Identifying Candidates for Master Regulators of Transcriptome Changes in Murine Hippocampi Following Subarachnoid Hemorrhage

Friederike Dündar & Paul Zumbo

Applied Bioinformatics Core, Weill Cornell Medicine

RNA-seq

Sequenced DNA reads were aligned with default parameters to the mouse reference genome (mm9) using `STAR` (Dobin et al. 2013). Gene expression estimates were obtained with `featureCounts` using composite gene models (union of the exons of all transcript isoforms per gene) from UCSC mm9 annotation from Illumina's iGenomes (Liao, Smyth, and Shi 2014, @gencode).

Differentially expressed genes

Differentially expressed genes (DEG) were determined with `DESeq2` (Anders and Huber 2010) using the standard workflow starting from integer read counts. We treated naive and sham samples as a one "control" condition, and contrasted the SAH samples against these. There were 1,040 differentially expressed genes detected (FDR < 0.10): 642 were up-regulated in the SAH samples; and 398 were up-regulated in the control samples. Only genes with an FDR < 0.10 were used for all motif analyses decsribed below.

Identifying putative regulators based on the genomic location of the DEG

RegulatorTrail

We used the web-based software `RegulatorTrail` to perform over-representation analysis (Kehl et al. 2017). Differentially expressed genes were split by the direction of their fold-change ("up" and "down") and uploaded to RegulatorTrail's `Over-representation analysis` tool, which is designed to find transcriptional regulators whose set of target genes have a significant overlap with differentially expressed genes relative to their in-house collection of regulator-target interactions, which are based on third-party resources that contain information on TF binding motifs or interactions between regulators and associated target genes. These resources include ChEA, ChIP-Atlas, ChipBase, ENCODE, JASPAR, SingaLink, and TRANSFAC, and they encompass information about transcription factor binding site motifs and bindings sites determined via ChIP-seq. Typically, a regulator protein is assigned to a gene if the binding site is within an interval around the TSS.

For our analysis, default parameters were used.

Ingenuity Pathway Analysis Studio

Differentially expressed genes were analyzed using `Ingenuity Pathway Analysis` (IPA) (QIAGEN Inc., n.d.) via their "Core Analysis" Feature. IPA has 4 causal analysis features (descriptions from (Krämer et al. 2014)):

- causal network
  - Most of IPA's causal analyses rely on a 'master' network which is derived from the Ingenuity Knowledge Base, where nodes are genes and edges reflect cause-effect relationships mined from the literature. The differential gene expression data is mapped onto that network and subsequent p-values and activation z-scores are computed.
- upstream regulator analysis (URA)
  - Identify molecules upstream of the genes in the dataset that potentially explain the observed expression changes.
- mechanistic networks
  - The goal is to determine those network edges between pre-determined upstream regulators for which there is statistical evidence that the corresponding relationship is likely relevant for the causal mechanism behind the dataset. The most significant causal edges between regulators are then used to construct networks downstream of a 'master' regulator in order to indicate possible causal (e.g. signaling) mechanisms.
- downstream effects analysis
  - This aims to identify those biological processes and functions that are likely to be causally affected by up- and down-regulated genes additionally predicting whether those processes are increased or decreased. This is based on the calculation of an activation z-score for which the information from the DEG is mapped onto the pathways underlying the Ingenuity Knowledge Base.

Using IPA's "Upstream Analysis" tab within the results, we then mined for putative upstream regulators by filtering for genes which had an expression log ratio greater than or less than 0 and which are known transcription regulator (by selecting that option in the "Molecule Type" filter).

Testing for over-represented motifs in the promoters of DEG

HOMER

Motif analysis was run separately for differentially expressed genes with positive or negative fold-changes. We used `HOMER` (v4.9.1) (Benner, Heinz, and Glass 2017) with the flags `-start -1000 -end 50`, which instructed `HOMER` to only look for motifs enriched within the promoters of DEG relative to all other promoters.
A promoter was defined as -1000bp to +50bp relative to each gene's transcription start site (TSS).

Pscan

We also subjected genes with a positive fold-change to transcription factor-binding motif analysis using the web-based software `Pscan`. Pscan is a software tool that scans a set of sequences (e.g. promoters) from co-regulated or co-expressed genes with motifs describing the binding specificity of known transcription factors and assesses which motifs are significantly over- or under-represented (Zambelli, Pesole, and Pavesi 2009).

The "Jaspar 2018_NR" (Khan et al. 2018) database of transcription factor binding sequences was analyzed using enriched groups of 950 base pair (bp) sequences to +50 bp of the 5' upstream promoters.

MEME suite

The MEME suite offers several tools for *de novo* motif finding as well as over-representation and enrichment analyses.

We extracted the DNA sequences representing 500 bp upstream of each TSS of both up- and down-regulated DEG.

