THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Hugh Kaul Precision Medicine Institute

A mediKanren-based drug repurposing pipeline using patients' RNA sequencing data

Demo case: CHAMP1

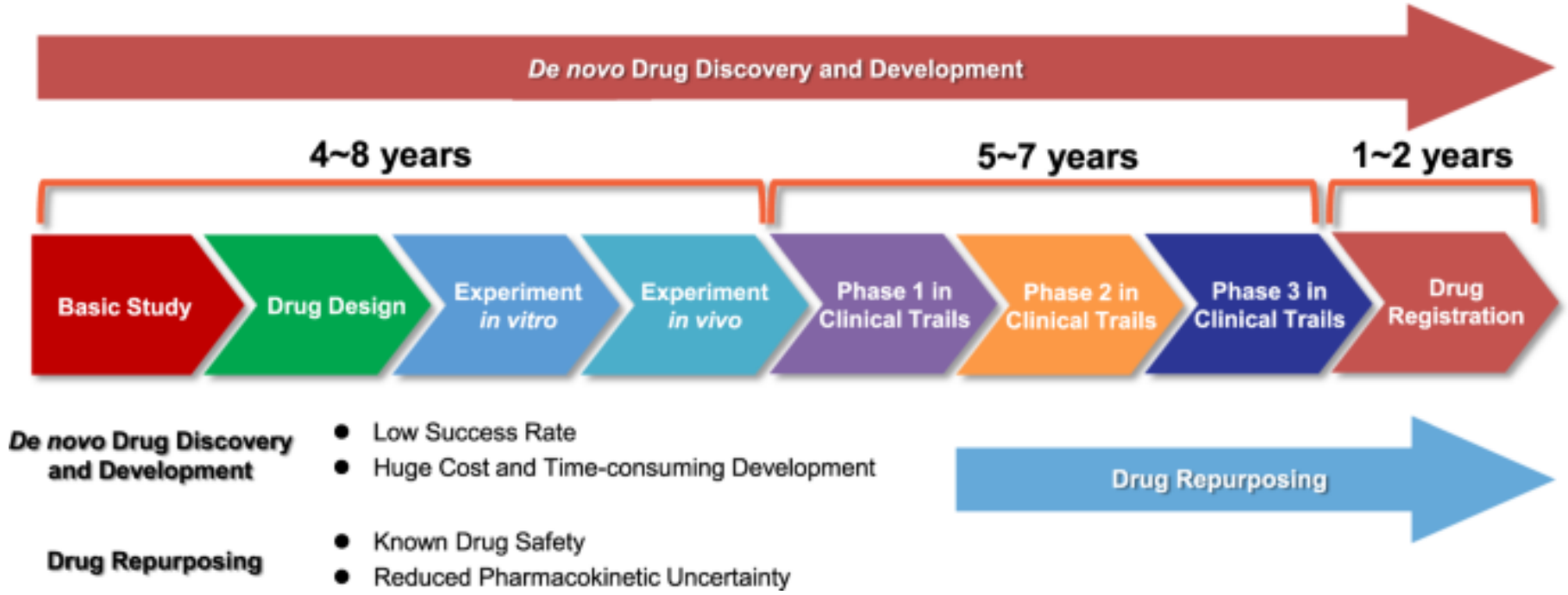
Presented by: Thi K.Tran-Nguyen, Ph.D.
Hugh Kaul Precision Medicine Institute

email: kimthi@uab.edu
twitter: kimthi1011

Outline

1. Background
   - drug repurposing for rare genetic diseases
   - transcriptome to model genetic disorders
2. CHAMP1 Background – a Case Study
3. Workflow
   - Fastq -> gene count -> DEG -> GSEA -> mediKanren queries
4. Explore the data
5. Three Strategies to Find Drugs with mediKanren
6. Drugs Curation

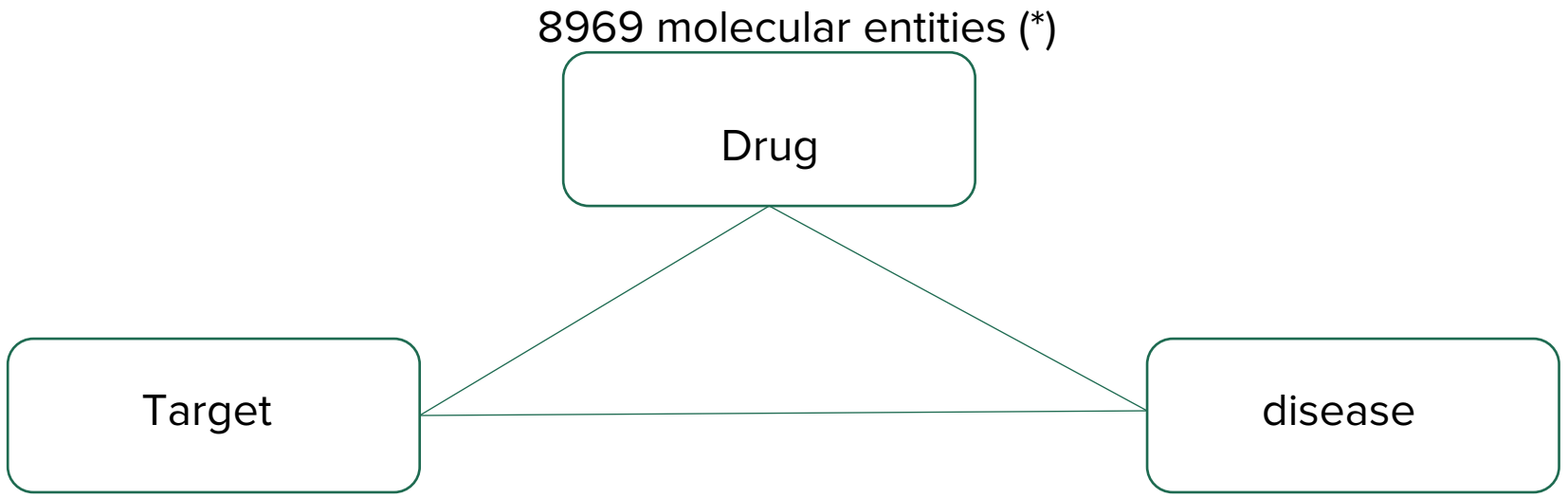
Drug repurposing for rare genetic diseases



https://www.nature.com/articles/s41392-020-00213-8

Drug repurposing for rare genetic diseases

- **rare/orphan/neglected diseases** : limited commercial interest for *de novo* drug discovery
- **drug repurposing:** identify new indications for FDA-approved/investigational drugs



~ 400 known drug targets
~ 25000 human genes

~ 5000 treated diseases
~ 7000 rare diseases
? total human diseases

- (*) NCGC Pharmaceutical Collection
- https://rarediseases.info.nih.gov/diseases/pages/31/faqs-about-rare-diseases

Patients' RNA sequencing data

- WES and WGS can only detect causal variants 25-50% of the cases
- RNA-seq can detect transcriptional abnormalities caused by DNA variants
  - ❖ e.g., splice- altering mutations (splice-disruption, splice-gain, exon skipping, exon extension)
  - ❖ increase diagnostic yield

Improving genetic diagnosis in Mendelian disease with transcriptome sequencing

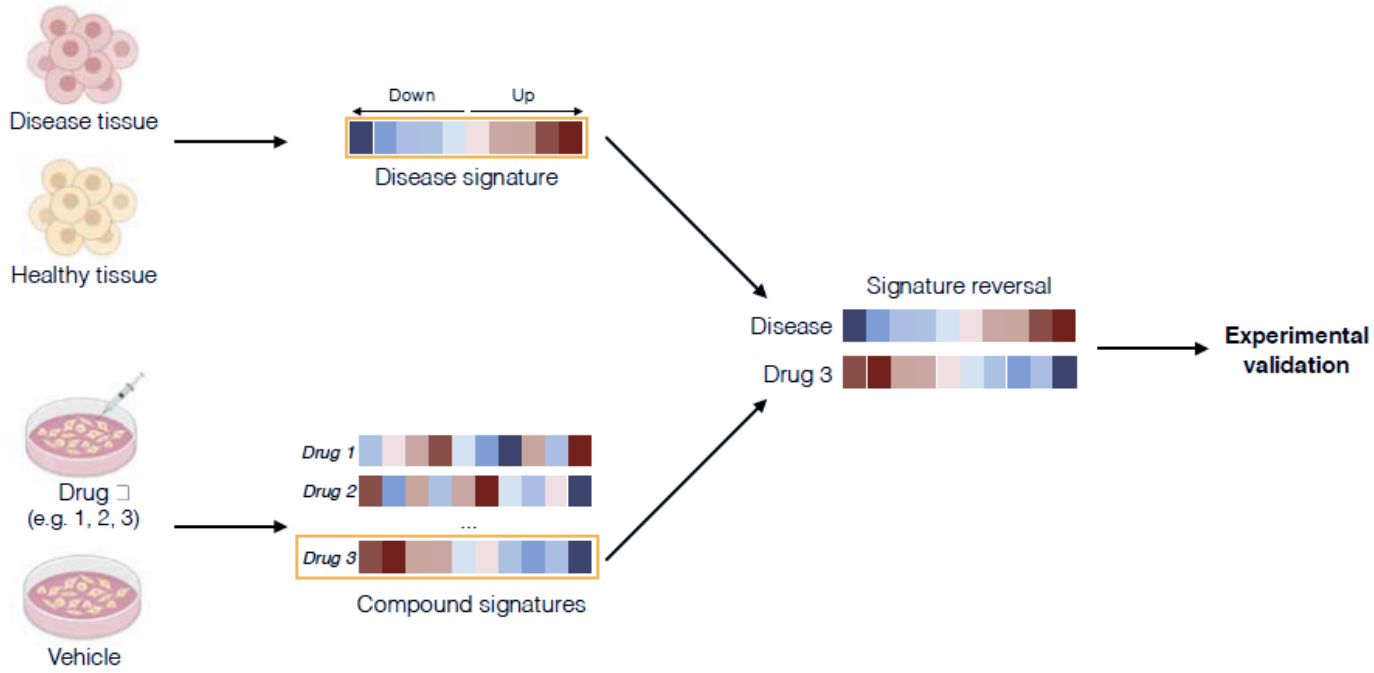
Beryl B. Cummings[1,2,3], Jamie L. Marshall[1,2], Taru Tukiainen[1,2], Monkol Lek[1,2,4,5], Sandra D...
+ See all authors and affiliations

Science Translational Medicine 19 Apr 2017:
Vol. 9, Issue 386, eaal5209
DOI: 10.1126/scitranslmed.aal5209

- RNA-seq data in combination with WES and WGS
- muscle biopsies (n=50 rare muscle disorders)
- GTEX reference panel n=184 skeletal muscle
- identify additional 35% patients with negative results from WES and WGS

Patients' RNA sequencing data considerations

- Patients RNA-seq reflect a snapshot of gene expressions (time/space limit)
- Different cell/tissue expressed different genes and alternative transcripts
- Important to sequence disease-relevant tissues/cells
  - ❖ Challenges to get access to disease-relevant tissues
- In situ biopsies vs cell culture vs organoids
- Transcriptome Reversal Approach:



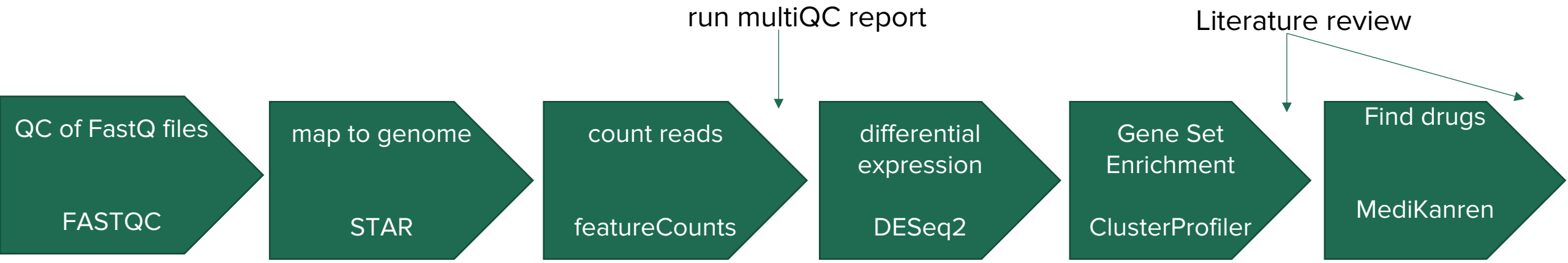
https://www.biorxiv.org/content/10.1101/2020.05.13.093468v1

CHAMP1 genetic disorder

- CHAMP1: chromosome alignment-maintaining phosphoprotein 1
- 13q34, contains 3 exons (two 5'-unstranslated exons + one coding exon)
- encodes a zin-finger protein (812 aa), maintains kinetochore-microtubule attachment during mitosis and regulates proper chromosome segregation
- typically caused by *de novo* frameshift or nonsense mutation that leads to loss of functionally important zinc finger domain at the C-terminus
- mutation affects cell division and brain development
- phenotypes: intellectual disability, impaired speech, muscular hypotonia, facial anomalies, motor developmental delay, seizures

Isidor *et al*. **Human Mutation** 2016

WORK-FLOW FOR DRUG REPURPOSING USING RNASeq



- index genome
- Align -> BAM files
- Check for strandedness
- Sort BAM by names

- raw count matrix

Unix command line

R studio

Racket

Fastq file format

```
@HWI-ST330:304:H045HADXX:1:1101:1111:61397
CACTTGTAAGGGCAGGCCCCCTTCACCCTCCCGCTCCTGGGGGANNNNNNNNNNANNNCGAGGCCCTGGGGTAGAGGGNNNNNNNNNNNNNNGATCTTGG
+
@?@DDDDDDHHH?GH:?FCBGGB@C?DBEGIIIIAEF;FCGGI##########################################################
```

Line	Description
1	Always begins with '@' and then information about the read
2	The actual DNA sequence
3	Always begins with a '+' and sometimes the same info in line 1
4	Has a string of characters which represent the quality scores; must have same number of characters as line 2

- 2 folders provided by Dr. Lessel (Germany)
  - CHAMP1 folder : HH17_1.fq.gz, HH17_2.fq.gz, HH18_1.fq.gz, HH18_2.fq.gz
  - CONTROL folder: HH07_1.fq.gz, HH07_2.fq.gz, HH19_1.fq.gz, HH19_2.fq.gz, HH27_1.fq.gz, HH27_2.fq.gz
- paired end
- unstranded

FASTQ quality encoding

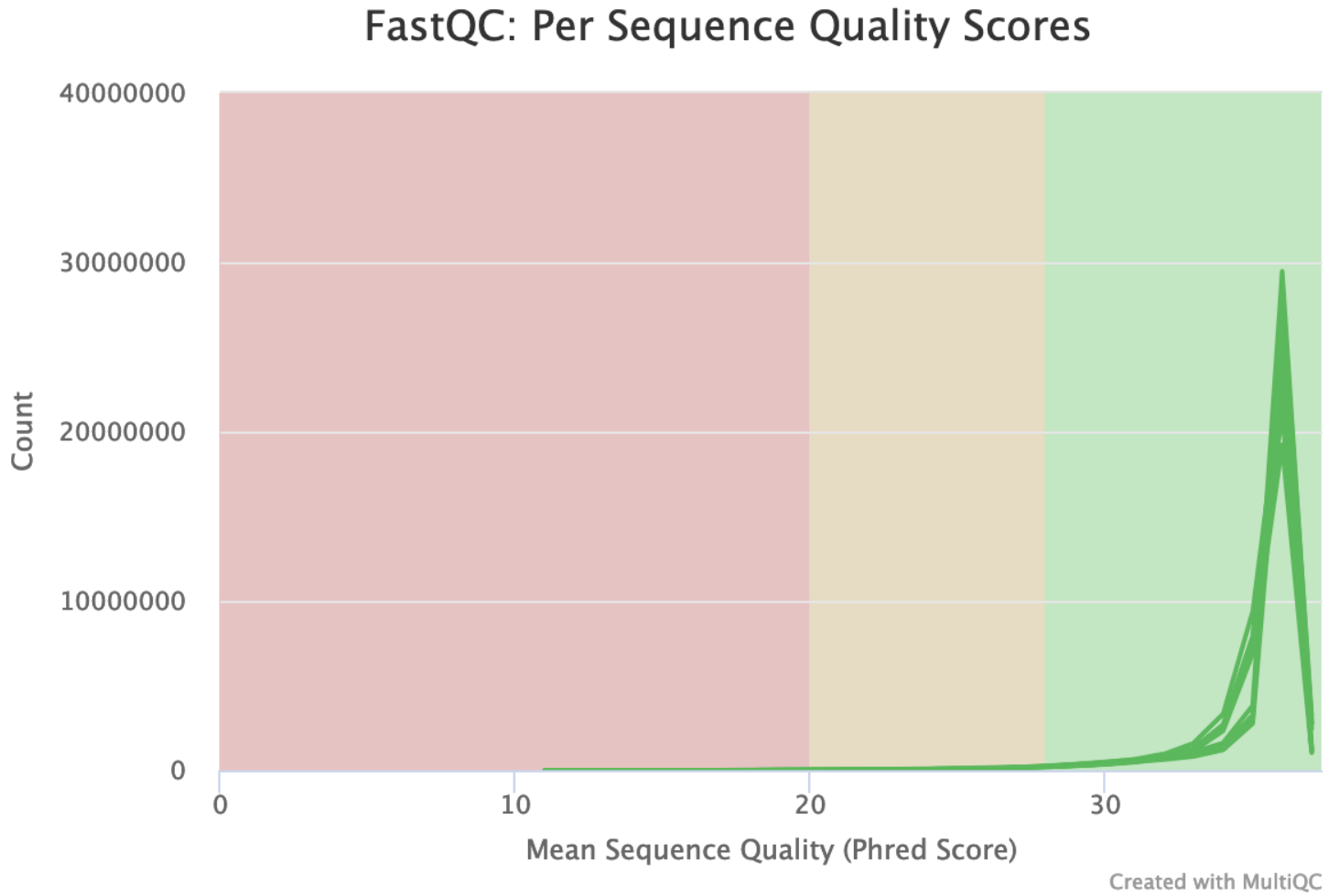
- Quality score is encoded by letters of the ASCII table
- Each quality score measures the probability that the nucleotide is correctly called

```
Q = -10 x log10(P), where P is the probability that a base call is erroneous
```

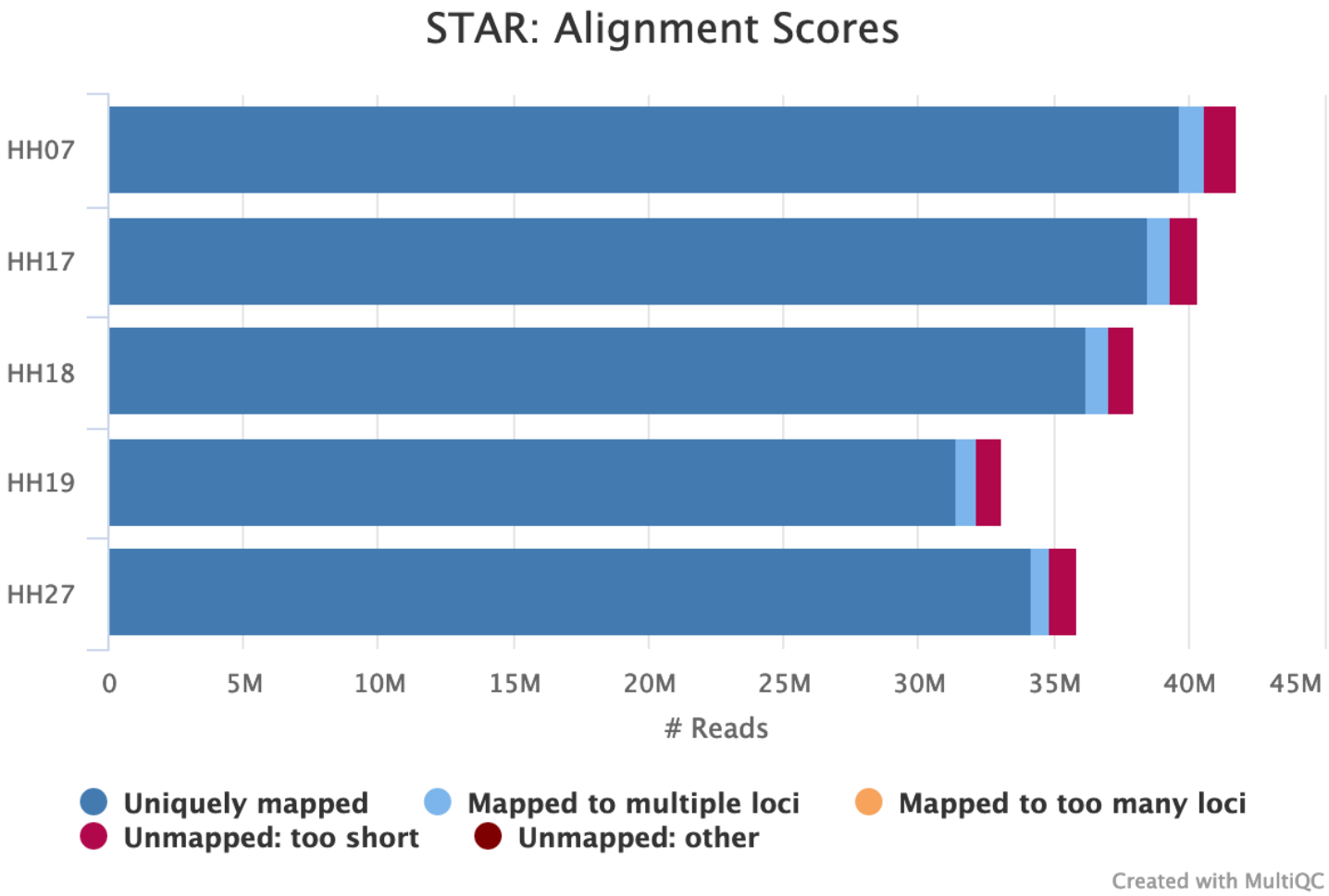
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

https://github.com/hbctraining/Intro-to-rnaseq-hpc-O2/blob/master/lessons/02_assessing_quality.md

