | 2 |  |
| 21. Def between | 2 |  |
| 22. Pick hub gene | 2 |  |
| 23. Hub gene loc. | 2 |  |
| 24. Bar Graph | 8 |  |
| 25. Hub G. fasta | 2 |  |
| 26. Global Alignment | 4 |  |
| 27. Dot plot | 2 |  |
| 28. Motif | 2 |  |
| 29. PTM | 1 |  |
| 30. FUNCTION | 1 |  |
| 31. drug1 | 2 |  |
| 32. drug2 | 2 |  |