DREME We first ran `DREME`, which discovers short, ungapped motifs that are relatively enriched in sequences of interest compared to control sequences (Bailey 2011). In contrast to MEME, DREME is able to handle relatively large sets of sequences (>50), which is necessary here because of the number of DEG that we were testing.

We ran DREME mostly with default settings, using 500 randomly sampled promoter regions of genes that showed no significant change as control regions:

```
dreme -verbosity 1 -oc . -dna -p genesFDR01_up_500bp_upstreamTSS_100bpChunks.bed.fa
-n genes_noChange_500bp_upstreamTSS_Rand500_100bpChunks.bed.fa
-t 18000 -e 0.05 -dfile description
```

CentriMo We then ran [`CentriMo`](http://meme-suite.org/tools/centrimo), which identifies known or user-provided motifs that show a significant preference for particular locations in your sequences. CentriMo takes a set of motifs and a set of equal-length sequences and plots the positional distribution of the best match of each motif (Lesluyes et al. 2014). We used 500 randomly sampled promoter regions of genes that showed no significant change as control regions.

```
centrimo --oc . --verbosity 1 --dfile description --local --score 5.0
--ethresh 10.0 --bfile genesFDR01_up_500bp_upstreamTSS.bed.fa.bg
--neg genes_noChange_500bp_upstreamTSS_Rand500.bed.fa
genesFDR01_up_500bp_upstreamTSS.bed.fa dreme_upRegulated.txt
```

CentriMo is extremely sensitive, therefore, when the provided sequence set is large (100s of sites), even motifs that are only slightly similar to the actual motif can show significant enrichment. The shape of the graph and the best bin width are subjective, so we prefer to use the E-value as a discriminator between different degrees of central (local) enrichment.

AME AME identifies user-provided motifs that are either relatively enriched in the sequences of interest compared with control sequences (McLeay and Bailey 2010),(Machanick and Bailey 2011). In contrast to CentriMo, the location of the motif within the supplied set of promoter sequences is not taken into account.

We ran AME via the web server using the following settings to test for enrichment of known mouse motifs:

```
ame --verbose 1 --oc . --scoring avg --method fisher --hit-lo-fraction 0.25
--evalue-report-threshold 10.0 --control genes_noChange_500bp_upstreamTSS_Rand500.bed.fa
genesFDR01_up_500bp_upstreamTSS.bed.fa db/EUKARYOTE/jolma2013.meme
db/JASPAR/JASPAR2018_CORE_vertebrates_non-redundant.meme db/MOUSE/uniprobe_mouse.meme
```

References

Anders, Simon, and Wolfgang Huber. 2010. "DESeq: Differential expression analysis for sequence count data." *Genome Biology* 11: R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Bailey, Timothy L. 2011. "DREME: Motif discovery in transcription factor ChIP-seq data." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btr261.

Benner, Christopher, Sven Heinz, and Christopher K Glass. 2017. "HOMER - Software for motif discovery and next generation sequencing analysis." http://homer.Ucsd.Edu/.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast universal RNA-seq aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Harrow, Jennifer, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, et al. 2012. "GENCODE: The reference human genome annotation for the ENCODE Project." *Genome Research* 22 (9): 1760–74. https://doi.org/10.1101/gr.135350.111.

Kehl, T., L. Schneider, F. Schmidt, D. Stockel, N. Gerstner, C. Backes, E. Meese, A. Keller, M. H. Schulz, and H. P. Lenhof. 2017. "RegulatorTrail: a web service for the identification of key transcriptional regulators." *Nucleic Acids Res.* 45 (W1): W146–W153.

Khan, A., O. Fornes, A. Stigliani, M. Gheorghe, J. A. Castro-Mondragon, R. van der Lee, A. Bessy, et al. 2018. "JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework." *Nucleic Acids Res.* 46 (D1): D1284.

Krämer, Andreas, Jeff Green, Jack Pollard, and Stuart Tugendreich. 2014. "Causal analysis approaches in ingenuity pathway analysis." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btt703.

Lesluyes, Tom, James Johnson, Philip Machanick, and Timothy L. Bailey. 2014. "Differential motif enrichment analysis of paired ChIP-seq experiments." *BMC Genomics*. https://doi.org/10.1186/1471-2164-15-752.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. "FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features." *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

Machanick, Philip, and Timothy L. Bailey. 2011. "MEME-ChIP: Motif analysis of large DNA datasets." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btr189.

McLeay, Robert C., and Timothy L. Bailey. 2010. "Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data." *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-11-165.

QIAGEN Inc. n.d. https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis.

Zambelli, F., G. Pesole, and G. Pavesi. 2009. "Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes." *Nucleic Acids Res.* 37 (Web Server issue): W247–252.