Quality Control of Fastq files using FastQC

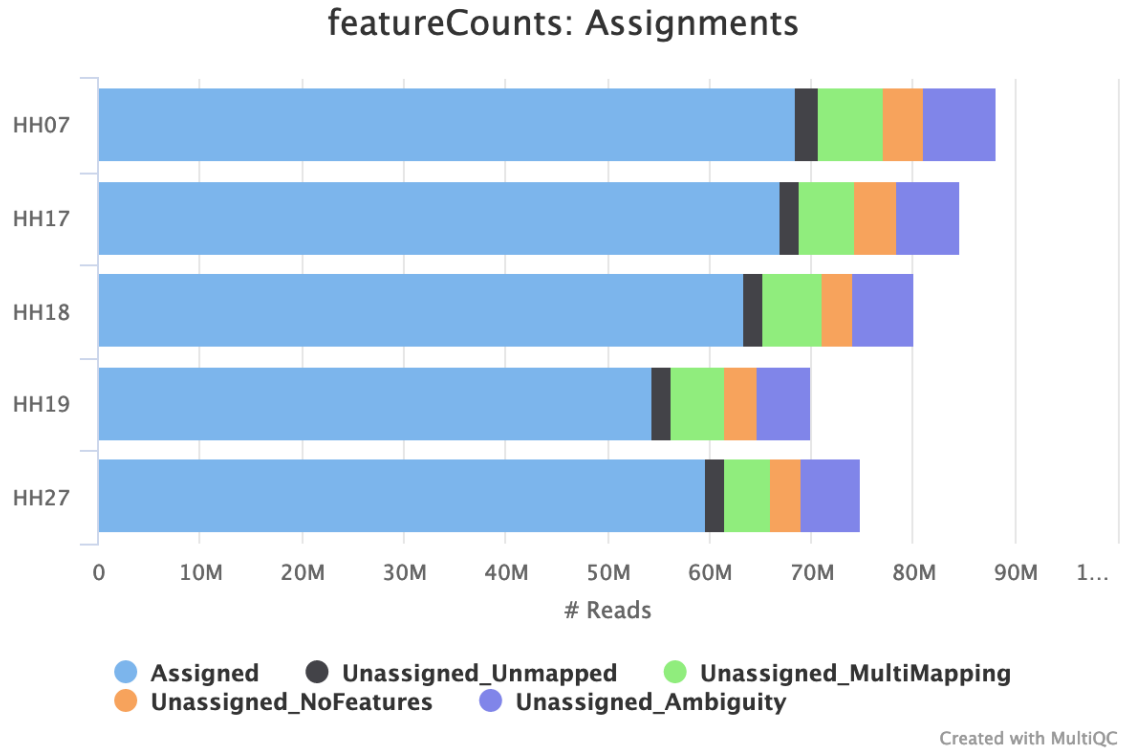


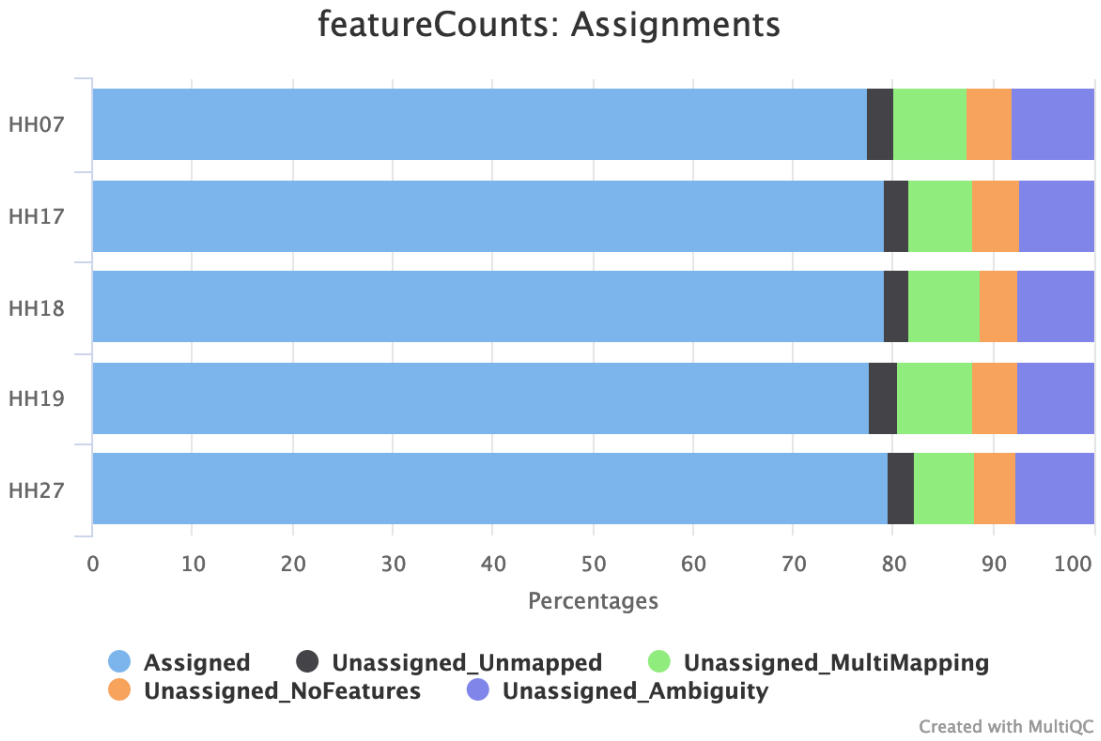
Read Alignment with STAR



Count Reads that are mapped to a gene with featureCounts

- **input:** multiple BAM + 1 GTF file
- **output:** raw count matrix (gene as rows and samples as columns)





Count Normalization by DESeq2

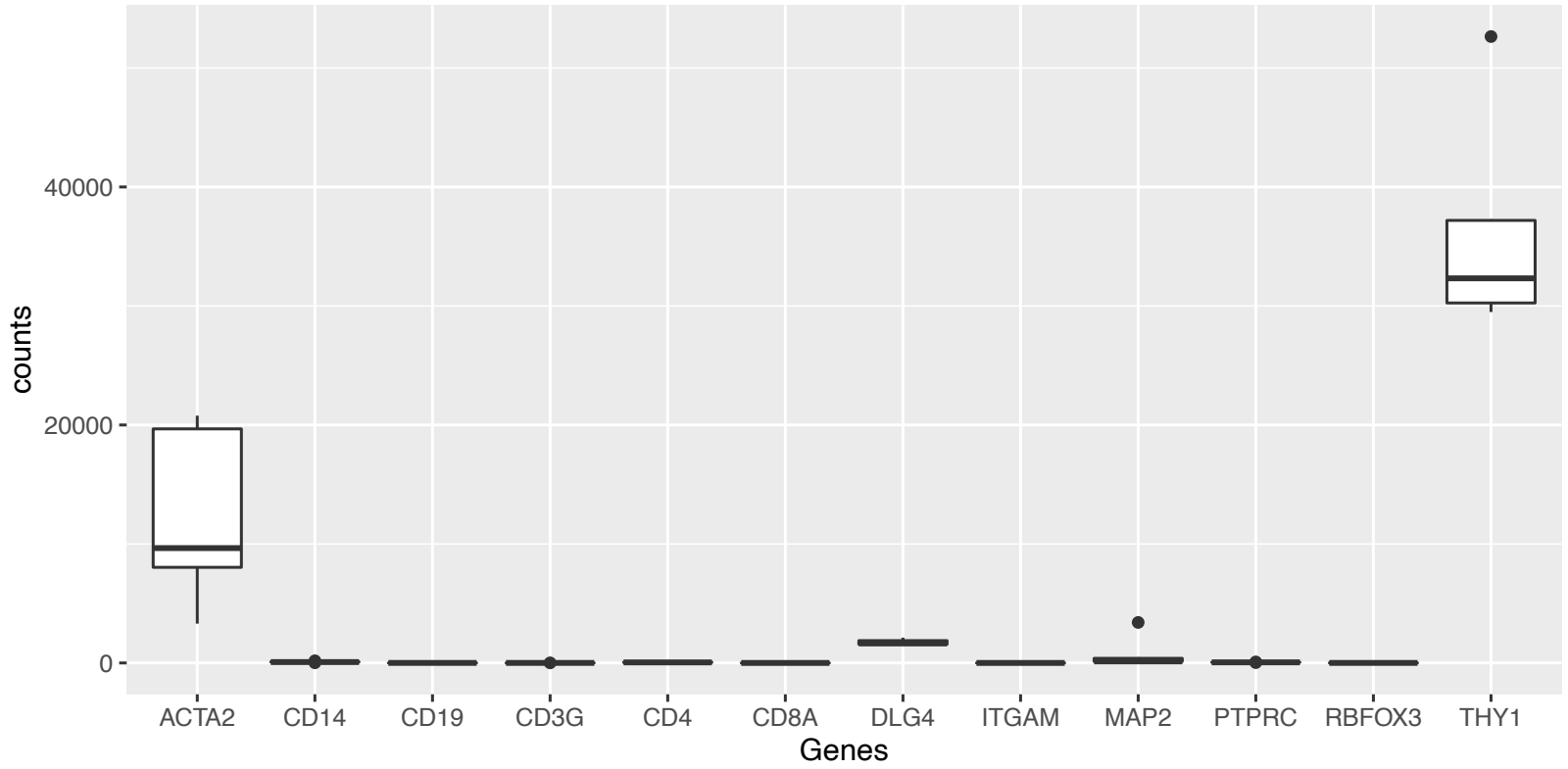
- Mean of Ratios method
- account for sequencing depth and RNA composition
- does not account for gene length (not necessary since we are comparing the same genes across # samples)

Procedure:

- Step 1: For each gene, a pseudo reference sample is created – geometric mean across all samples
- Step 2: Calculate ratio of each sample to the reference
- Step 3: Calculate the normalization factor for each sample (median of all ratios for each sample) => should be clos
- Step 3: Calculate the normalized count using the normalization factor :
    - for each gene in each sample, divide raw count by the normalization factor

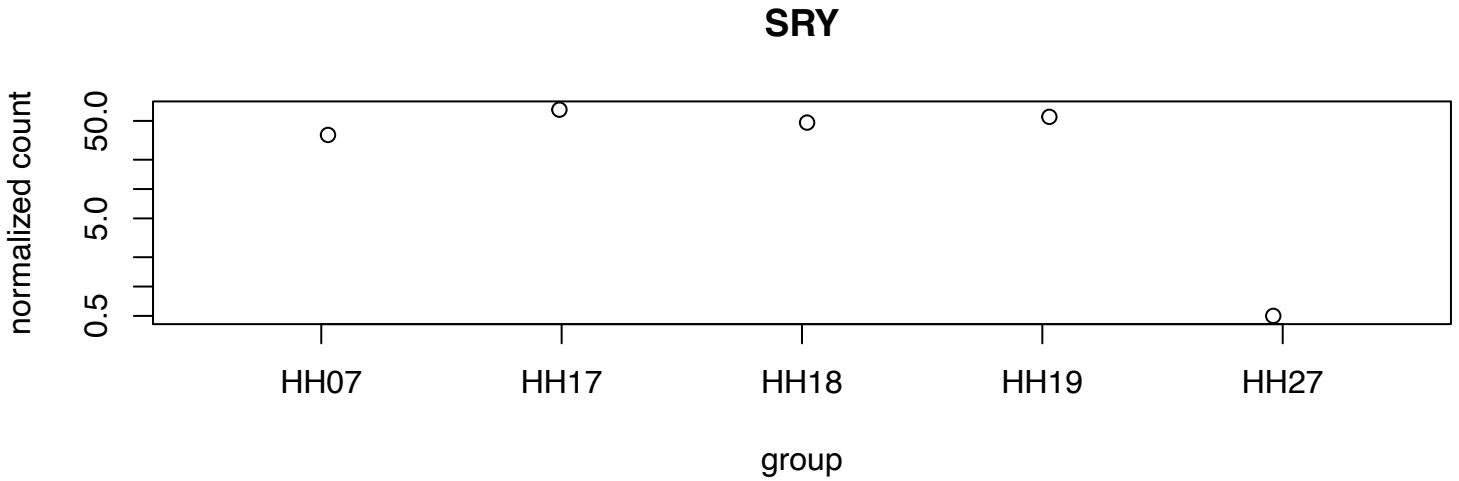
Check cell/tissue types

Given the bulk RNAseq data, can we infer what tissue/cell types of the original samples?



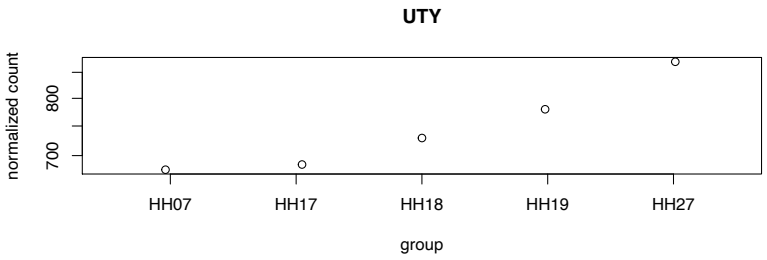
- Fibroblast: CD90 (THY1)
- Smooth muscle cell: Alpha-SMA (ACTA
- Leukocyte: PTPRC
- Monocyte: CD14
- Myeloid: ITGAM
- B cells: CD19
- T cells: CD3
- Helper T cells: CD4
- Cytotoxic T cells: CD8
- Neurons: DLG4, MAP3, RBFOX3

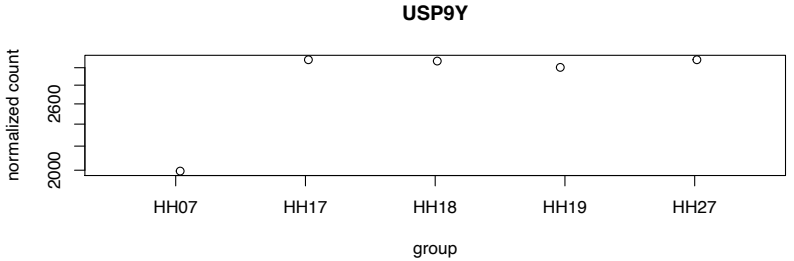
Check sample sexes

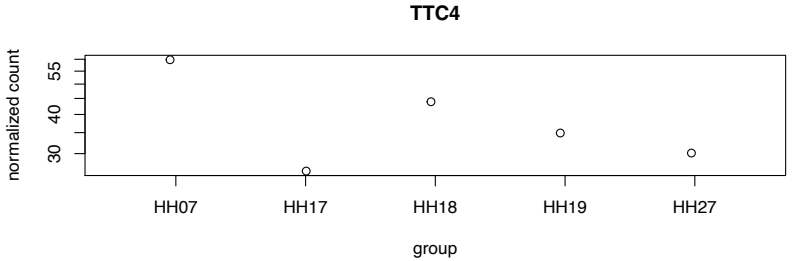


- **CHAMP1**: HH17 and HH18
- **Control**: HH07, HH19 and HH27

- SRY gene is crucial for the development of a fetus into a male
- ➢ all samples are male, except for HH27 (female)







- many genes on the Y chromosome are actually present in both X and Y (pseudoautosomal regions)
- pseudoautosomal regions are where X and Y chromosome pair + crossover
- inherited just like autosomal genes

Check CHAMP1 variant types

Check CHAMP1 variant types

Variant details

NM_032436.4(CHAMP1):c.1192C>T (p.Arg398Ter)

Allele ID:
204638

Variant type:
single nucleotide variant

Variant length:
1 bp

Cytogenetic location:
13q34

Genomic location:
13: 114325034 (GRCh38) GRCh38 UCSC
13: 115090509 (GRCh37) GRCh37 UCSC

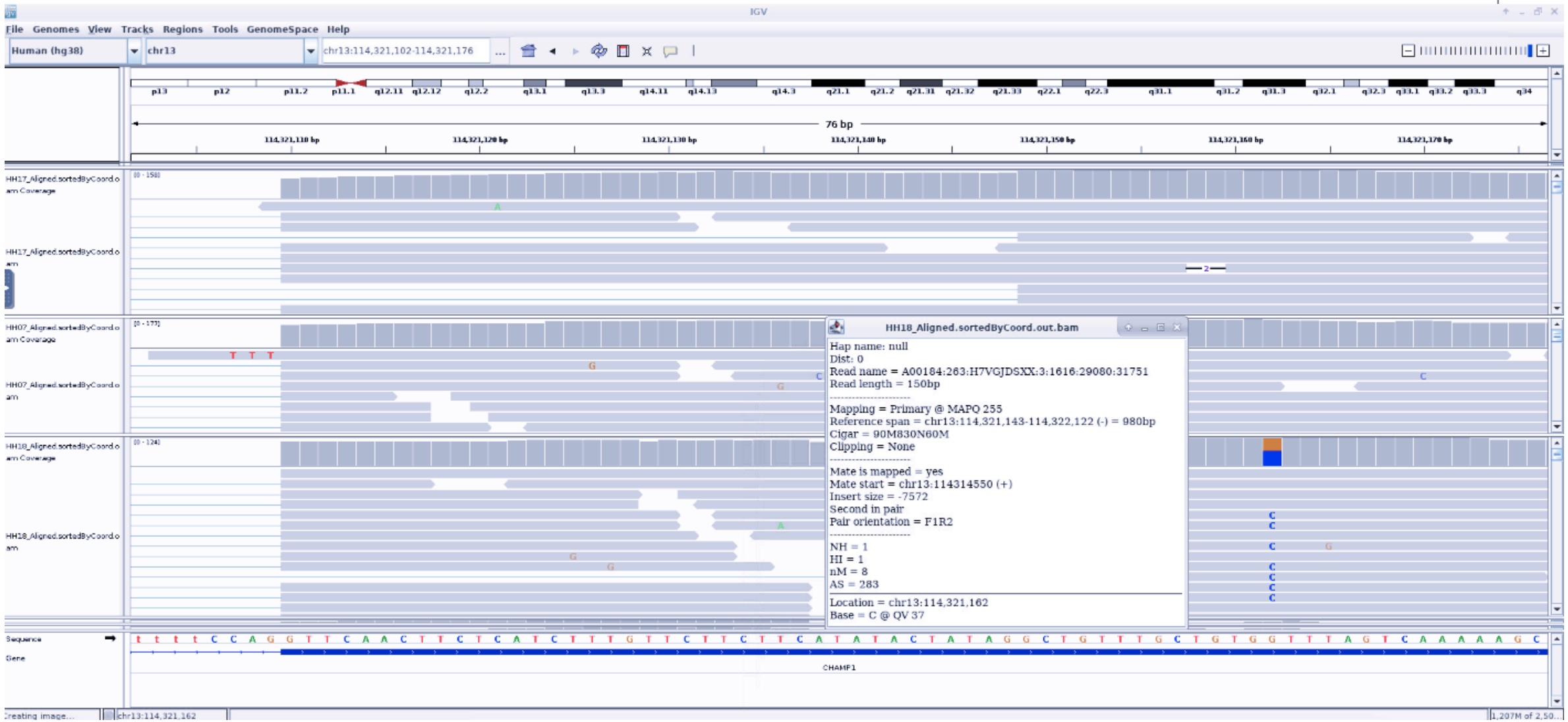
HGVS:

Nucleotide	Protein	Molecular consequence
NC_000013.10:g.115090509C>T		
NC_000013.11:g.114325034C>T		
NM_032436.4:c.1192C>T MANE SELECT	NP_115812.1:p.Arg398Ter	nonsense

... more HGVS

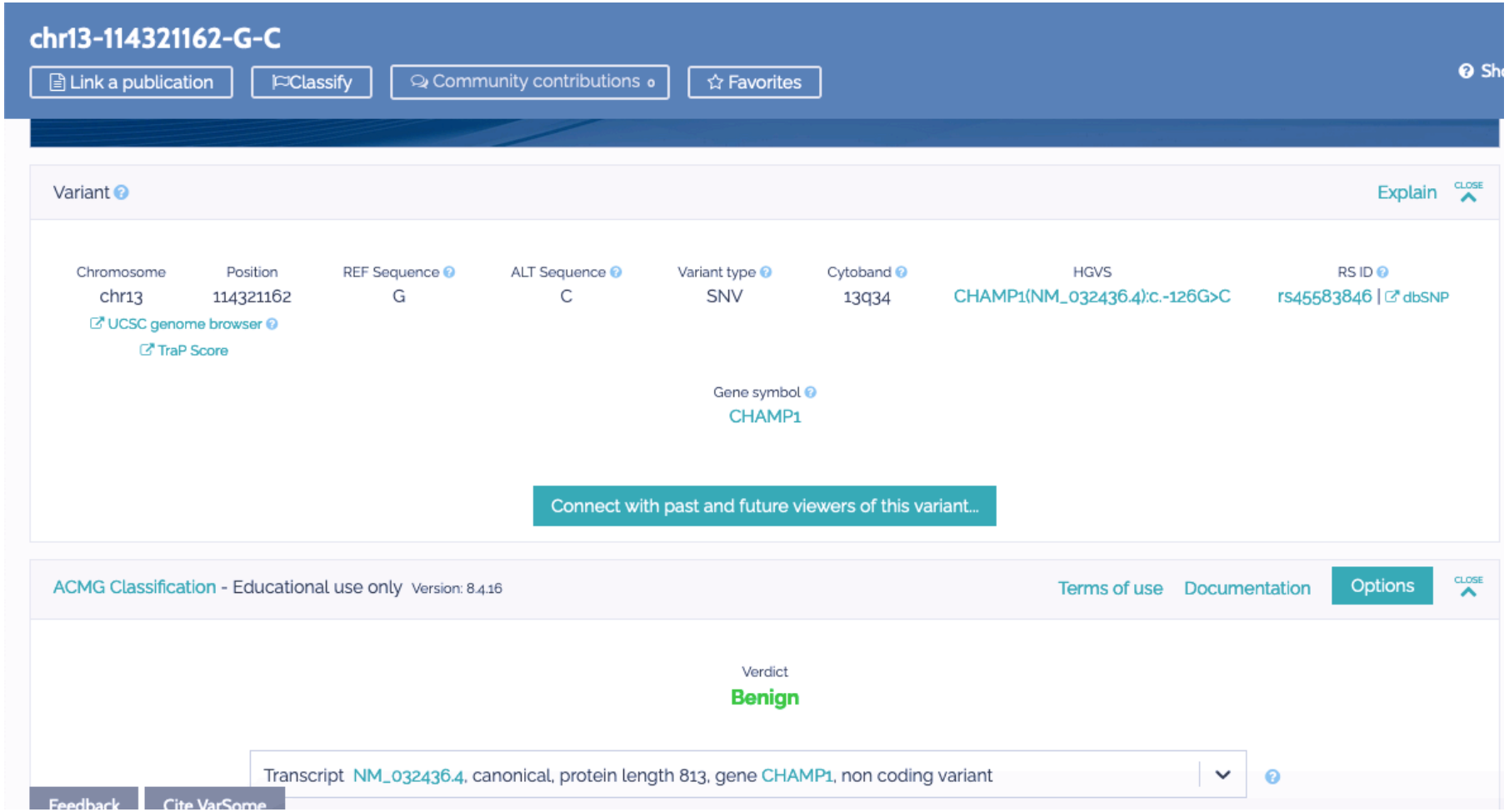
- HH17 (CHAMP1) contains the CHAMP1 mutation that is annotated to be pathogenic
- Clingen: some evidence for dosage sensitivity
- likely to haploinsufficient

Check CHAMP1 samples' variant types

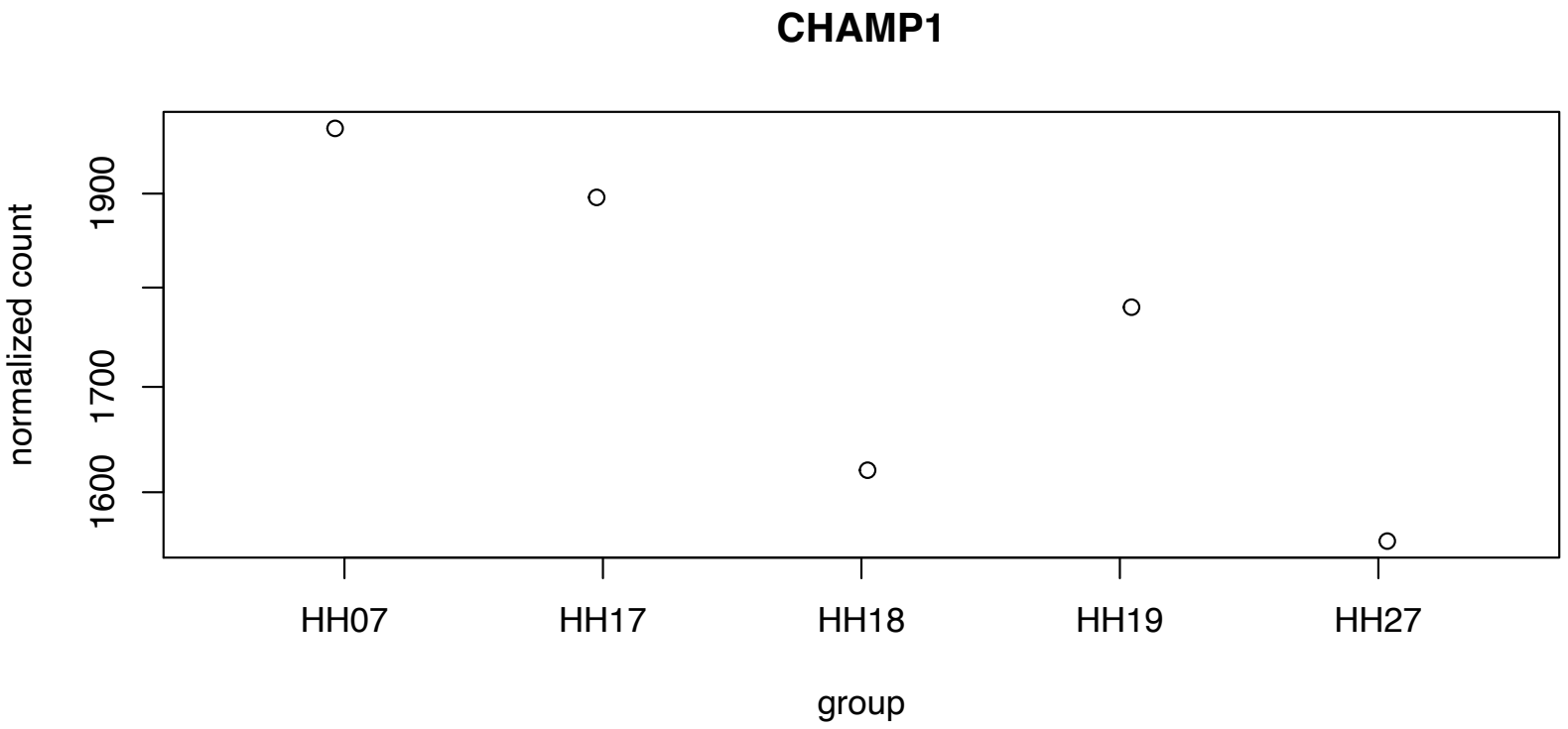


Check CHAMP1 variant types-HH18

- point mutation in exon 2
- non-coding variant



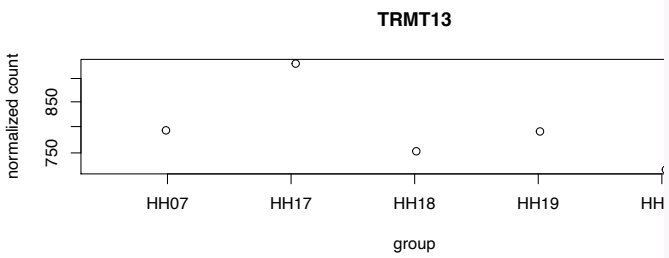
Check CHAMP1 gene expression and interacting partners

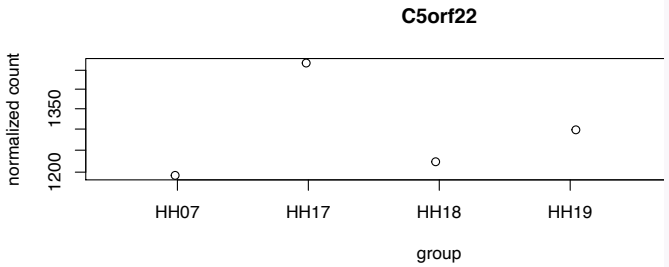


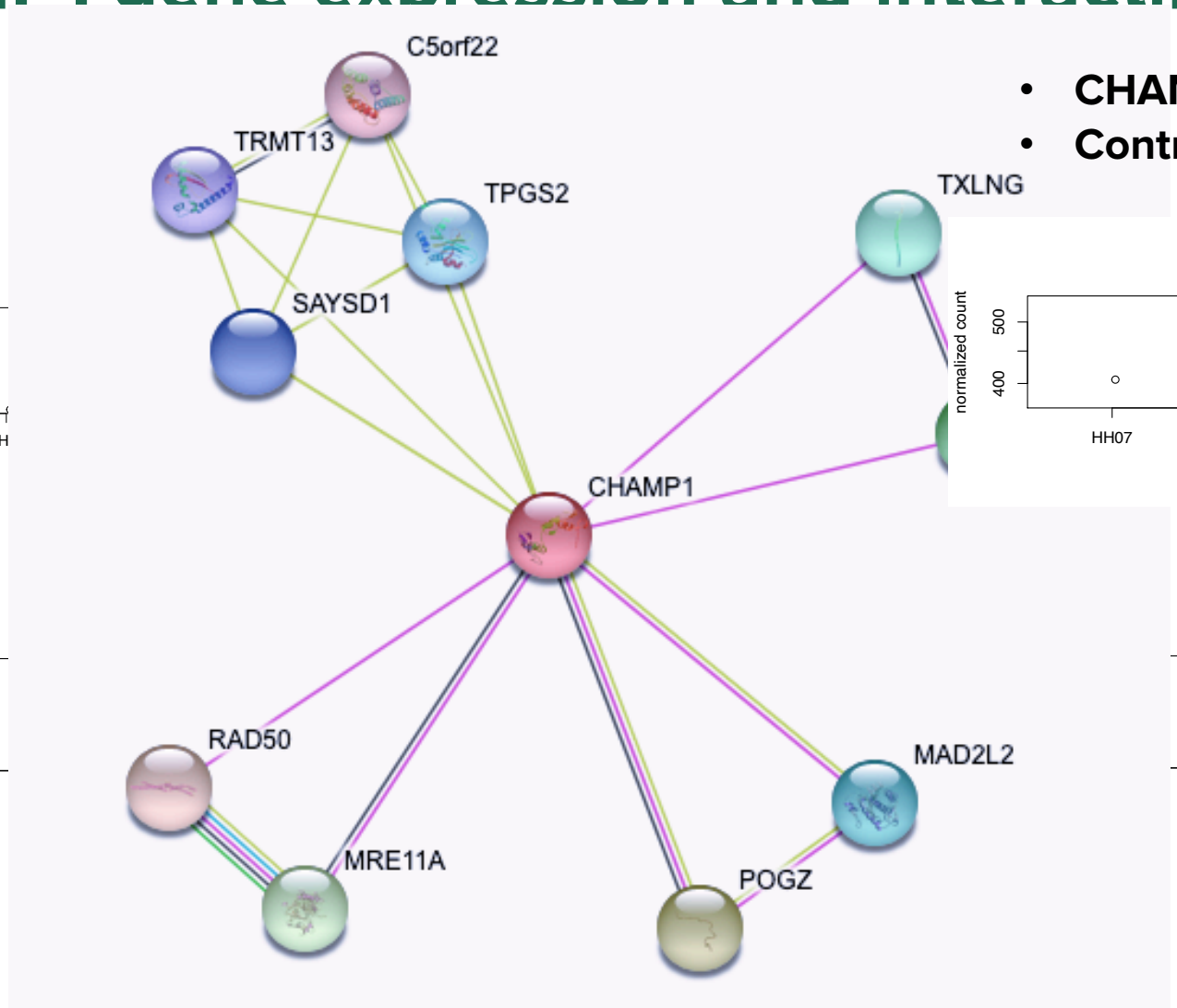
- **CHAMP1**: HH17 and HH18
- **Control**: HH07, HH19 and HH27

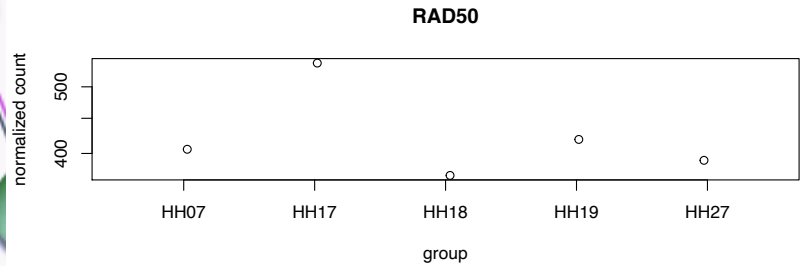
Check CHAMP1 gene expression and interacting partners

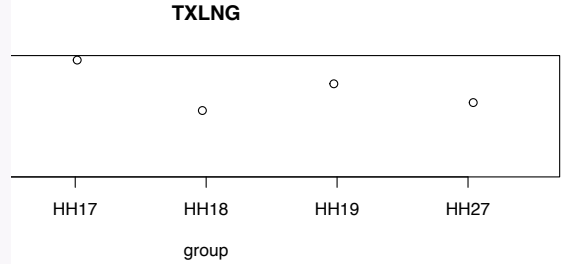
- **CHAMP1**: HH17 and HH18
- **Control**: HH07, HH19 and HH27









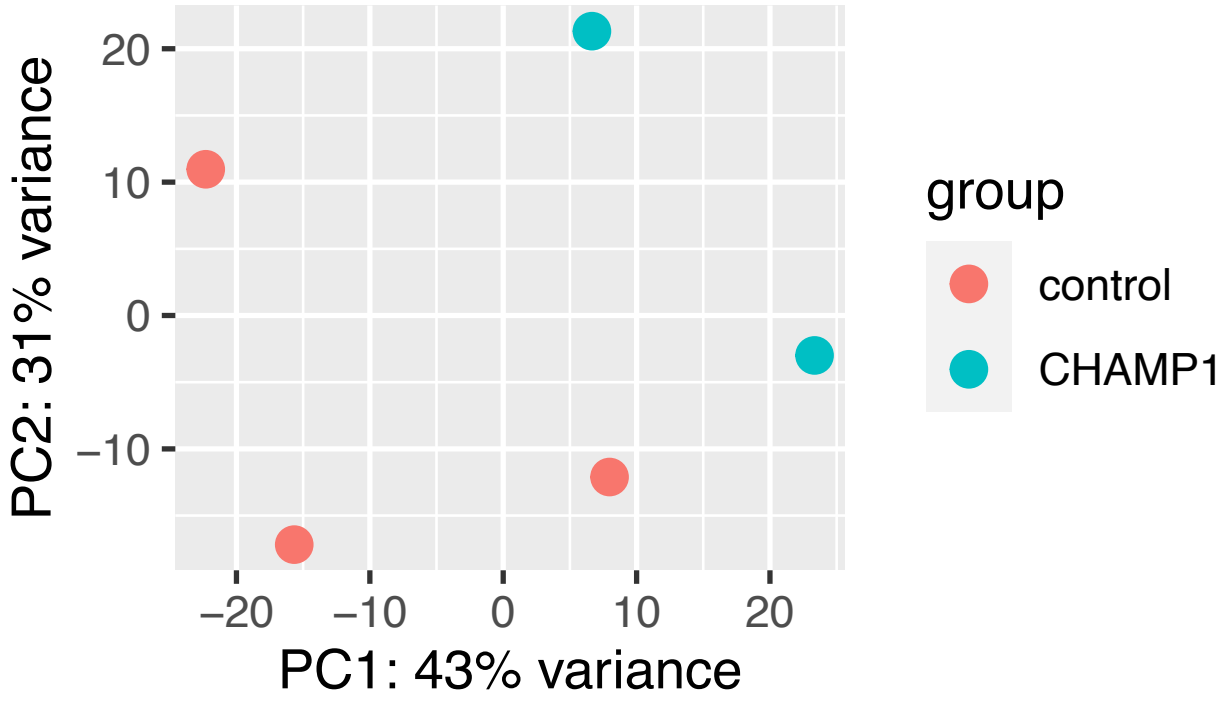


- **CHAMP1**: HH17 and HH18
- **Control**: HH07, HH19 and HH27

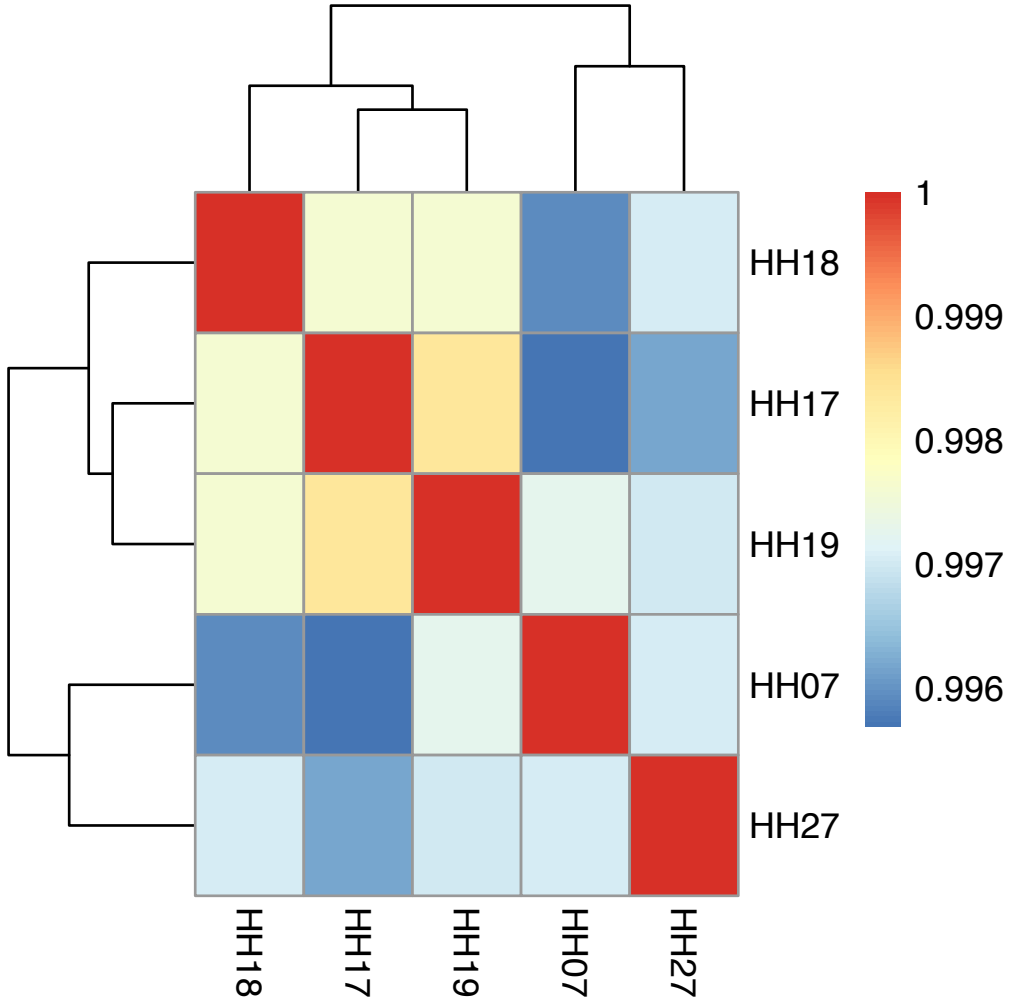
Exploring the data

- **CHAMP1**: HH17 and HH18
- **Control**: HH07, HH19 and HH27

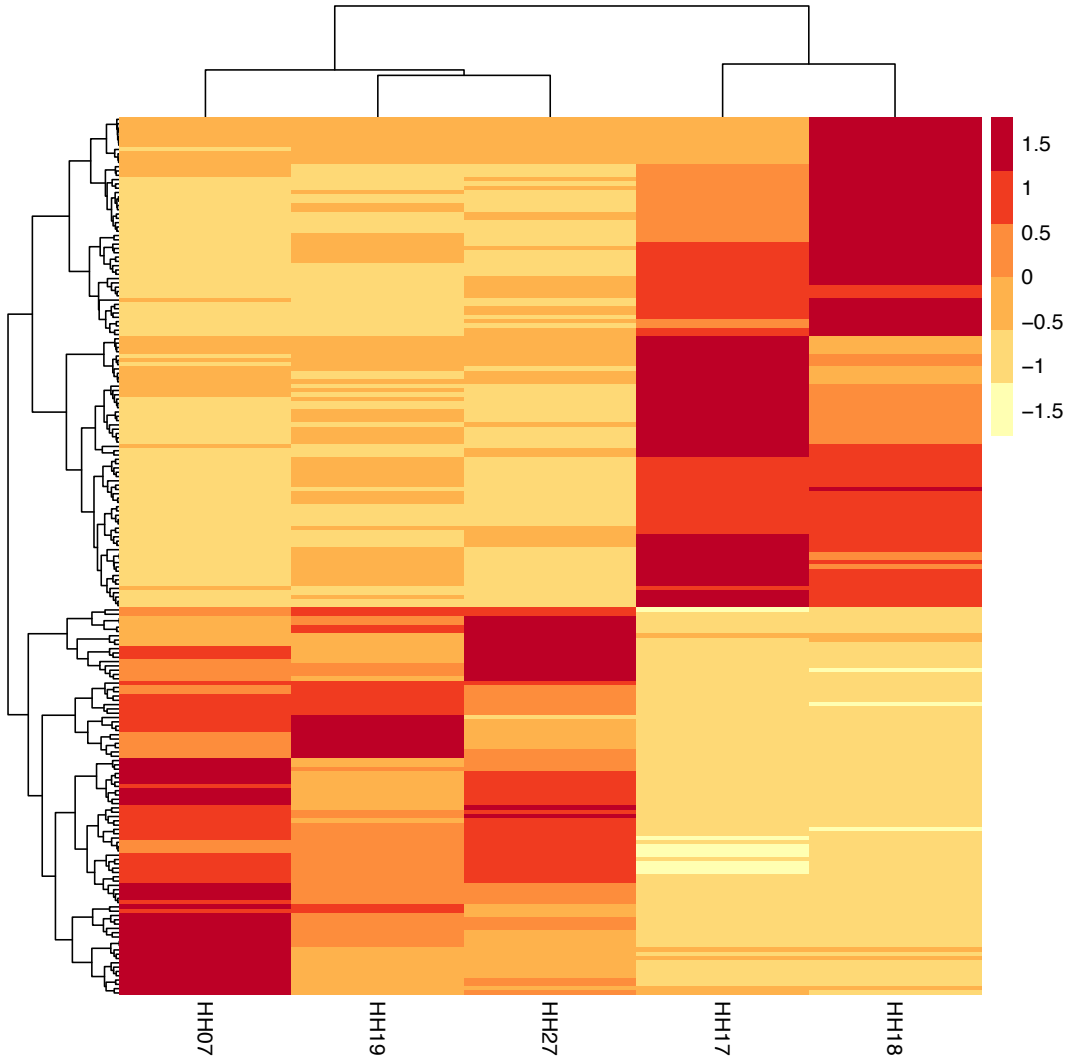
Principle Component Analysis (PCA)



Unsupervised Hierarchical Clustering



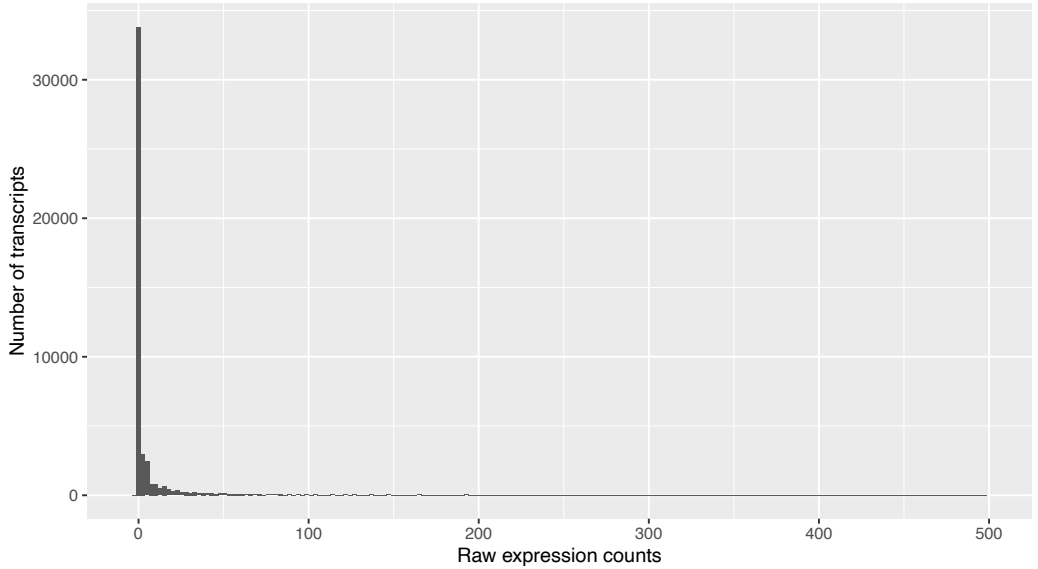
Clustering based on DEG genes



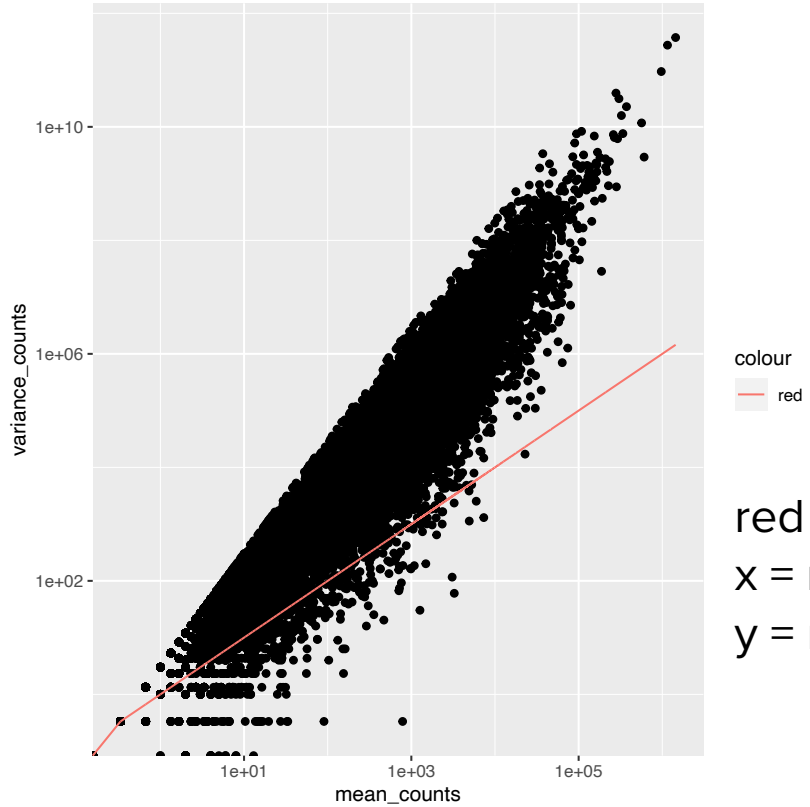
Differential gene expression analysis with DESeq2

- **null hypothesis:** is the log fold change between disease and control is zero?
- DEseq2 model the raw counts using the negative binomial distribution to perform statistical inferences on the differences

count distribution in HH07



mean vs variance (3 controls)



red line
x = mean_counts
y = mean_counts

http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#methods-changes-since-the-2014-deseq2-paper

Differential gene expression analysis with DESeq2

DESeq2 pipeline:

Model raw counts for each genes:

- Estimate size factors
- Estimate gene-wise dispersion
- Fit curve to gene-wise dispersion estimates
- Shrink gene-wise dispersion estimates
- GLM fit for each gene

Shrink log2 fold changes

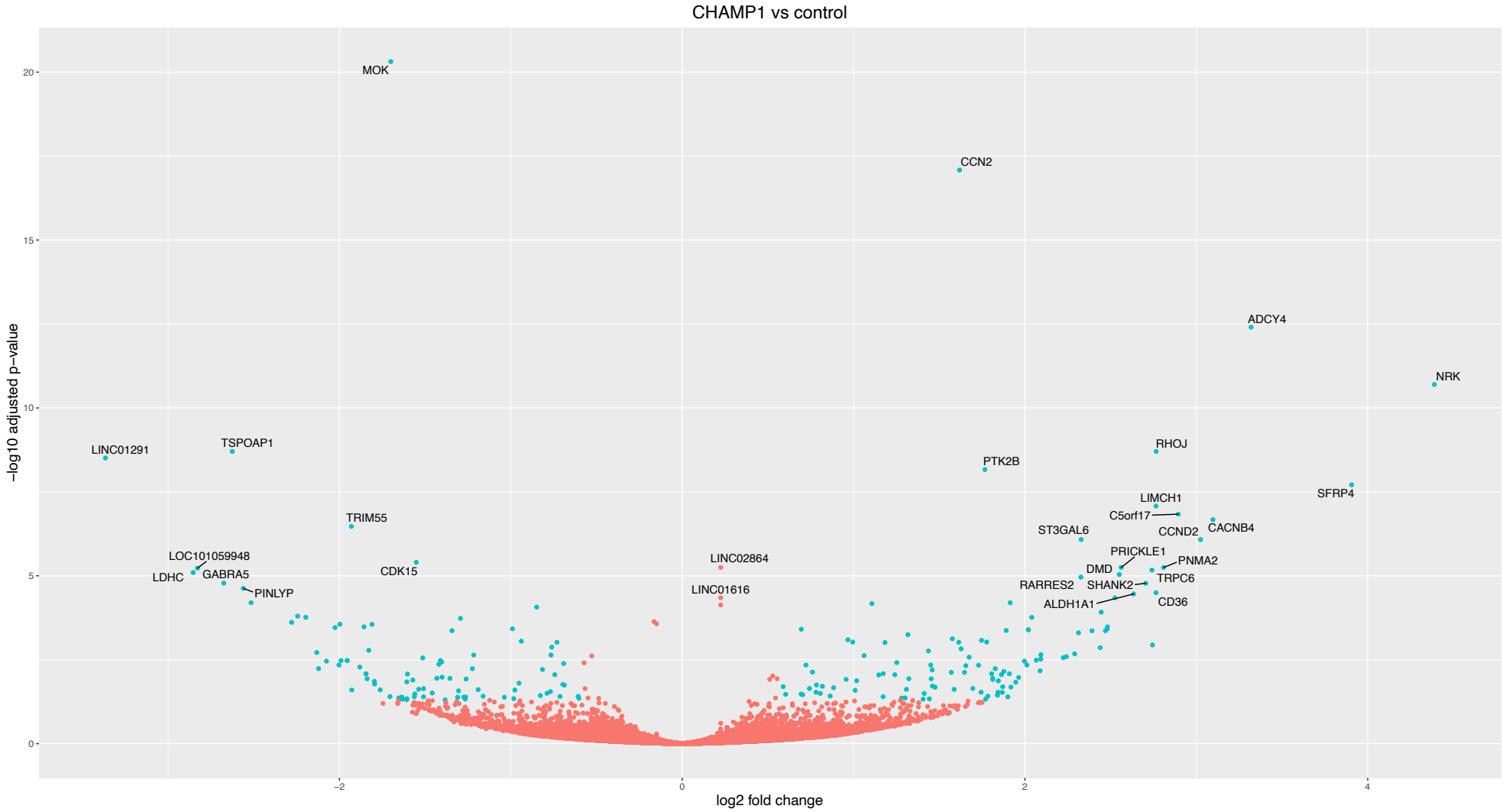
Test for differential expression

http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#methods-changes-since-the-2014-deseq2-paper

CHAMP1 vs control
All Differentially Expressed (DE) genes

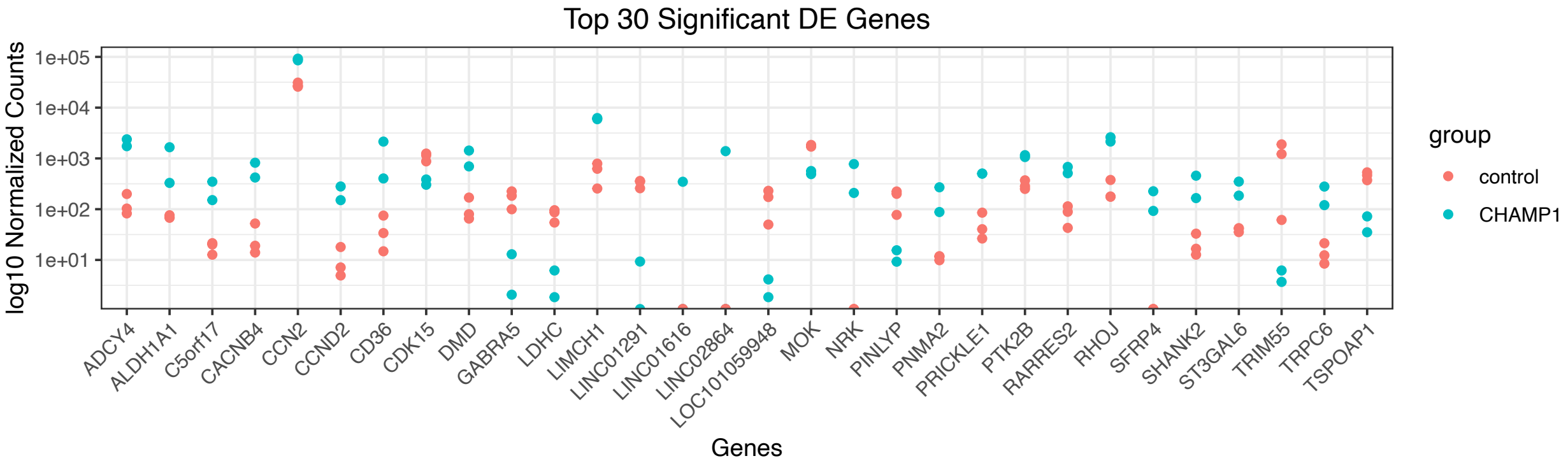
- padj < 0.05

& |log2FoldChange| >= 0.58



Top 30 genes with the highest log2Foldchange are labeled

Top 30 significant DE genes



What May Cause the Differential Gene Expressions?

- Variations due to DNA codes:
  - different types of CHAMP1 variants
  - other variants than CHAMP1
- Demographics: age, sex, ethnicity
- Different cell types express different genes
- Cell culture conditions;
  - ❖ cell passage numbers
  - ❖ media composition (base media, types/lots/batches of serum, growth factors, vitamins, amino acids, antibiotics, antifungal)
  - ❖ incubator conditions: temperature, levels of CO_2, H_2O
  - ❖ cell plating density
  - ❖ contamination (mycoplasma, bacteria, virus, fungus)
  - ❖ time collected (circadian rhythm)
- RNA extraction, library preparation and sequencing methods

Thoughts

- Do CHAMP1 patients cultured fibroblast an appropriate model for the disease?
- Can we mine patients' RNA sequencing data for characterizing molecular dysregulations in the disease?
- Do we have enough molecular evidence for drug recommendations?
- How to efficiently use mediKanren for drug-repurposing?

3 drug-repurposing strategies using mediKanren biomedical reasoner

Strategy 1:

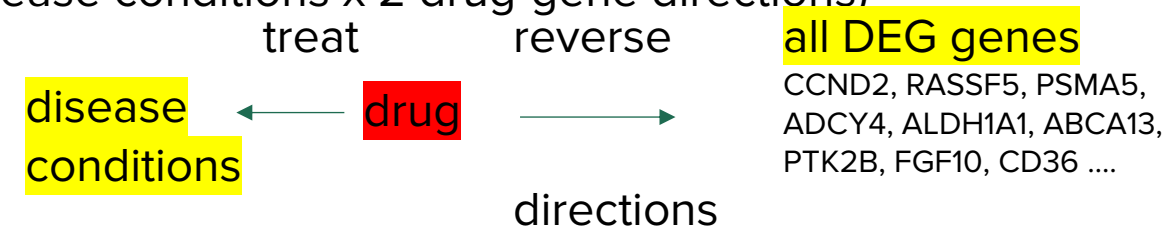
- **Step 1:** GSEA (ClusterProfiler) -> enriched functional categories -> significantly altered genes
- **Step 2:** query/graph to find drugs that reverse each enriched gene in the desired direction
- result in 2 lists (up-regulated vs down regulated genes)

Strategy 2:

- **Step 1:** Of all the human genes, using run/graph select the ones that are related to CHAMP1 disease conditions (n =5 hypotonia, cerebral palsy, autism, development delay and epilepsy)
- **Step 2:** Plot differential expression on these disease-relevant genes only
- **Step 3:** query/graph to find drugs that reverse each of these genes in the desired direction
- result in 2 lists (up-regulated vs down regulated genes)

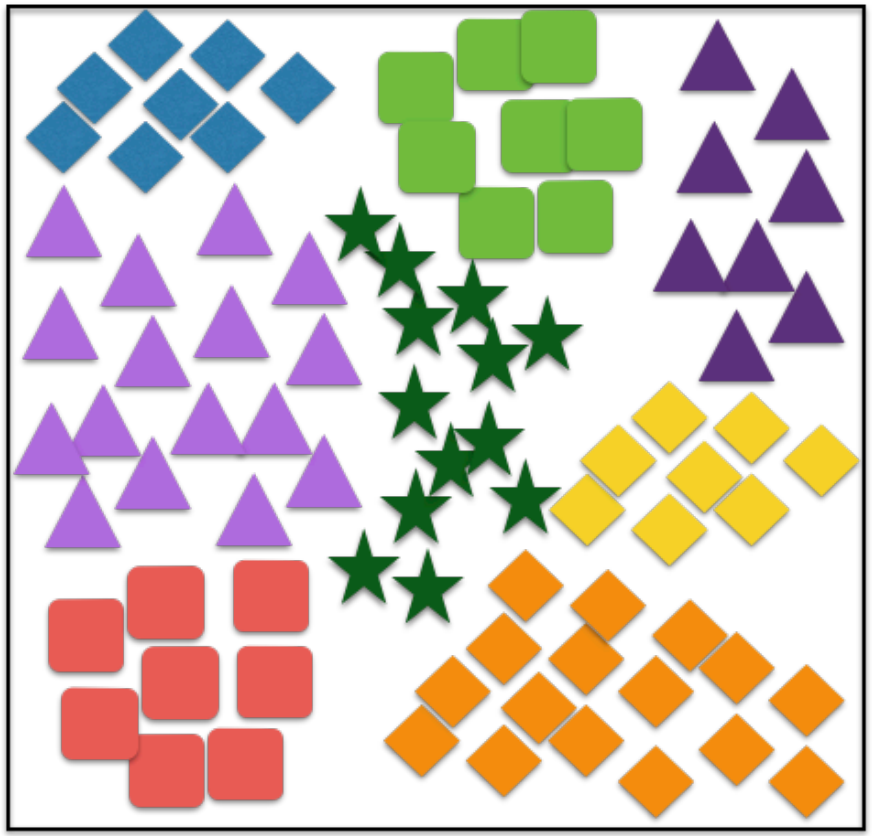
Strategy 3:

- 2-hop queries to find drugs that treat CHAMP1 disease conditions (n=5) and reverse the gene expression directions for all significant DE genes
- result in 10 lists (5 disease conditions x 2 drug-gene directions)

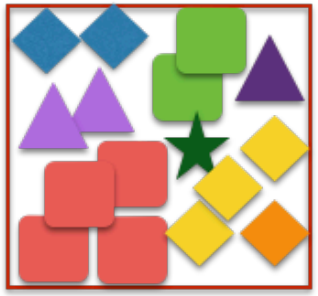


Gene Set Enrichment Analysis (GSEA)

All known genes in a species
(categorized into groups)



GO
KEGG
reactome
Wikipathways



DEGs

https://github.com/hbctraining/DGE_workshop/blob/master/lessons/09_functional_analysis.md

Gene Set Enrichment Analysis (GSEA) by Over Representation Analysis

Genes categories	Organism-specific Background	DE results	Over-represented?
Functional category 1	35/13000	25/1000	Likely
Functional category 2	56/13000	4/1000	Unlikely
Functional category 3	90/13000	8/1000	Unlikely
Functional category 4	15/13000	10/1000	Likely
…			
…			

Hypergeometry testing:
e.g. : What is the probability of 25 genes (k) being associated with Functional category 1 for all genes in the gene list (n=1000), from a population of all genes in the entire genome (N=13000) which contains 35 genes (K) associated with Functional Category 1?

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

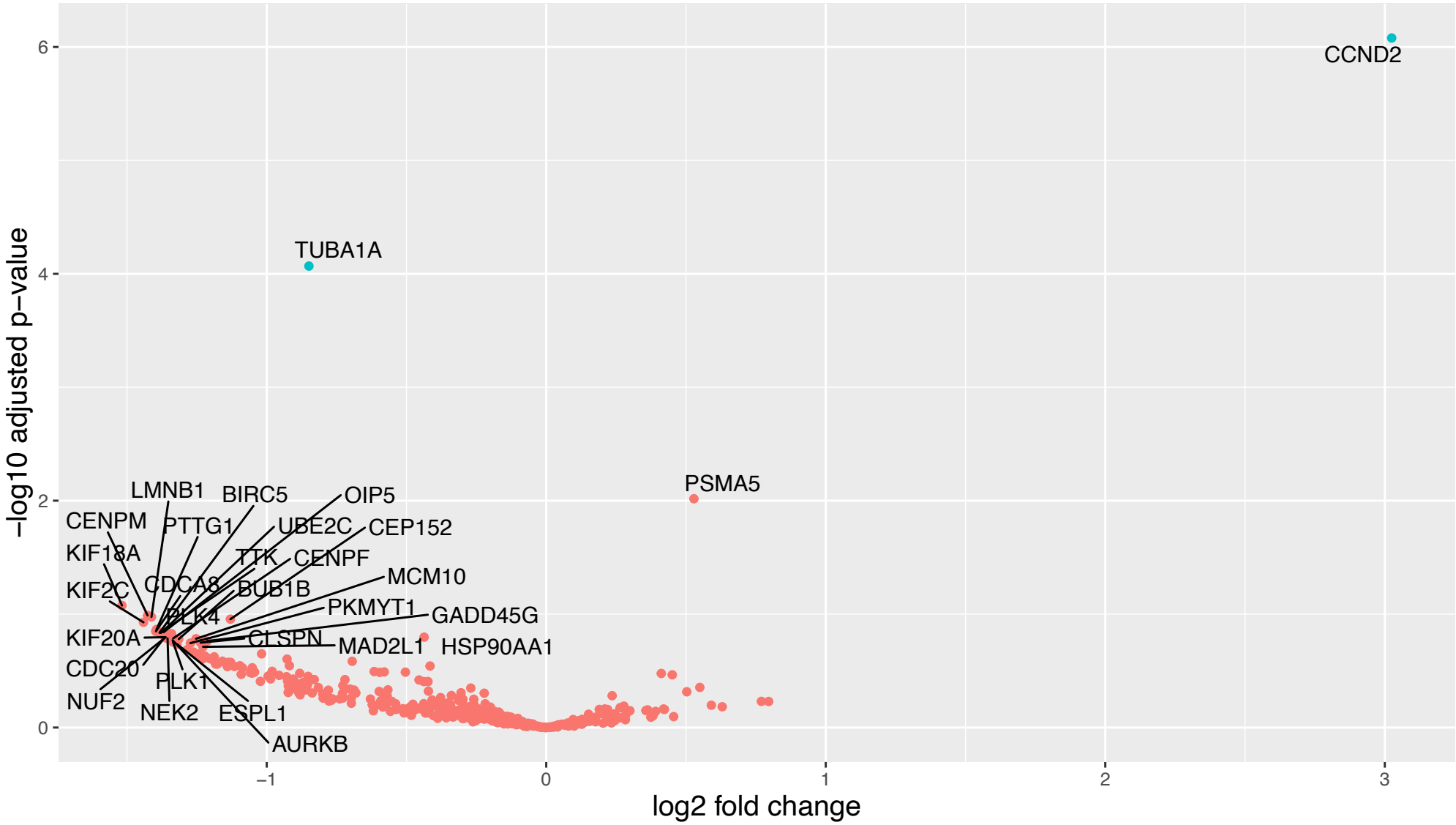
https://github.com/hbctraining/DGE_workshop/blob/master/lessons/09_functional_analysis.md

CHAMP1 GSEA (Geneset Enrichment analyses) with ClusterProfiler using KEGG database

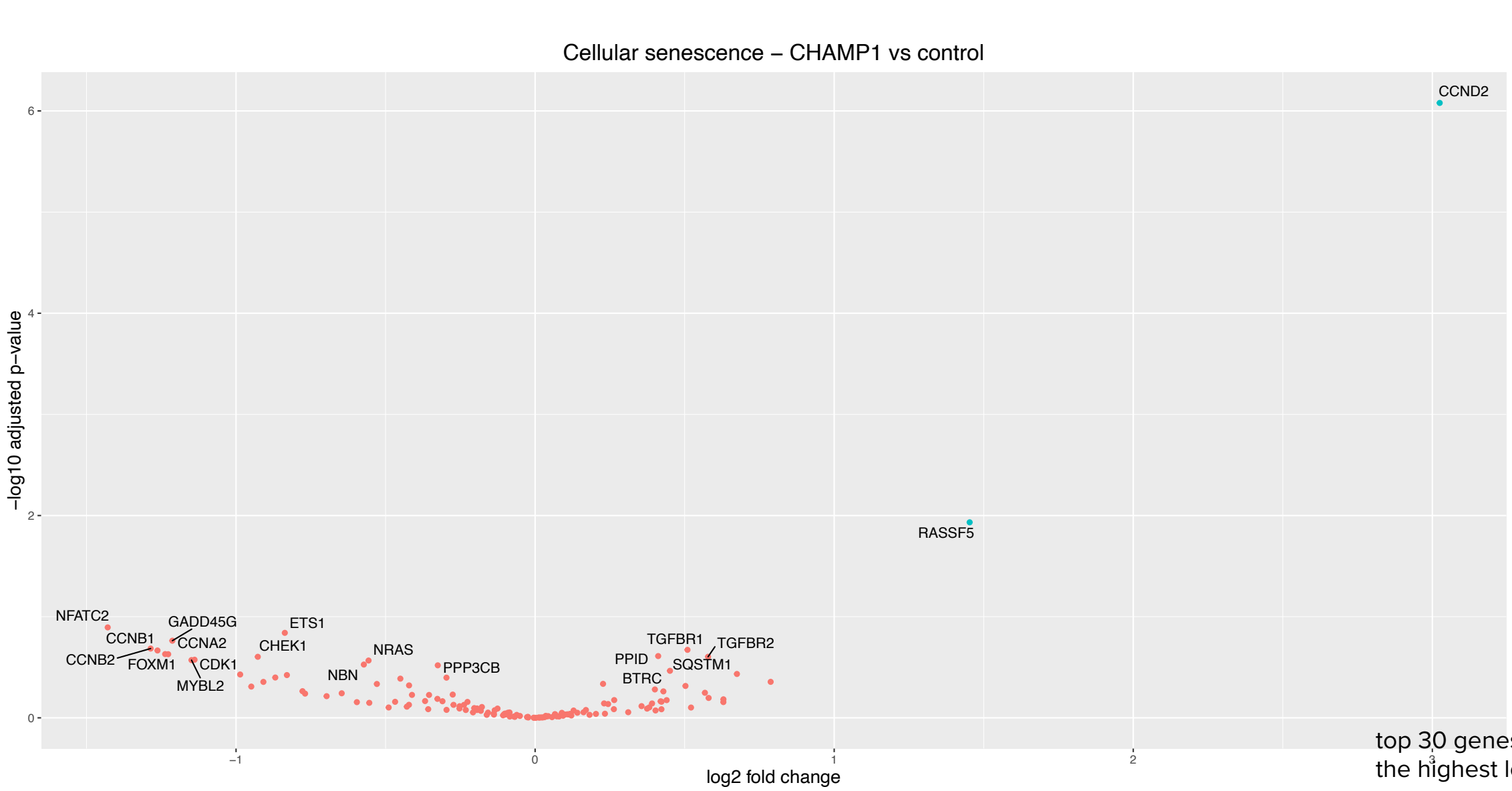
ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalues	rank	leading_edge	core_enrichment	
hsa04110	Cell cycle	123	-0.591859332	-2.333165629	1.00E-10	3.31E-08	2.89E-08	2189	tags=39%	list=12%	signal=35%
hsa03040	Spliceosome	140	-0.542501179	-2.174777645	1.13E-09	1.86E-07	1.63E-07	5145	tags=59%	list=29%	signal=42%
hsa03030	DNA replication	36	-0.75091223	-2.373702545	1.11E-08	1.22E-06	1.07E-06	1699	tags=61%	list=9%	signal=55%
hsa03460	Fanconi anemia pathway	52	-0.660638601	-2.246700622	2.85E-07	2.36E-05	2.06E-05	2714	tags=46%	list=15%	signal=39%
hsa04114	Oocyte meiosis	116	-0.48090957	-1.865461477	2.27E-05	0.00150306	0.00131449	2402	tags=30%	list=13%	signal=26%
hsa03008	Ribosome biogenesis in eukaryotes	76	-0.50541597	-1.834428155	0.00026437	0.01458464	0.01275489	5154	tags=58%	list=29%	signal=41%
hsa03430	Mismatch repair	23	-0.651077434	-1.831558114	0.00072275	0.02799541	0.02448318	2638	tags=52%	list=15%	signal=45%
hsa03013	RNA transport	165	-0.390553066	-1.601624364	0.00073485	0.02799541	0.02448318	5400	tags=48%	list=30%	signal=34%
hsa00982	Drug metabolism - cytochrome P450	42	0.562747499	1.812782865	0.00083266	0.02799541	0.02448318	3263	tags=50%	list=18%	signal=41%
hsa00830	Retinol metabolism	37	0.584161514	1.843423759	0.00084578	0.02799541	0.02448318	3263	tags=49%	list=18%	signal=40%
hsa03440	Homologous recombination	41	-0.563358633	-1.840232703	0.00165731	0.04981289	0.0435635	3368	tags=49%	list=19%	signal=40%
hsa02010	ABC transporters	42	0.541673078	1.744895672	0.00180591	0.04981289	0.0435635	3486	tags=52%	list=19%	signal=42%
hsa03410	Base excision repair	32	-0.583856754	-1.802135059	0.00209794	0.05341686	0.04671533	4114	tags=59%	list=23%	signal=46%
hsa04914	Progesterone-mediated oocyte maturation	92	-0.428568224	-1.61935026	0.00255145	0.05754889	0.05032897	2700	tags=27%	list=15%	signal=23%
hsa05170	Human immunodeficiency virus 1 infection	188	-0.361097354	-1.508041707	0.00260796	0.05754889	0.05032897	3481	tags=28%	list=19%	signal=23%
hsa05014	Amyotrophic lateral sclerosis	320	-0.318571203	-1.400396386	0.00296064	0.06124817	0.05356415	4134	tags=34%	list=23%	signal=27%
hsa03050	Proteasome	44	-0.52408741	-1.710934488	0.00419838	0.07968796	0.06969054	6186	tags=73%	list=34%	signal=48%
hsa05016	Huntington disease	264	-0.318619755	-1.370353869	0.00433348	0.07968796	0.06969054	4892	tags=37%	list=27%	signal=27%
hsa05206	MicroRNAs in cancer	199	-0.3428597	-1.442117674	0.00501398	0.08734877	0.07639025	3113	tags=27%	list=17%	signal=22%
hsa04810	Regulation of actin cytoskeleton	204	-0.342669029	-1.439608693	0.00577197	0.09552607	0.08354164	3783	tags=34%	list=21%	signal=27%
hsa05166	Human T-cell leukemia virus 1 infection	206	-0.339558908	-1.431292192	0.00624742	0.09847122	0.0861173	2429	tags=23%	list=13%	signal=20%
hsa05214	Glioma	72	-0.437003794	-1.570812596	0.00676594	0.10089914	0.08824062	4833	tags=40%	list=27%	signal=30%
hsa04929	GnRH secretion	59	-0.461321444	-1.611517856	0.00701112	0.10089914	0.08824062	1827	tags=20%	list=10%	signal=18%
hsa04512	ECM-receptor interaction	79	0.425126906	1.582537975	0.00875749	0.11770164	0.10293512	1431	tags=25%	list=8%	signal=23%
hsa04722	Neurotrophin signaling pathway	115	-0.383822151	-1.488439123	0.00923066	0.11770164	0.10293512	4833	tags=37%	list=27%	signal=28%
hsa04080	Neuroactive ligand-receptor interaction	234	-0.323300666	-1.381778656	0.00924545	0.11770164	0.10293512	1917	tags=19%	list=11%	signal=17%
hsa04218	Cellular senescence	149	-0.35027821	-1.413650114	0.0105832	0.12974222	0.11346514	2189	tags=23%	list=12%	signal=20%
hsa03015	mRNA surveillance pathway	90	-0.40246315	-1.508407989	0.01300936	0.15378923	0.13449527	4917	tags=41%	list=27%	signal=30%
hsa04020	Calcium signaling pathway	175	0.329177493	1.366605045	0.01668392	0.19042679	0.16653639	2348	tags=22%	list=13%	signal=20%

padj < 0.05 & |log2FoldChange| >= 0.58

cell cycle genes in CHAMP1 vs control

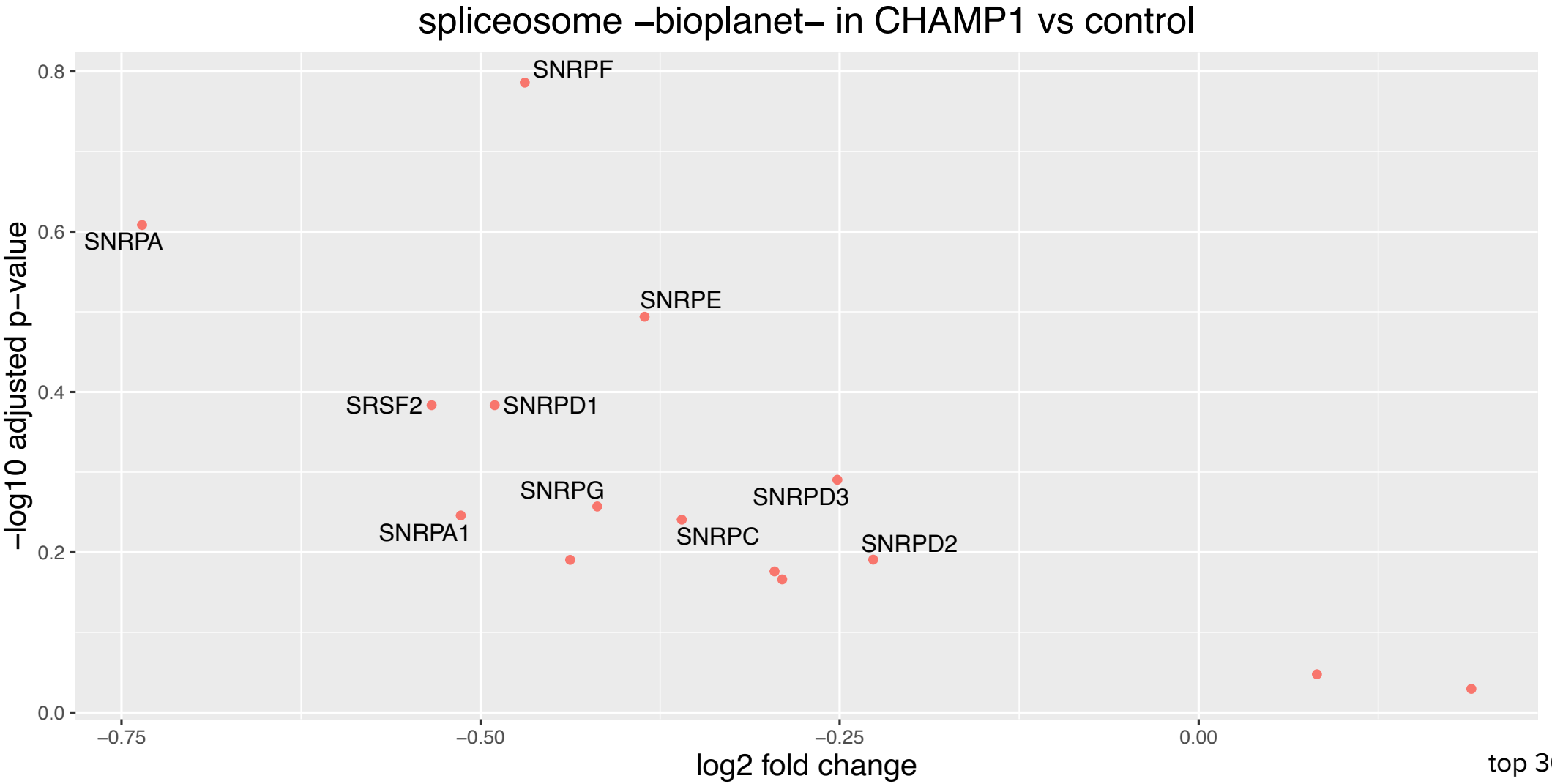


top 30 genes with the highest log2Foldchange are labeled



top 30 genes with the highest log2Foldchange are labeled

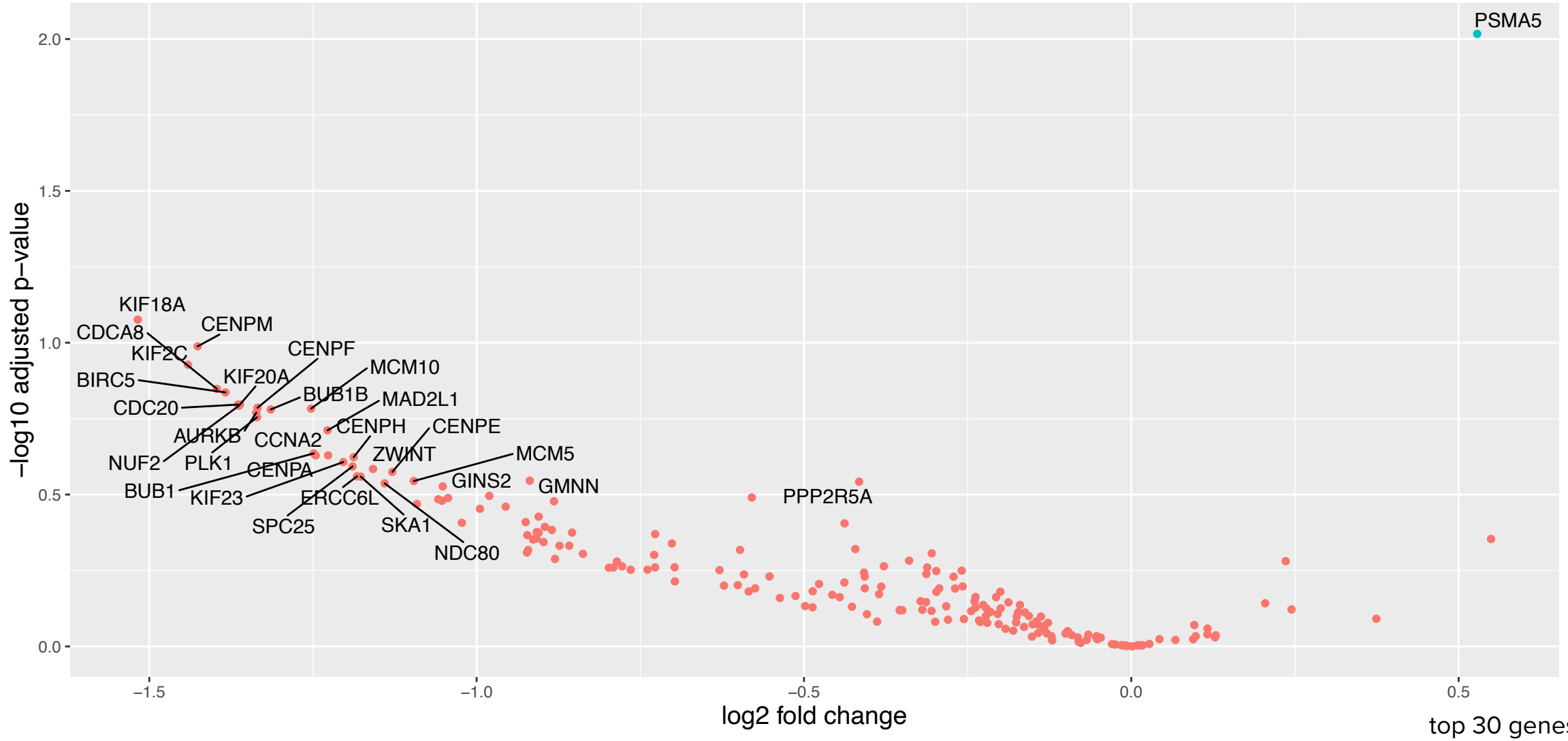
padj < 0.05 & |log2FoldChange| >= 0.58



top 30 genes with
the highest log2Foldchange
are labeled

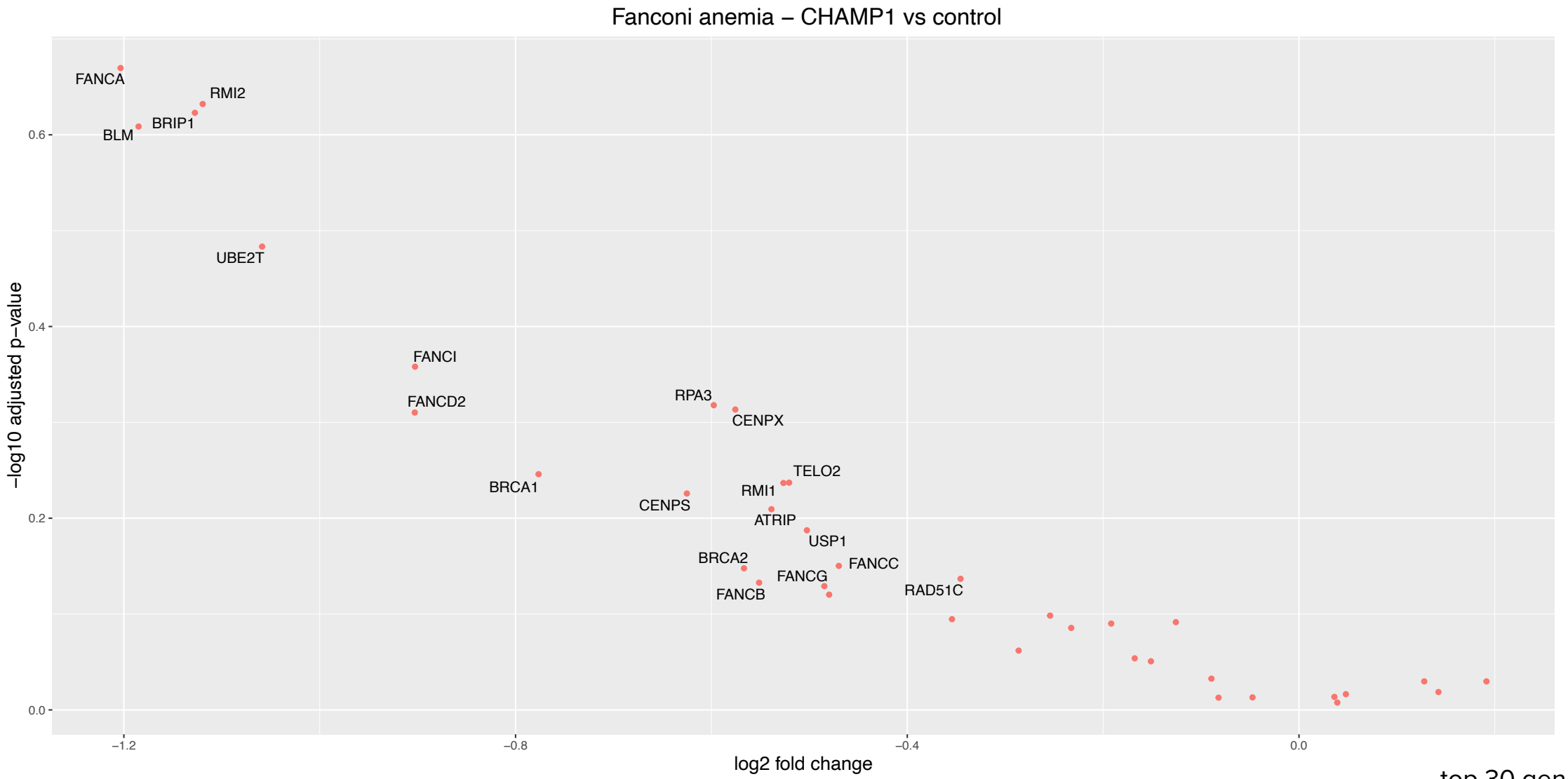
padj < 0.05 & |log2FoldChange| >= 0.58

DNA replication –bioplanet– CHAMP1 vs control



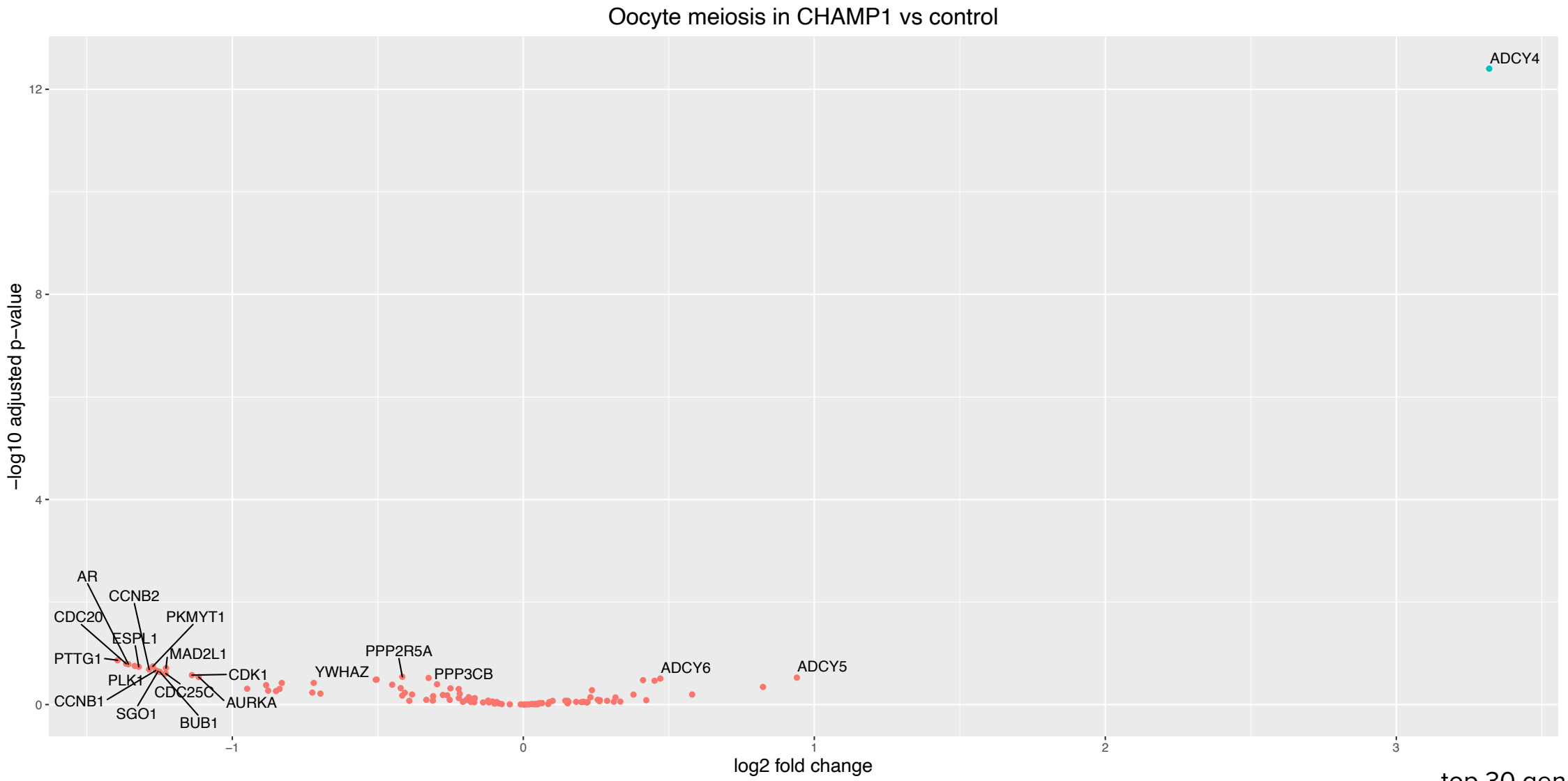
top 30 genes with the highest log2Foldchange are labeled

padj < 0.05 & |log2FoldChange| >= 0.58

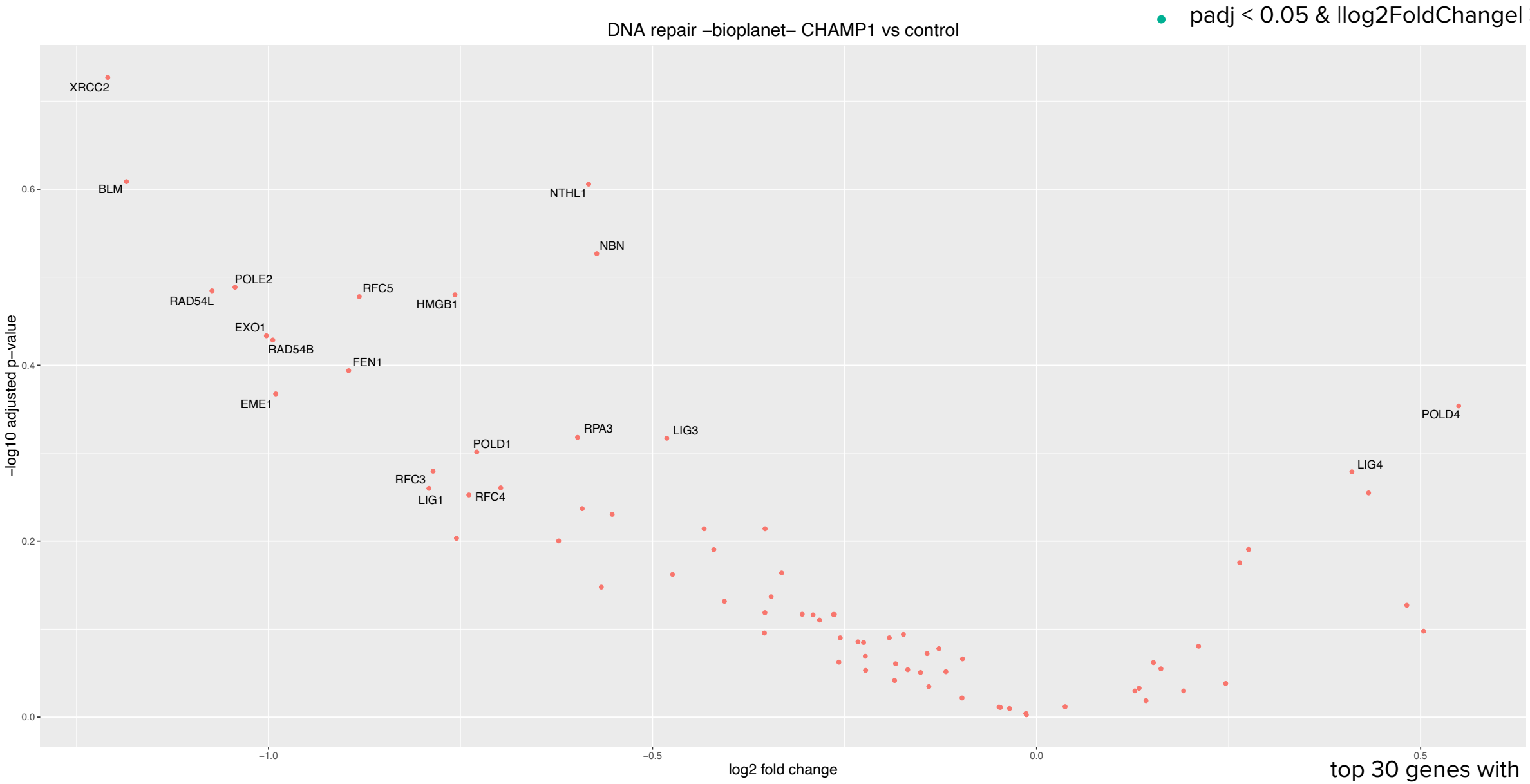


top 30 genes with
the highest log2Foldchange
are labeled

padj < 0.05 & |log2FoldChange| >= 0.58



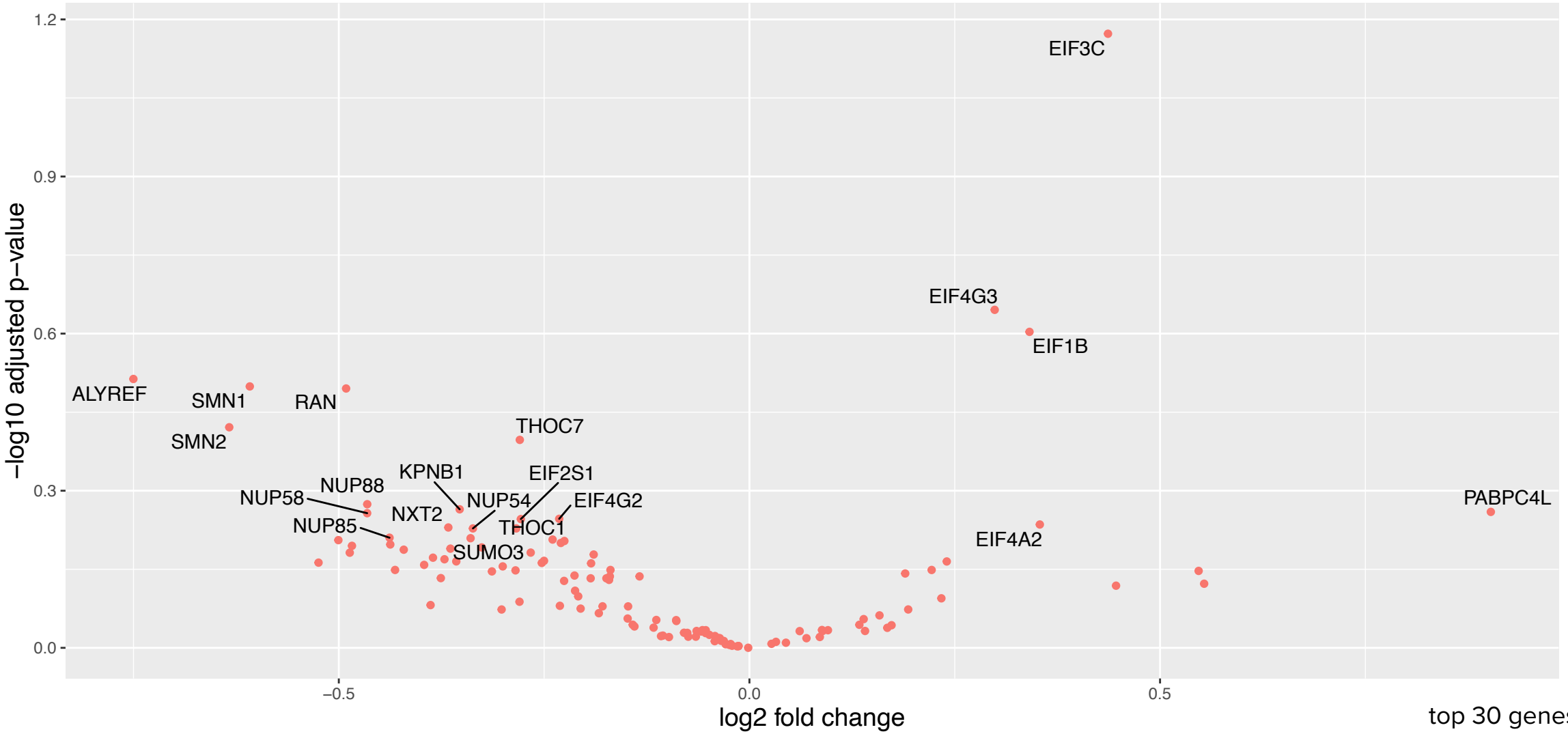
top 30 genes with
the highest log2Foldchange
are labeled



top 30 genes with the highest log2Foldchange are labeled

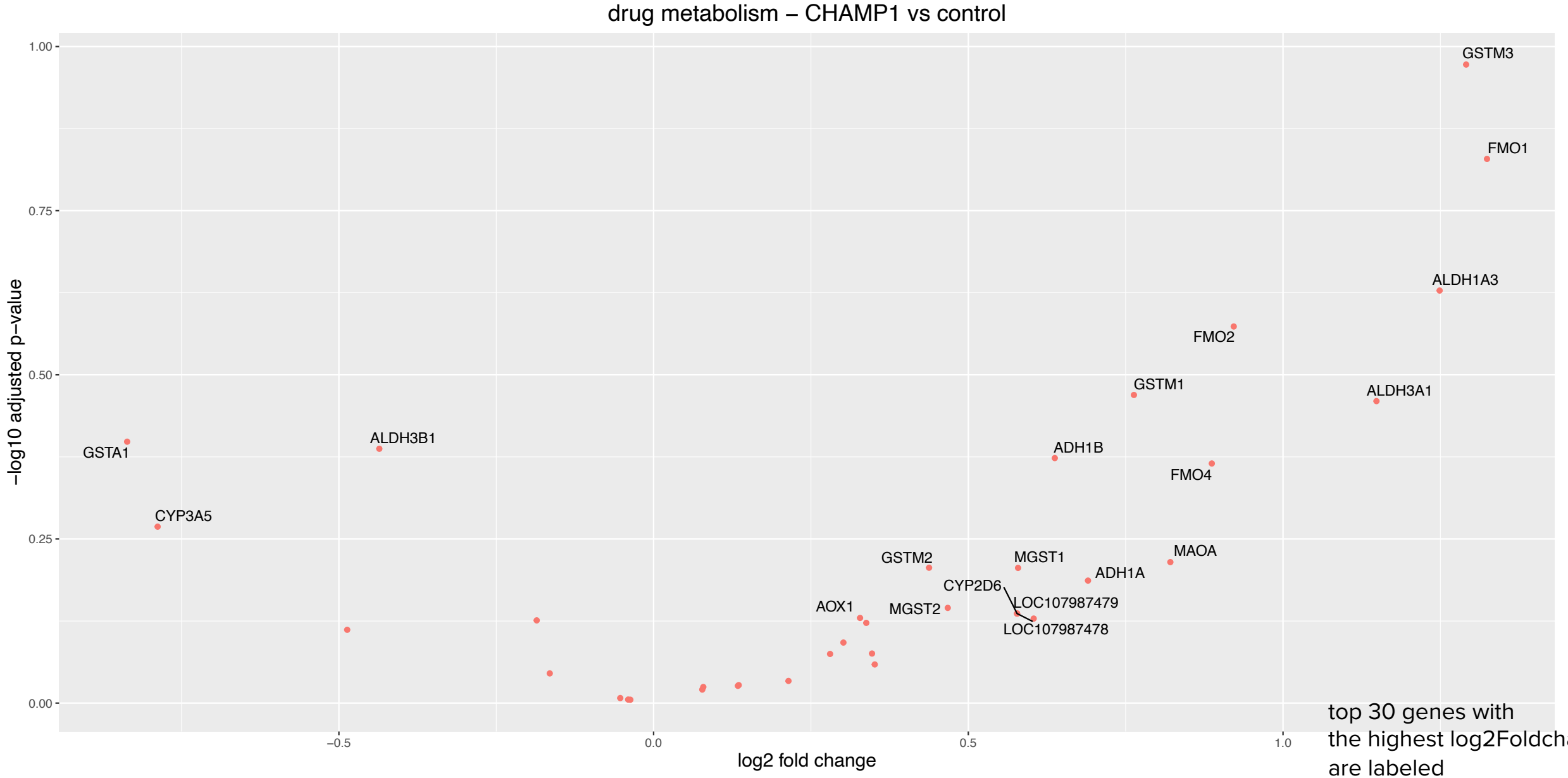
padj < 0.05 & |log2FoldChange| >= 0.58

RNA transport– CHAMP1 vs control



top 30 genes with the highest log2Foldchange are labeled

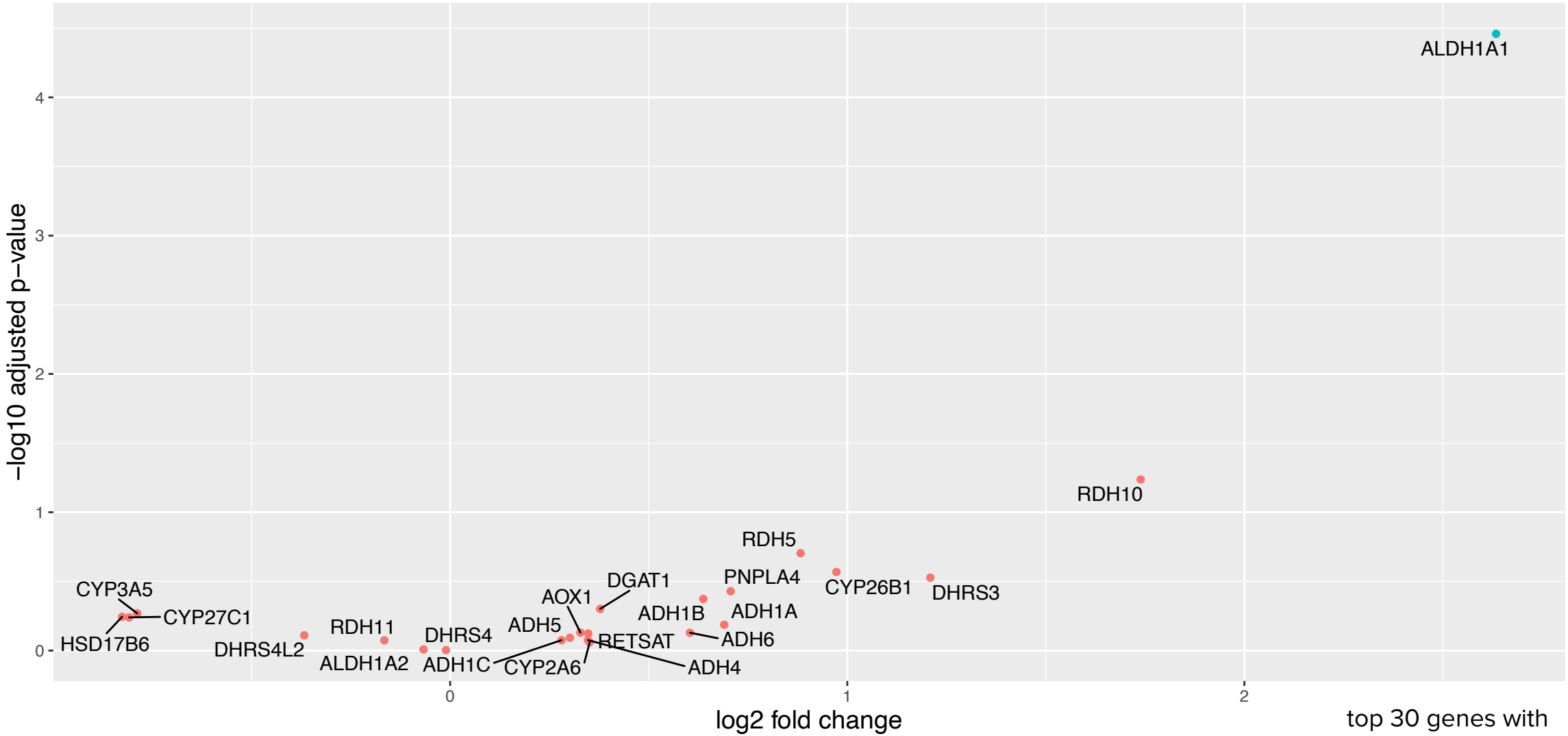
padj < 0.05 & |log2FoldChange| >= 0.58



top 30 genes with the highest log2Foldchange are labeled

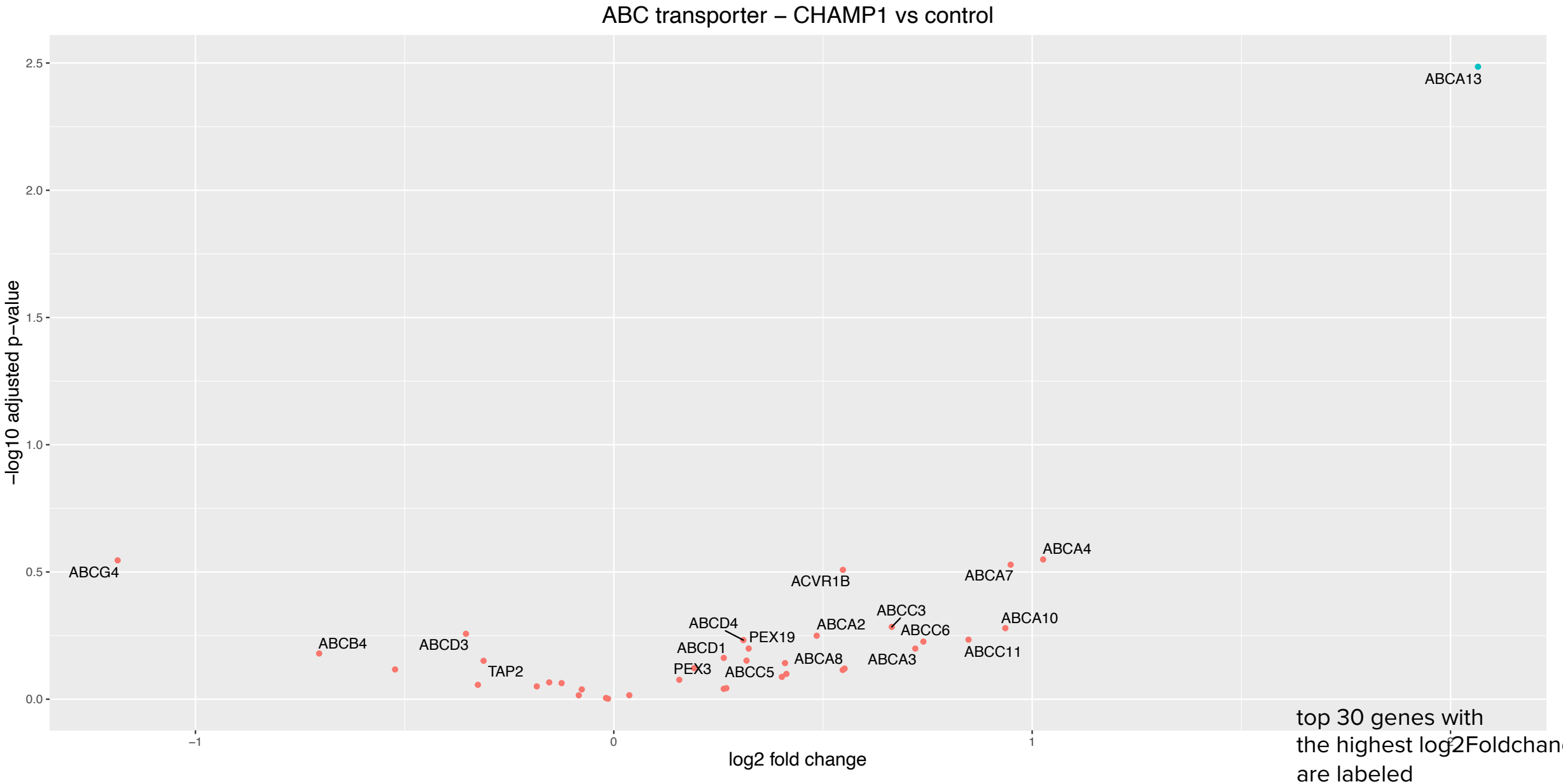
padj < 0.05 & |log2FoldChange| >= 0.58

retinol metabolism in CHAMP1 vs control



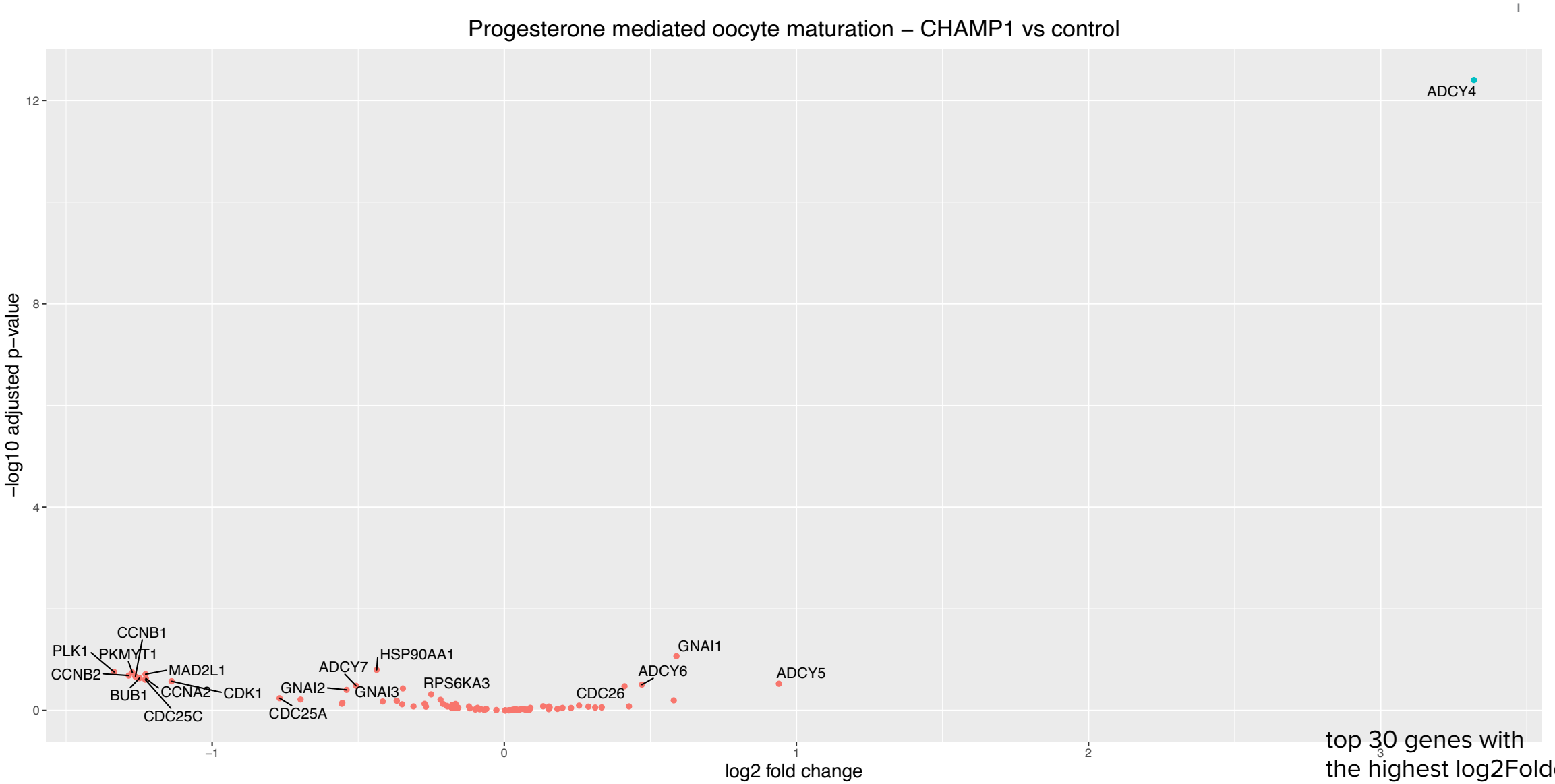
top 30 genes with the highest log2Foldchange are labeled

padj < 0.05 & |log2FoldChange| >= 0.58



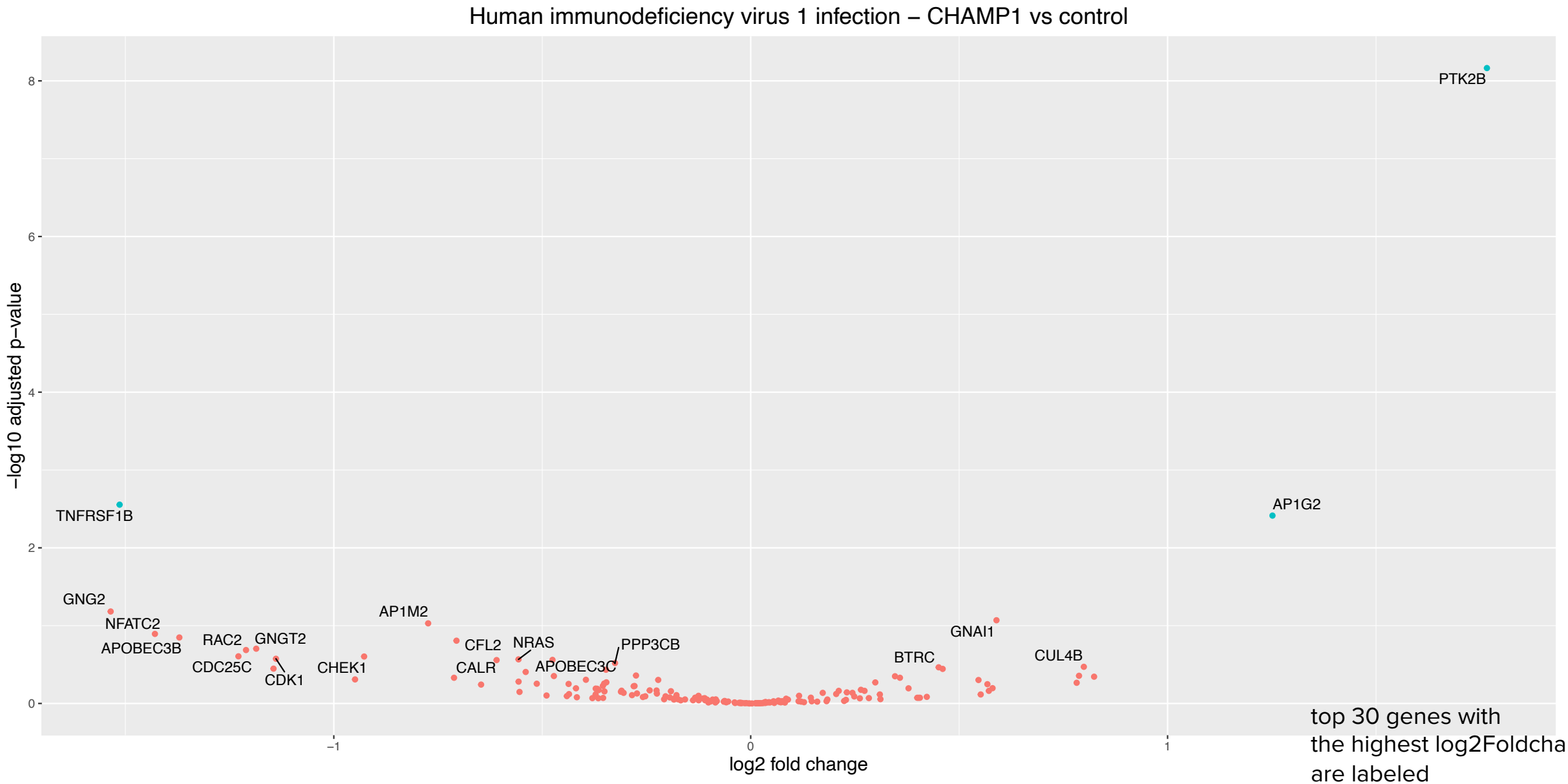
top 30 genes with the highest log2Foldchange are labeled

padj < 0.05 & |log2FoldChange| >= 0.58



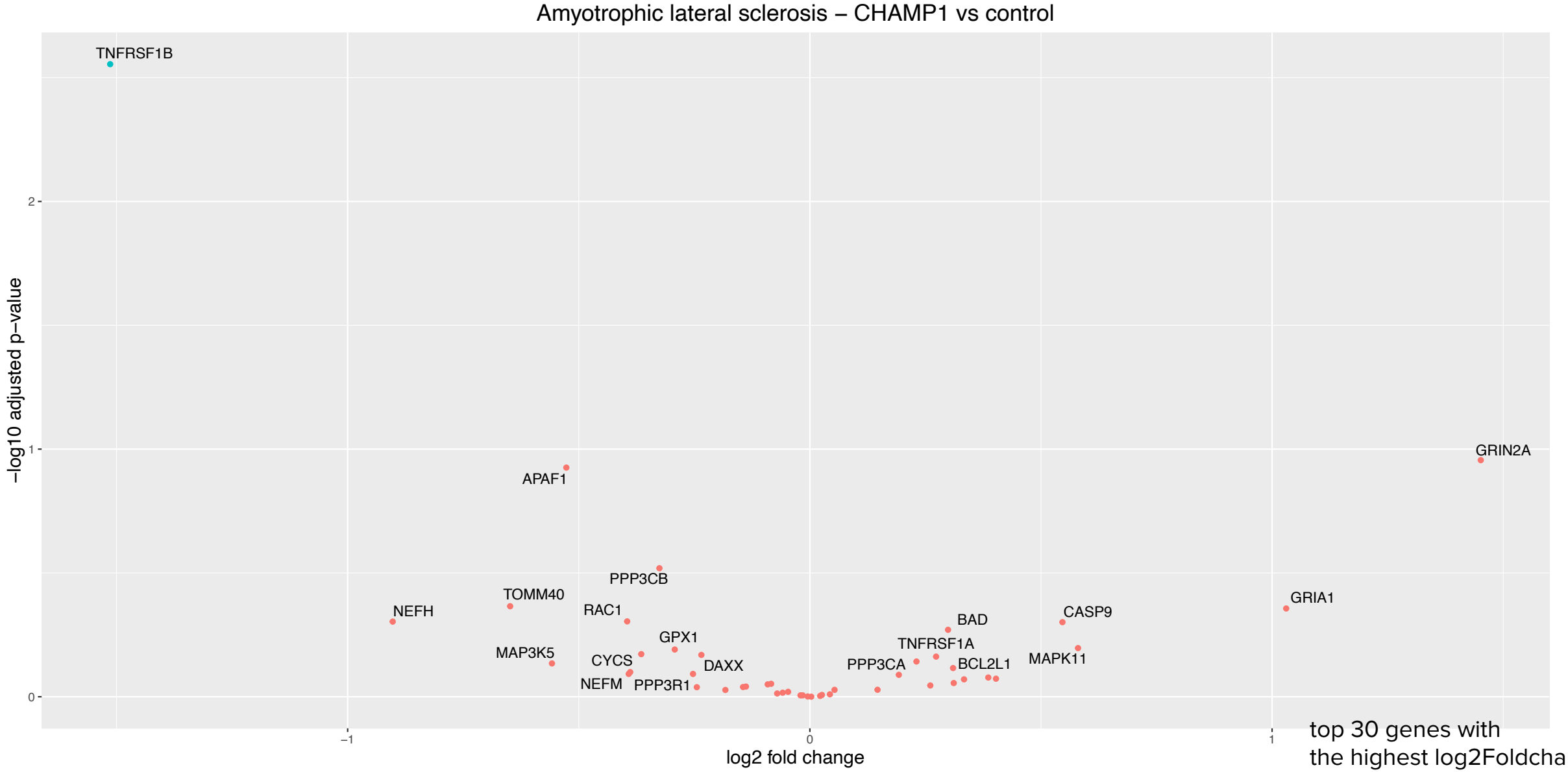
top 30 genes with the highest log2Foldchange are labeled

padj < 0.05 & |log2FoldChange| >= 0.58



top 30 genes with the highest log2Foldchange are labeled

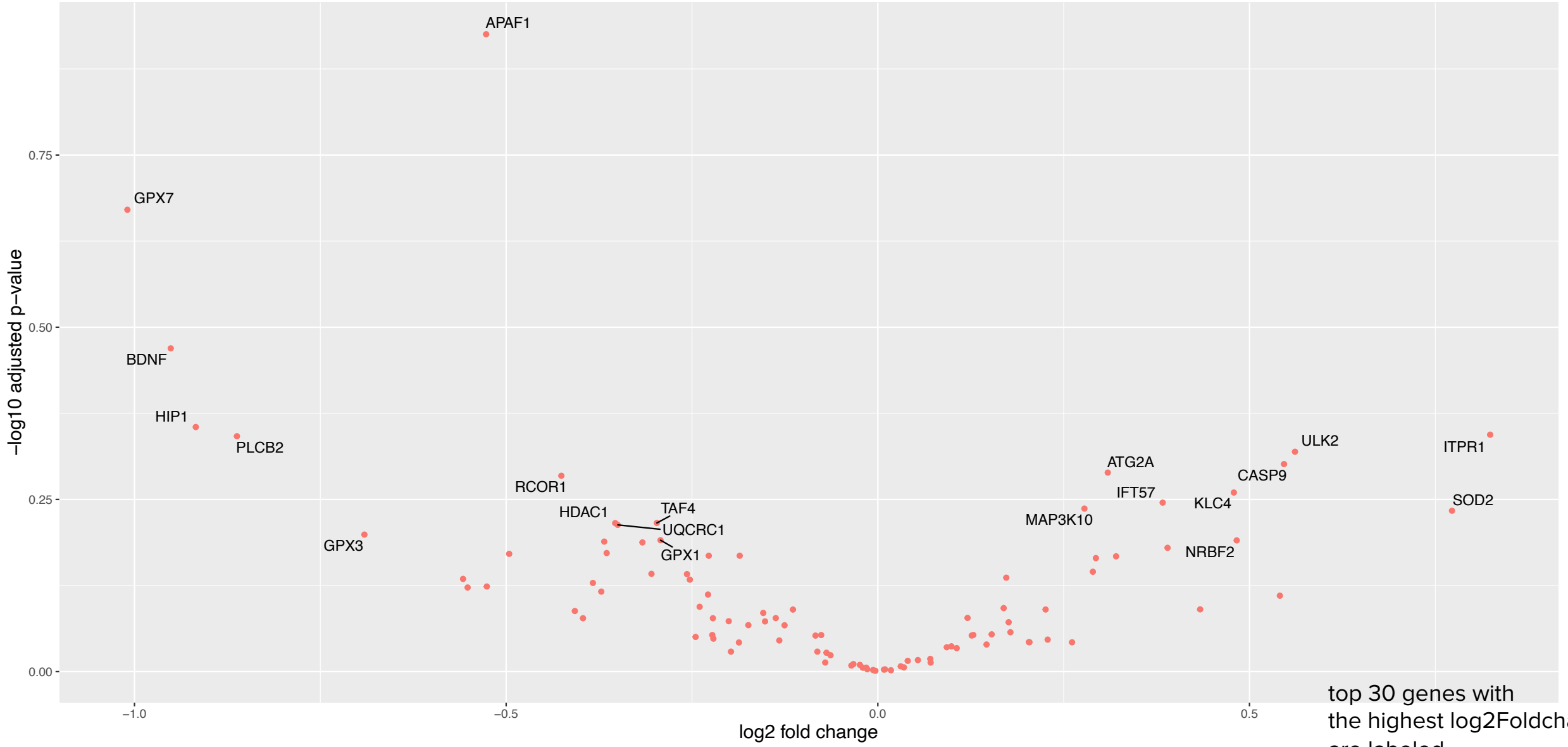
padj < 0.05 & |log2FoldChange| >= 0.58



top 30 genes with the highest log2Foldchange are labeled

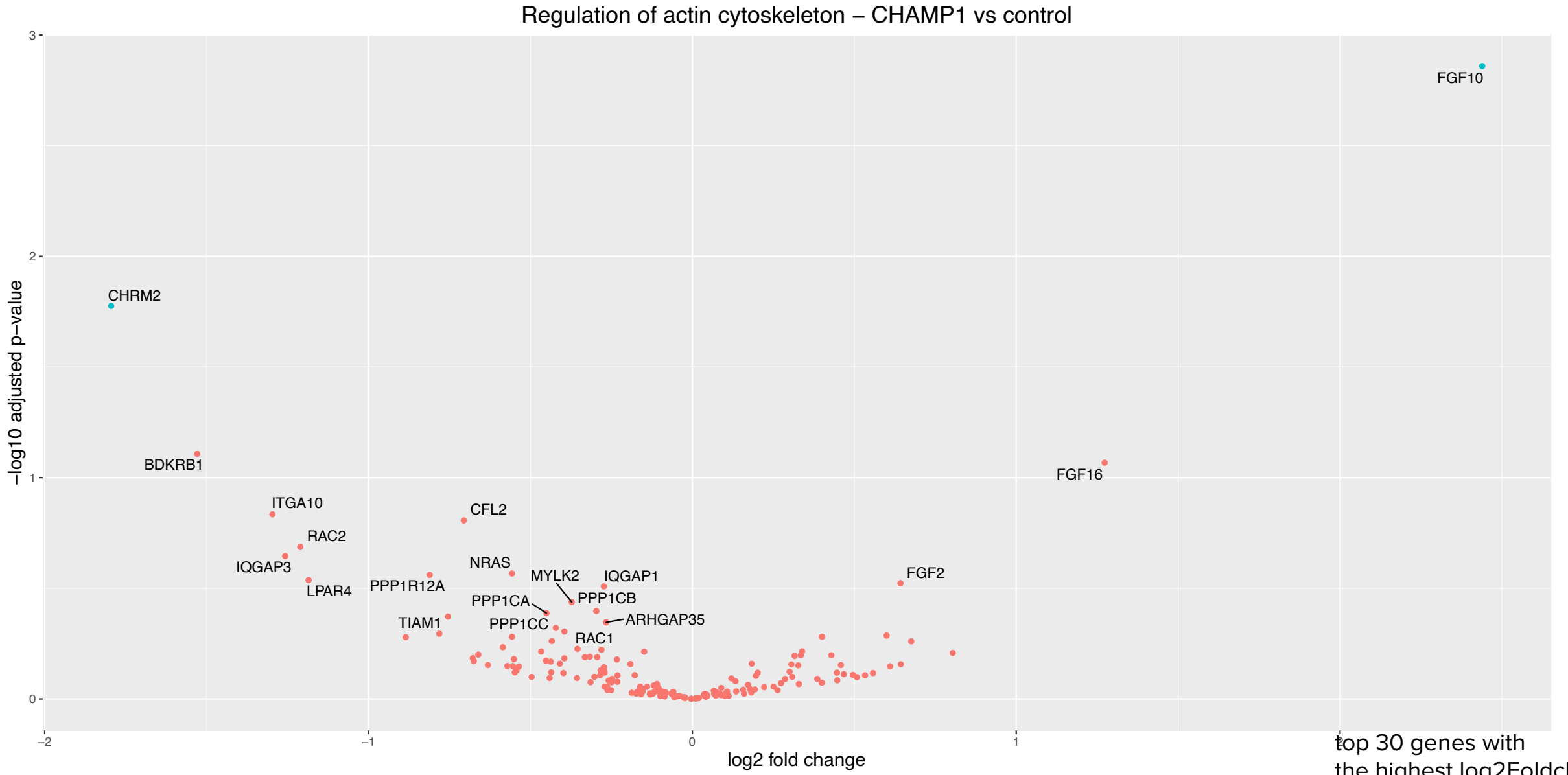
padj < 0.05 & |log2FoldChange| >= 0.58

Huntington disease – CHAMP1 vs control



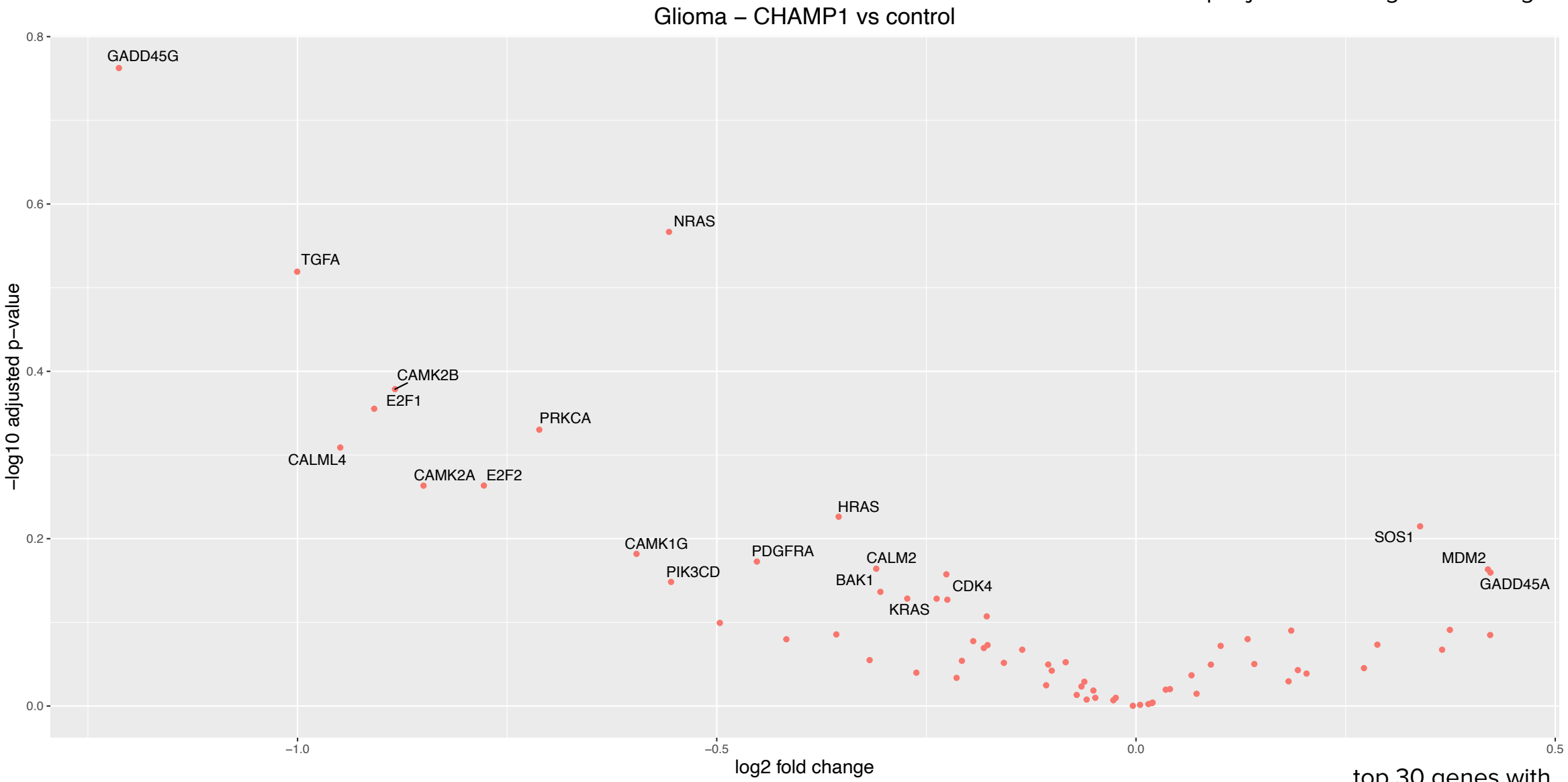
top 30 genes with the highest log2Foldchange are labeled

padj < 0.05 & |log2FoldChange| >= 0.58

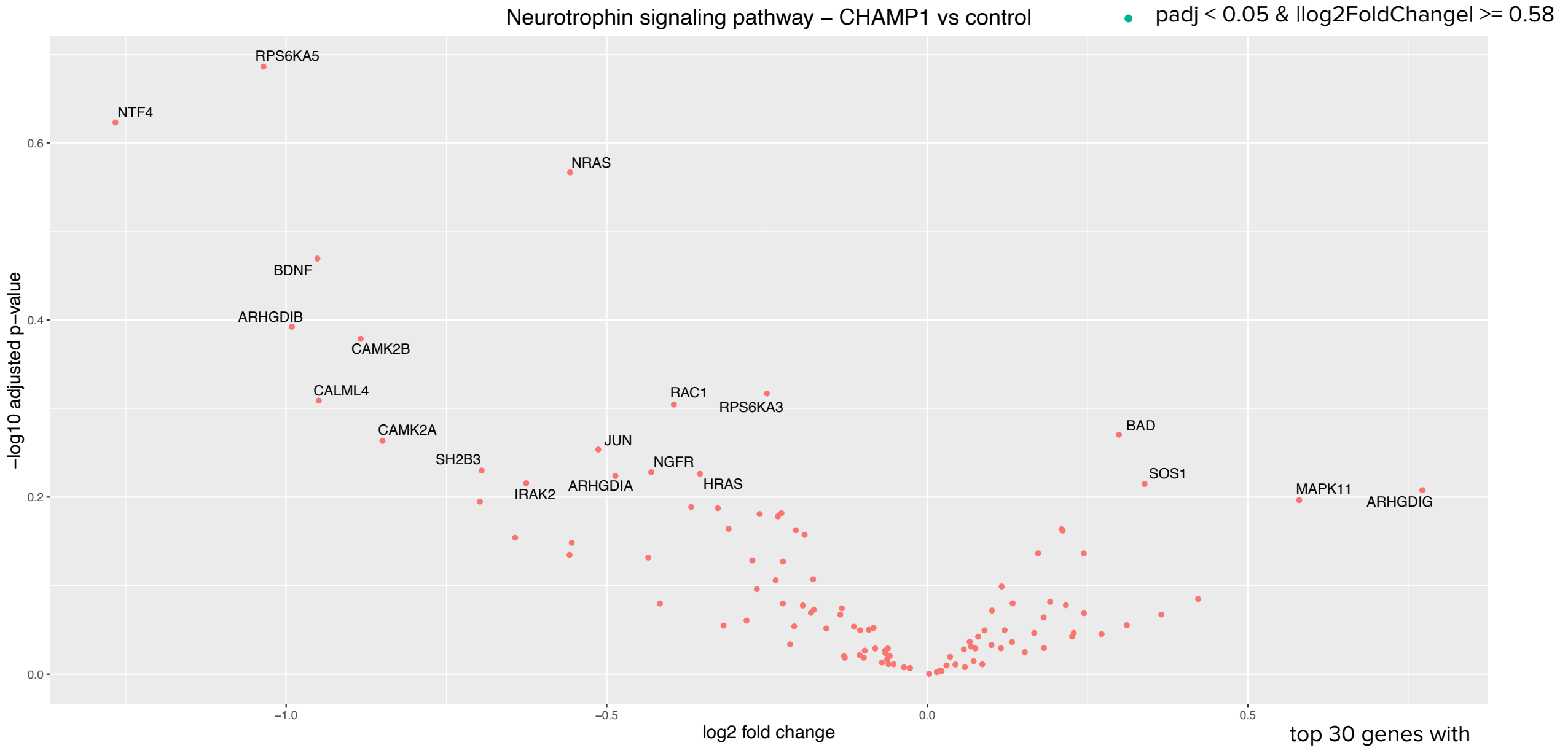


top 30 genes with the highest log2Foldchange are labeled

padj < 0.05 & |log2FoldChange| >= 0.58

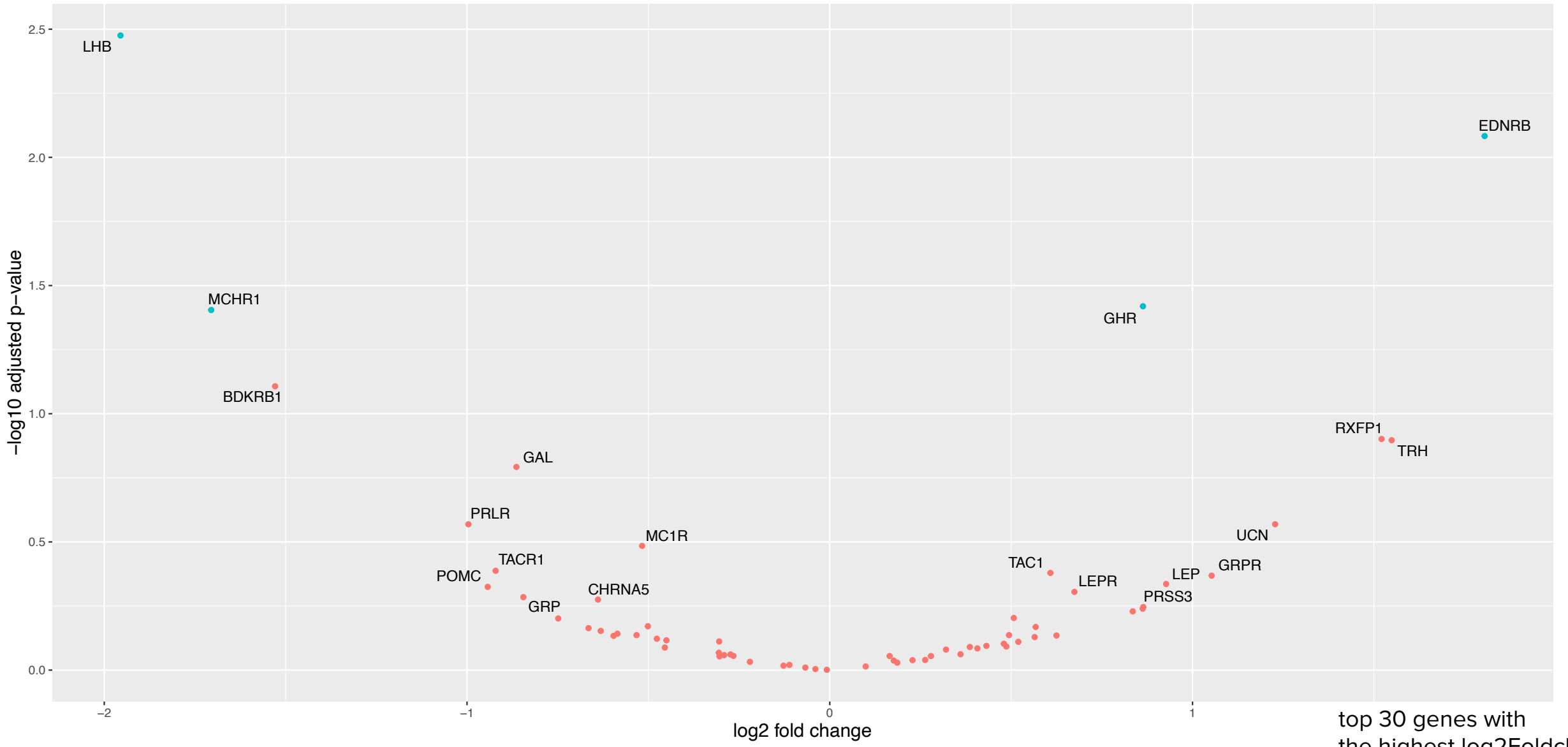


top 30 genes with the highest log2Foldchange are labeled

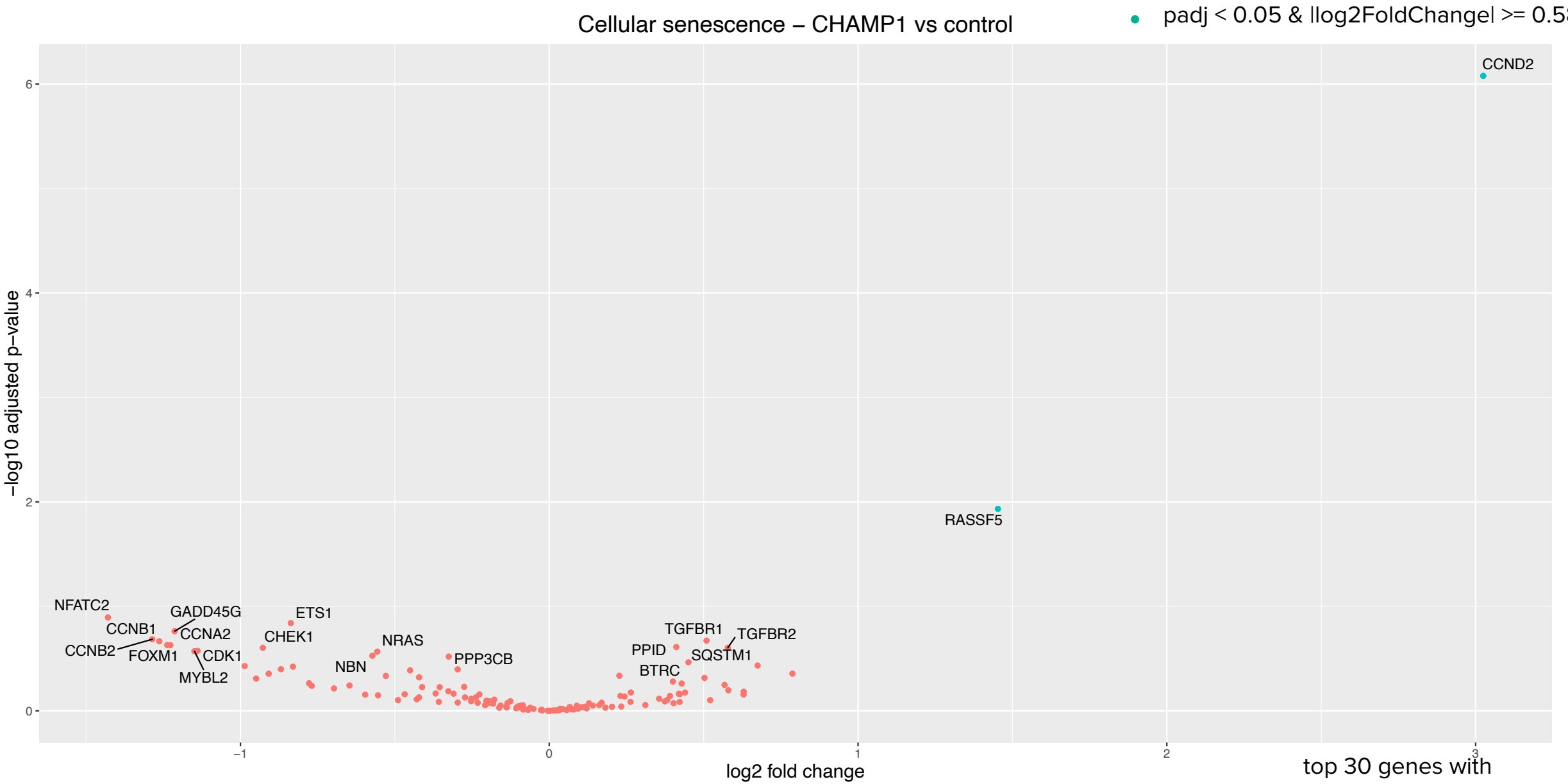


top 30 genes with the highest log2Foldchange are labeled

Neuroactive ligand–receptor interaction – CHAMP1 vs control

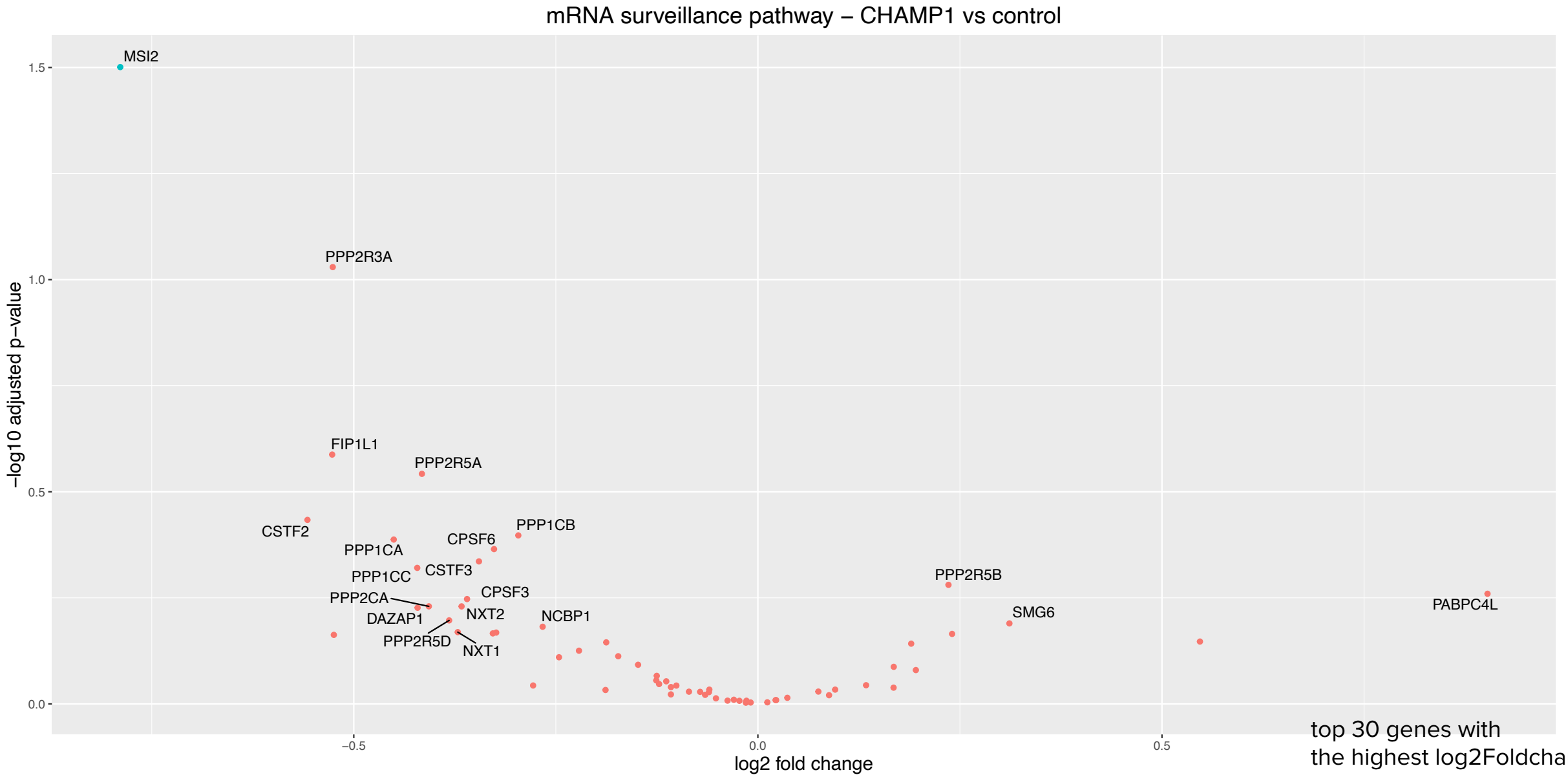


top 30 genes with
the highest log2Foldchange
are labeled



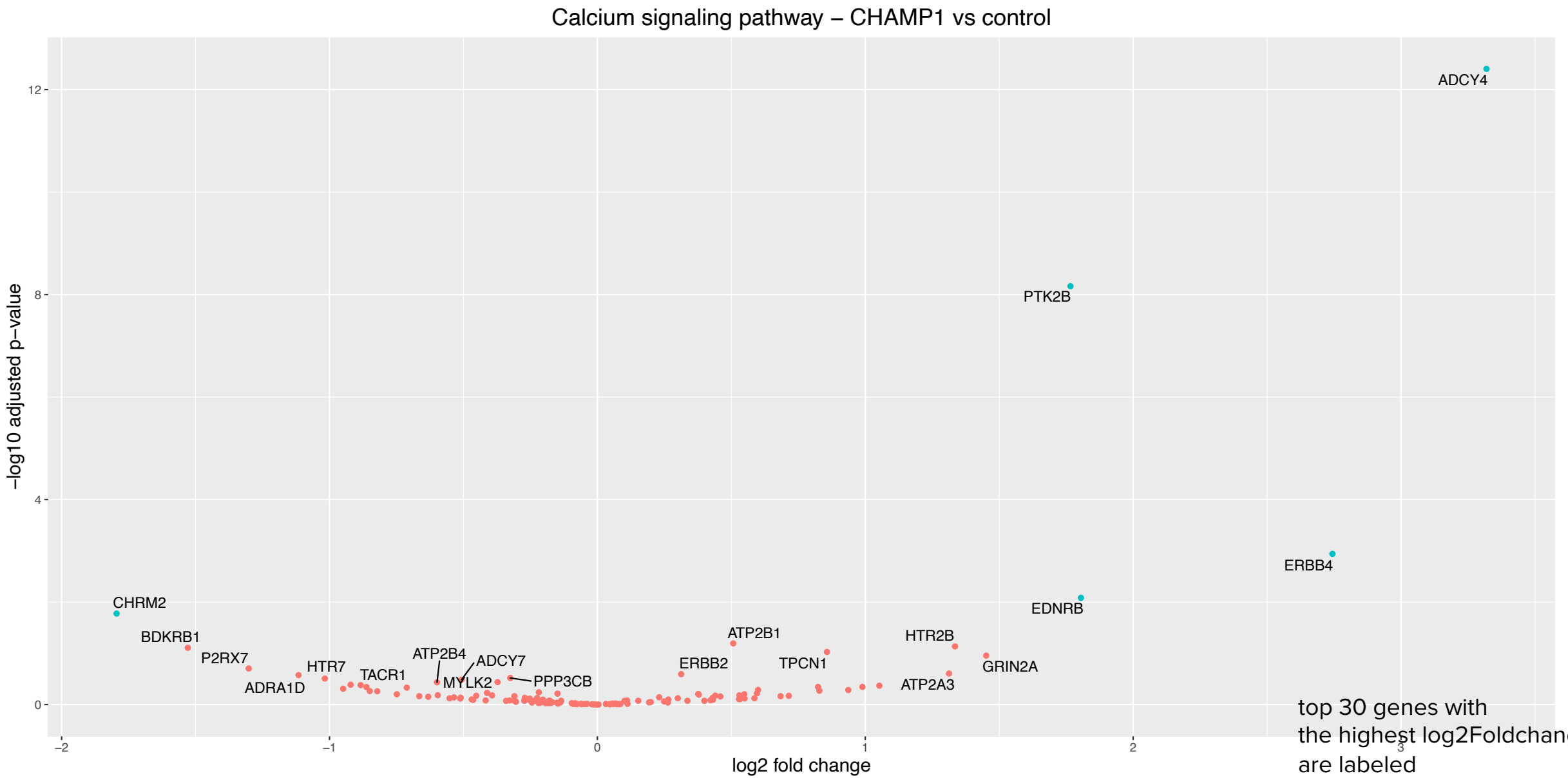
top 30 genes with the highest log2Foldchange are labeled

padj < 0.05 & |log2FoldChange| >= 0.58



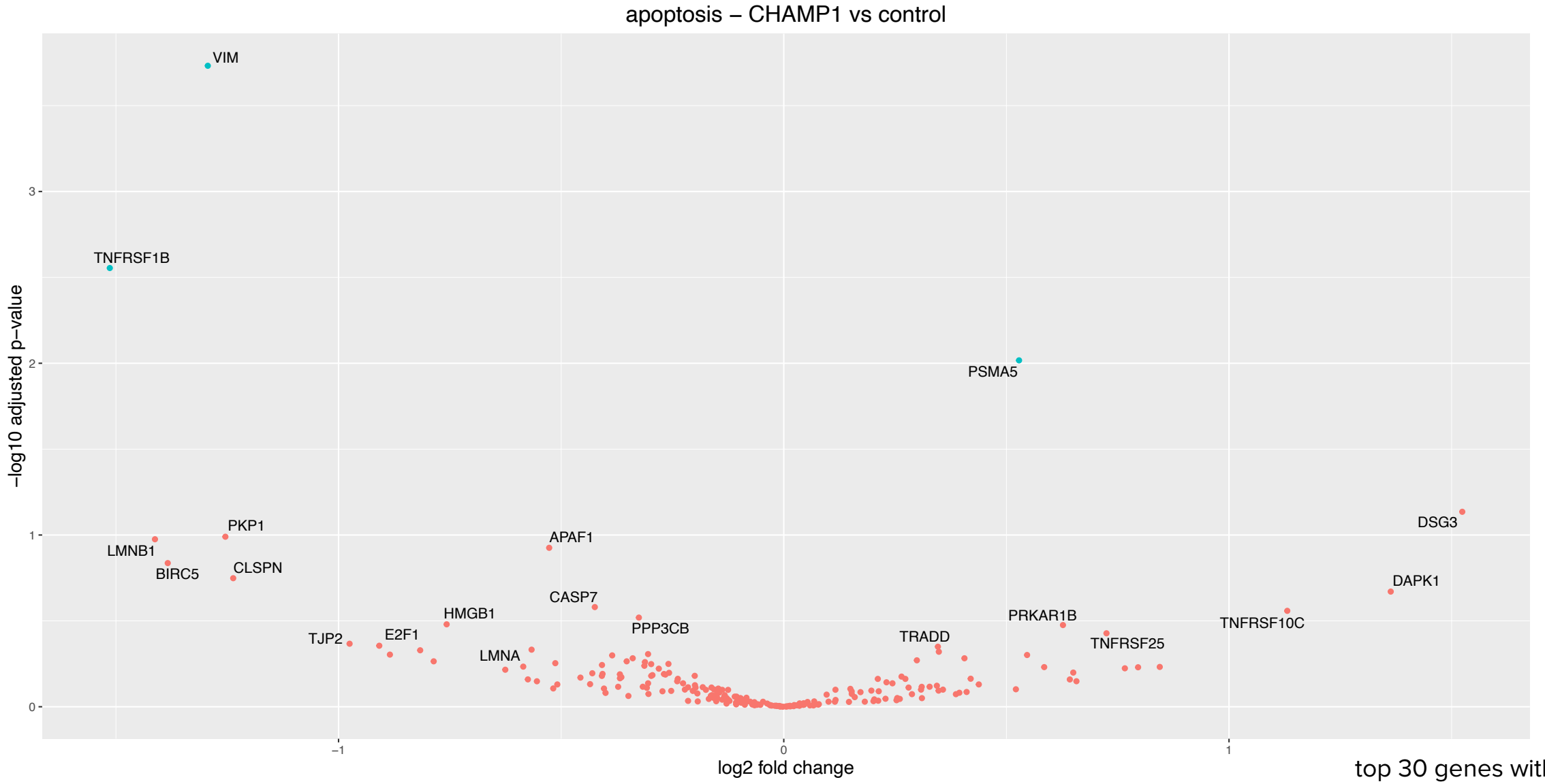
top 30 genes with the highest log2Foldchange are labeled

padj < 0.05 & |log2FoldChange| >= 0.58



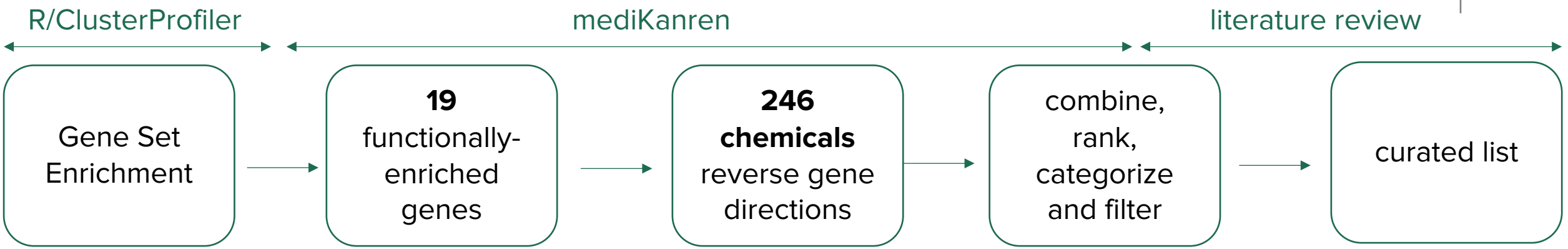
top 30 genes with the highest log2Foldchange are labeled

padj < 0.05 & |log2FoldChange| >= 0.58



top 30 genes with the highest log2Foldchange are labeled

Strategy I – find drugs purely based on gene expression information



Select genes in enriched functional categories:

- upregulated genes: CCND2 RASSF5 PSMA5 ADCY4 ALDH1A1 ABCA13 PTK2B AP1G2 FGF10 CD36 EDNRB ERBB4
- downregulated genes: TUBA1A TNFRSF1B CHRM2 KCNN4 LHB MSI2 VIM

Strategy I

- find drugs **purely** based on gene expression information
- reduce the number of genes to target by performing GSEA outside of mediKanren
- Genes nor drugs don't necessarily relate to CHAMP1 disease conditions

Filtering by categorizing chemicals

chemical substances CHEBI:

supplements/ food

- relatively safe
- although they are not tested for safety

drugs

- indicated for
- contraindicated for
- unintended effects (toxicity)

endogenous (non-drugs)

- eg, acetylcholine
- pyrimidines
- taurine
- cyclic AMP

(unsafe until proven otherwise)

- investigational compounds
- experimental drugs
- veterinary medicine
- pesticides
- toxicants
- other chemicals

What is a drug?

- drugs are substances intended to diagnose, treat or prevent diseases
- in the US, drugs are tested by the FDA by passing clinical trials

What is a supplement?

- Under the Dietary Supplement Health and Education Act (DSHEA), FDA treats supplements as food taken orally to supplement diets.

note : mediKanren can help categorize drugs vs non-drugs >90% accuracy but not other categories

Strategy I - prioritized drugs and supplements

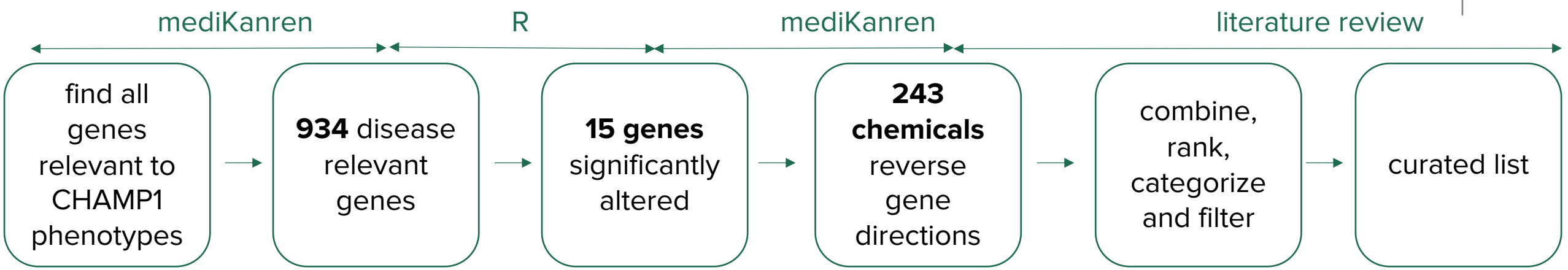
Prioritized drugs (filtered with reasonable toxicity)

- estradiol (6 enriched genes) : sex hormones
- progesterone (4): sex hormones
- testosterone (3): sex hormones
- losartan (3) : high blood pressure
- ~~sirolimus (3): immune suppressant~~
- indomethacin (3): anti-inflammatory/NSAIDs
- aspirin (2): anti-inflammatory, blood thinner
- flutamide (2): anti-androgen
- ~~Clozapine (2): mental disorders~~

Prioritized supplements

- calcium (3 enriched genes)
- quercetin (2)
- sodium butyrate (2)
- trans-resveratrol (2)
- rottlerin (2)
- curcumin (2)
- genistein (2) : phytoestrogen

Strategy II – find drugs for genes that are phenotypically relevant



CHAMP1 phenotypes: (hypotonia, cerebral palsy, autism, development delay and epilepsy)

Phenotype associated genes:

- upregulated genes: CACNB4 PRICKLE1 DMD SHANK2 MAP2 H19 BCHE EFNA5 CNTNAP2 ANKH SLC12A6
- downregulated genes: GABRA5 TNFRSF1B STX1B RBFA

Strategy II

- focus on phenotype-associated genes to reduce number of genes to target

Strategy II - gene targets (1)

CACNB4: Calcium Voltage-Gated Channel Auxiliary Subunit Beta 4

- encodes for the beta subunit of voltage-dependent calcium channel
- mediates influx of Ca2+ into the cell upon membrane polarization
- gene mutation is associated with epilepsy and ataxia

PRICKLE1: Prickle Planar Cell Polarity Protein 1

- encodes a nuclear receptor that is a negative regulator of the Wnt/beta-catenin signaling
- involved in nuclear trafficking of transcription repressor REST/NRSF and REST4
- implicated in dementia, Parkinson's, myoclonic seizure, abnormal peripheral nervous system morphology

DMD: Dystrophin

- component of the dystrophin-glycoprotein complex which bridges the inner cytoskeleton and the extracellular matr
- mutation causes Duchenne muscular dystrophy

SHANK2: SH3 and Multiple Ankyrin Repeat Domains 2

- a member of the Shank family of synaptic proteins
- molecular scaffolds in the postsynaptic density of excitatory synapses of the brain
- alterations in protein causes autism spectrum disorder

Strategy II - gene targets (2)

MAP2: Microtubule Associated Protein2

- microtubule assembly, essential step in neurogenesis
- implicate in determining and stabilizing dendritic shape during neuron development
- associated with Central Neurocytoma, Olivopontocerebellar Atrophy

H19: H19 imprinted Maternally Expressed Transcript

- encode for long non-coding RNA which functions as a tumor suppressor
- Beckwith-Wiedemann Syndrome, Wilms tumorigenesis, hemihyperplasia

BCHE: Butyryl cholinesterase

- enzyme of the type-B carboxylesterase/lipase family
- detoxification of poisons including organophosphate nerve agents and pesticides, metabolisms of drug (cocaine, heroin)
- implicated in abnormality of the nervous system/liver/cardiovascular
- involved in response to nutrient, learning, negative regulation of cell proliferation, neuroblast differentiation, choline metabolism

Strategy II - gene targets (3)

EFNA5 : Ephrin A5

- prevent axon bundling in cocultures of cortical neurons with astrocytes, a model of late-stage nervous system development and differentiation
- belong to the eph family tyrosine kinase, implicated in development of the nervous system
- crucial for migration, repulsion and adhesion during neuronal, vascular and epithelial development

CNTNAP2: Contactin Associated Protein 2

- member of neurexin family
- cell adhesion function in vertebrate nervous system
- contains epidermal growth factor repeats and laminin G domains
- localized at the juxtaparanodes of myelinated axons, mediates interaction between neurons and glia during nervous system development.
- involved in localization of K+ channel in differentiating axons
- directly bound to fork head box protein 2, a TF related to speech and language development
- implicated in multiple neurodevelopmental disorders, eg Gilles de la Tourette Syndrome, schizophrenia, epilepsy, autism, ADHD and ID.

Strategy II - gene targets (4)

ANKH: ANKH Inorganic Pyrophosphate Transport Regulator

- multipass transmembrane protein expressed in joints and other tissues and controls pyrophosphate levels in cultured
- inorganic phosphhate transmembrane transporter activity
- mutation associated with autosomal dominant craniometaphyseal dysplasia, Chrondrocalcinosis
- abnormality of head and neck, mouth, oral cavity, and dentition

SLC12A6: Solute Carrier Family 12 Member 6

- K-Cl cotransporter family
- lower intracellular chrolide level below the electrochemical equilibrium potential
- induced by cell swelling under hypotonic condition
- mutation affects agenesis of the corpus callosum with peripheral neuropathy, Charcot-Marie-Tooth Disease
- abnormality of head and neck, mouth, oral cavity, and dentition

Strategy II - gene targets (5)

GABRA5: Gamma-Aminobutyric Acid Type A Receptor Subunit Alpha 5

- receptor for GABA (a major inhibitory neurotransmitter)
- ligand-gated choride chanel
- related diseases: epileptic encephalopathy, Early Infantile, 79, Undetermined Early-Onset Epileptic Encephalophathy
- abnormality of body height, head and neck, mouth, oral cavity, and dentition

TNFRSF1B: TNF Receptor Superfamily Member 1B

- member of TNF super receptor family
- recruits 2 anti-apoptotic proteins, c-IAP1 and c-IAP2
- mice studies suggest that this protein protect neurons from apoptosis by stimulating antioxidative pathways
- abnormality of head and neck, face, orbital region, liver morphology

STX1B: Syntaxin 1B

- play role in exocytosis of synaptic vesicles. Vescicle exocytosis releases vesicular contents. In neurons, secretion of transmitt plays important role in synaptic transmission.
- mutation caues fever-associated epilepsy, myoclonic seizures, generalized non-convulsive status epilepticus without coma, infection-related seizures.
- also linked to Parkinson

RBFA: Ribosome Binding Factor A

- rRNA binding protein
- Disease associated with RBFA are Autosomal Dominant Non-Syndromic Intellectual Disability

Strategy II - prioritized drugs and supplements

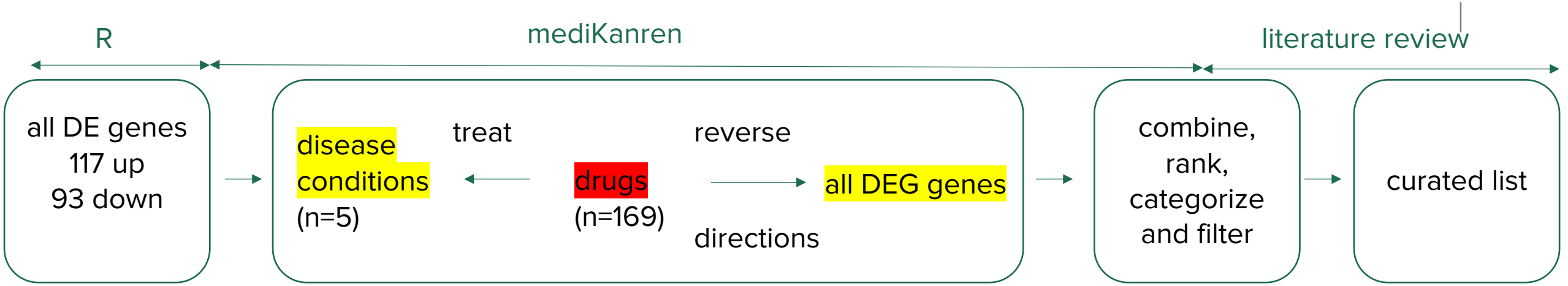
Prioritized drugs
(hit >=2 gene targets and little toxicity)

- valproic acid (anticonsulsant) (5) : GABRA5, CNTNAP2, EFNA5, BCHE, H19
- amitriptyline (tricyclic antidepressant) (3): TNFRSF1B, BCHE, H19
- simvastatin (cholesterol-lowering) (2): BCHE, DMD
- estradiol (sex hormones) (2): TNFRSF1B, DMD

Supplements (hit 1 phenotypically-relevant gene)

- Quercetin
- neferine
- luteolin
- acetyl carnitine
- citric acid monohydrate
- indole-3-carbinol
- phenylcaffeate
- kaempferol
- apigenin
- S-(+)-evodiamine
- myricetin
- epigallocatechingallate
- piperine
- calcium
- rutin
- curcumin
- genistein
- alpha Tocopherol
- Zinc Sulfate
- Choline

Strategy III – mediKanren 2-hop query



CHAMP1 phenotypes: (hypotonia, cerebral palsy, autism, development delay and epilepsy)

Strategy III

- leverage mediKanren's capacity to traverse multiple knowledge graphs to go from phenotypes -> genes -> drugs

Strategy III - prioritized drugs and supplements

Prioritized drugs

(hit >=2 gene targets and little toxicity)

- valproic acid (anti-convulsant) (17)
- estradiol (sex hormones) (7)
- progesterone (sex hormones) (7)
- metformin (glucose-lowering) (4)
- indometacin (NSAID) (4)
- dihydrotestosterone (sex hormones) (3)
- losartan (anti-hypertensive) (3)
- aspirin (NSAID) (3)
- simvastatin (cholesterol-lowering) (3)
- furosemide (diuretic) (2)
- amitriptyline (tricyclic antidepressant) (2)
- phenytoin (anti-convulsant) (2)
- testosterone (sex hormones) (2)
- all-trans-retinoic acid (vit A metab) (2)
- Statin (cholesterol-lowering) ((2)
- nifedipine (calcium channel blocker) (2)
- adenosine (muscle relaxation) (2)
- perindopril (ACEi, decreasing blood pressure) (2)
- Lovastatin (cholesterol-lowering) (2)
- bumetanide (diuretic)(2)

Supplements

- **trans-resveratrol (5)**
- **berberine (3)**
- **quercetin 3)**
- **curcumin (3)**
- **myricetin (2)**
- **apigenin (2)**
- **rutin (2)**
- **alpha Tocopherol**

(vitamin E) (2)

- magnesium (1)
- manganese (1)
- eugenol (1)
- epigallocatechingallate (1)
- luteolin (1)
- indole (1)
- piperine (1)
- carvacrol (1)
- ursolic acid (1)
- sodium butyrate (1)

Conclusion

- A tentative pipeline to analyze patient's RNAseq data for drug repurposing with mediKanren
- Many caveats but promising results
- Potential improvements:
  - Cell/RNA source level:
    - more relevant cell type/tissue for diseases (neuronal cells/muscle cells, microbrain organoids, biopsies etc)
    - better experiment design with relevant functional tests
  - DEG level: use other DEG tools (edgeR or limma) to gain more confidence in the DEG list
  - MediKanren: find a robust way to rank and filter drugs
  - build a reproducible pipeline with Snakemake

- Comments and Next Steps?