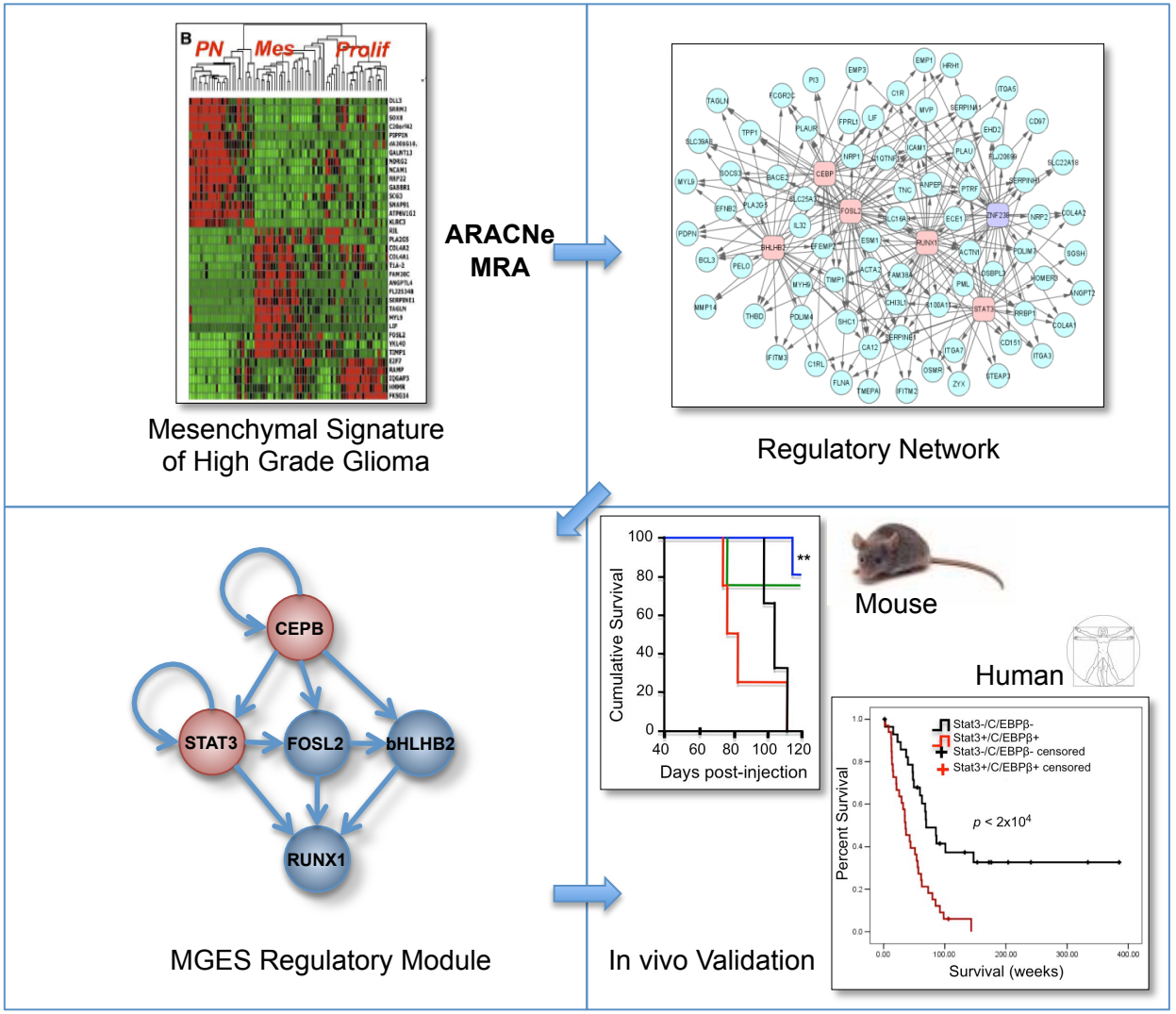
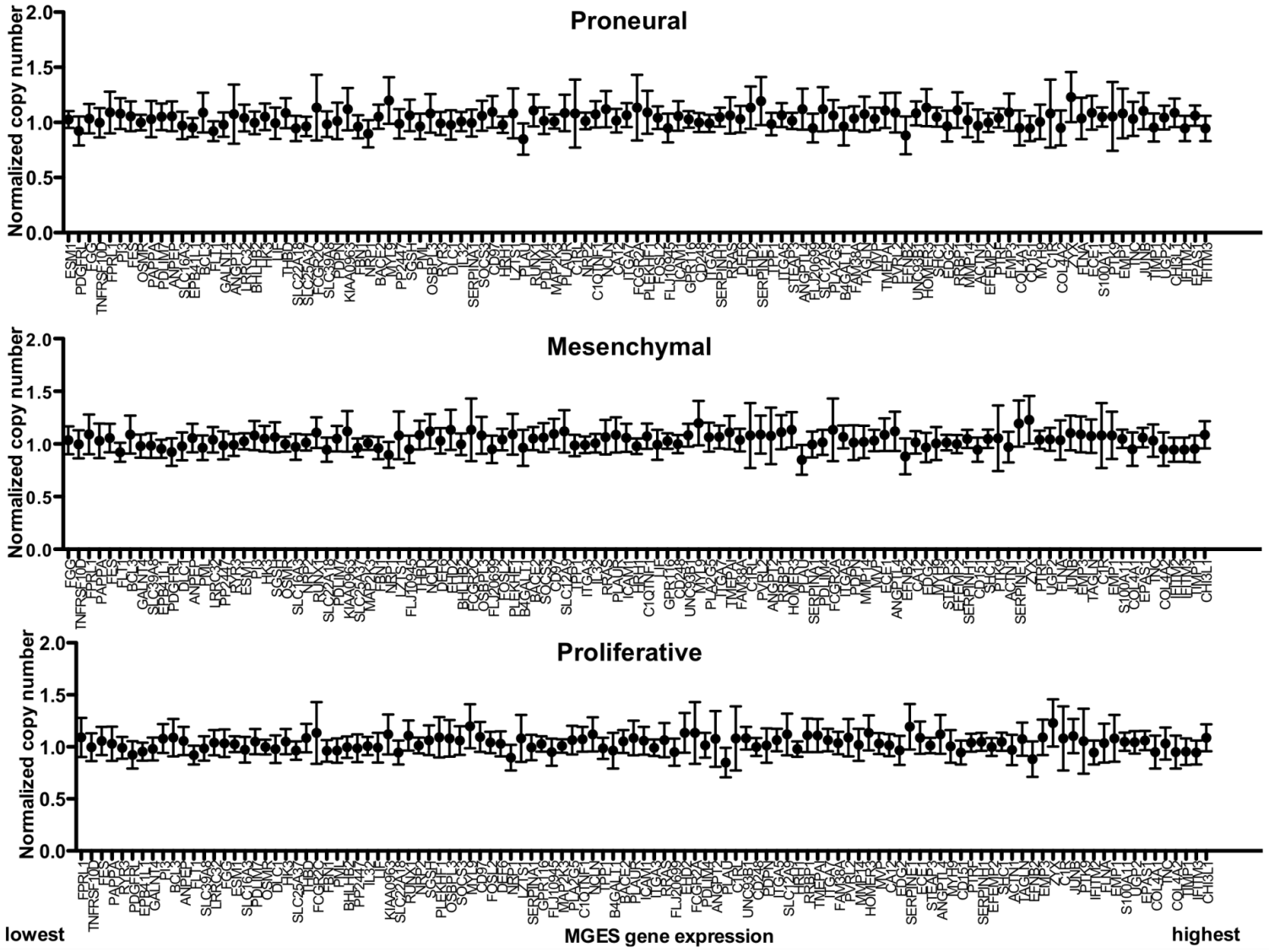
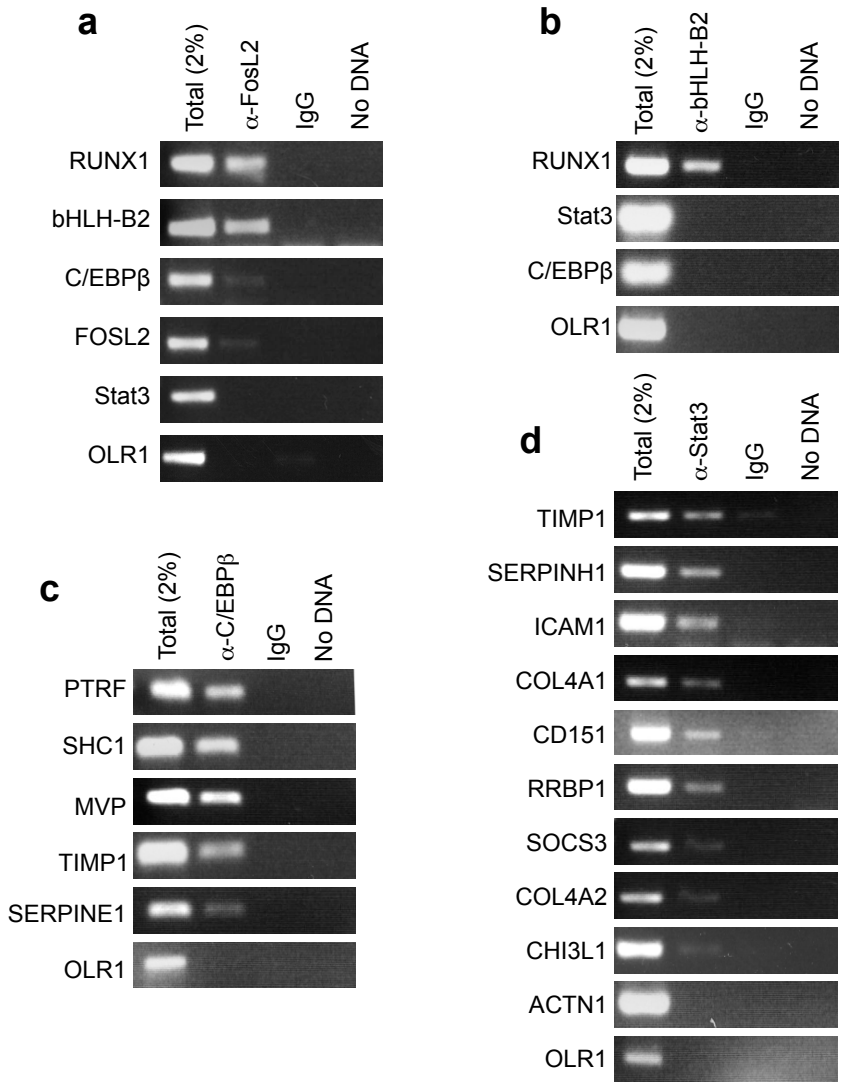


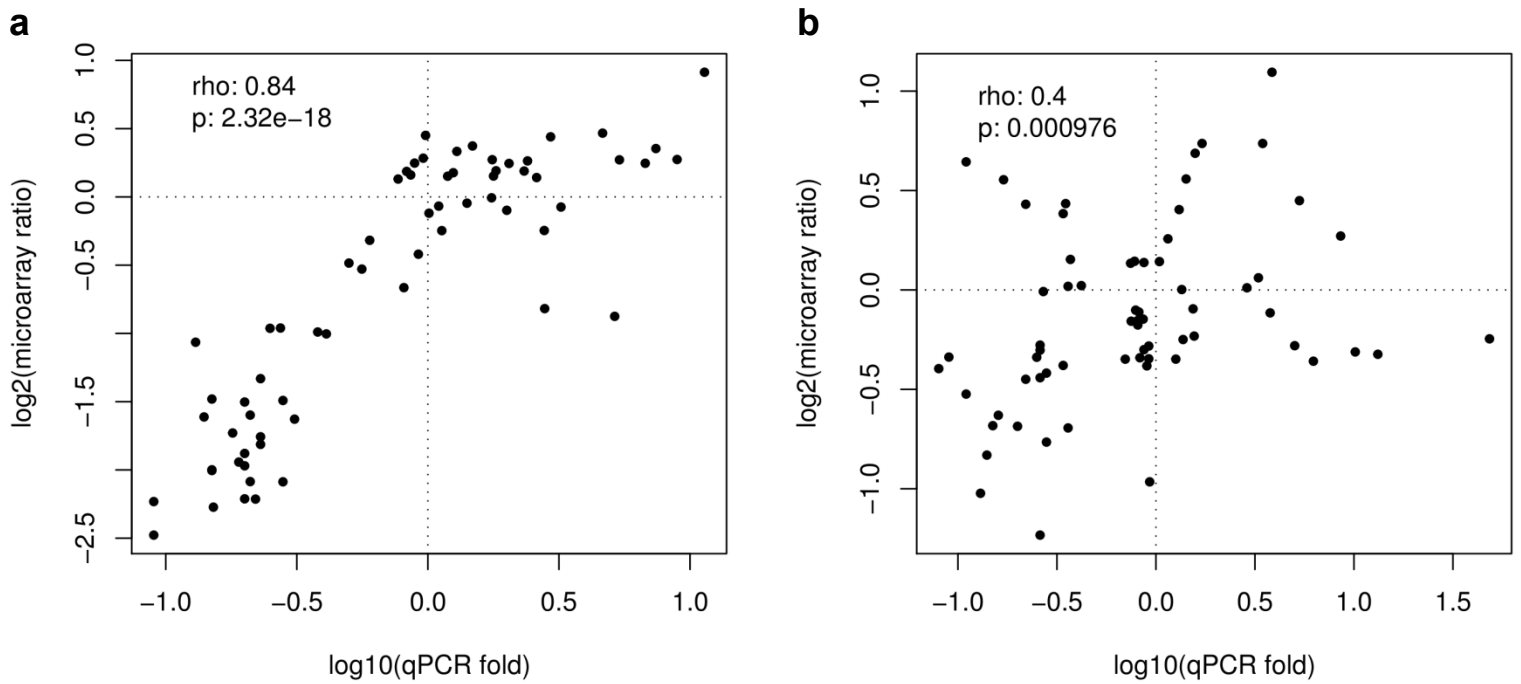
Supplementary Figure 1. Schematic diagram of the experimental strategy used to identify and experimentally validate the transcription factors that drive the mesenchymal phenotype of malignant glioma. Reverse-engineering of a high grade glioma-specific mesenchymal signature reveal the transcriptional regulatory module that activates expression of the mesenchymal genes. Two transcription factors (C/EBPβ and Stat3) emerge as synergistic master regulators of mesenchymal transformation. Elimination of the two factors in glioma cells leads to collapse of the mesenchymal signature and reduces tumor formation and aggressiveness in the mouse. In human glioma, the combined expression of C/EBPβ and Stat3 is a strong predicting factor for poor clinical outcome.



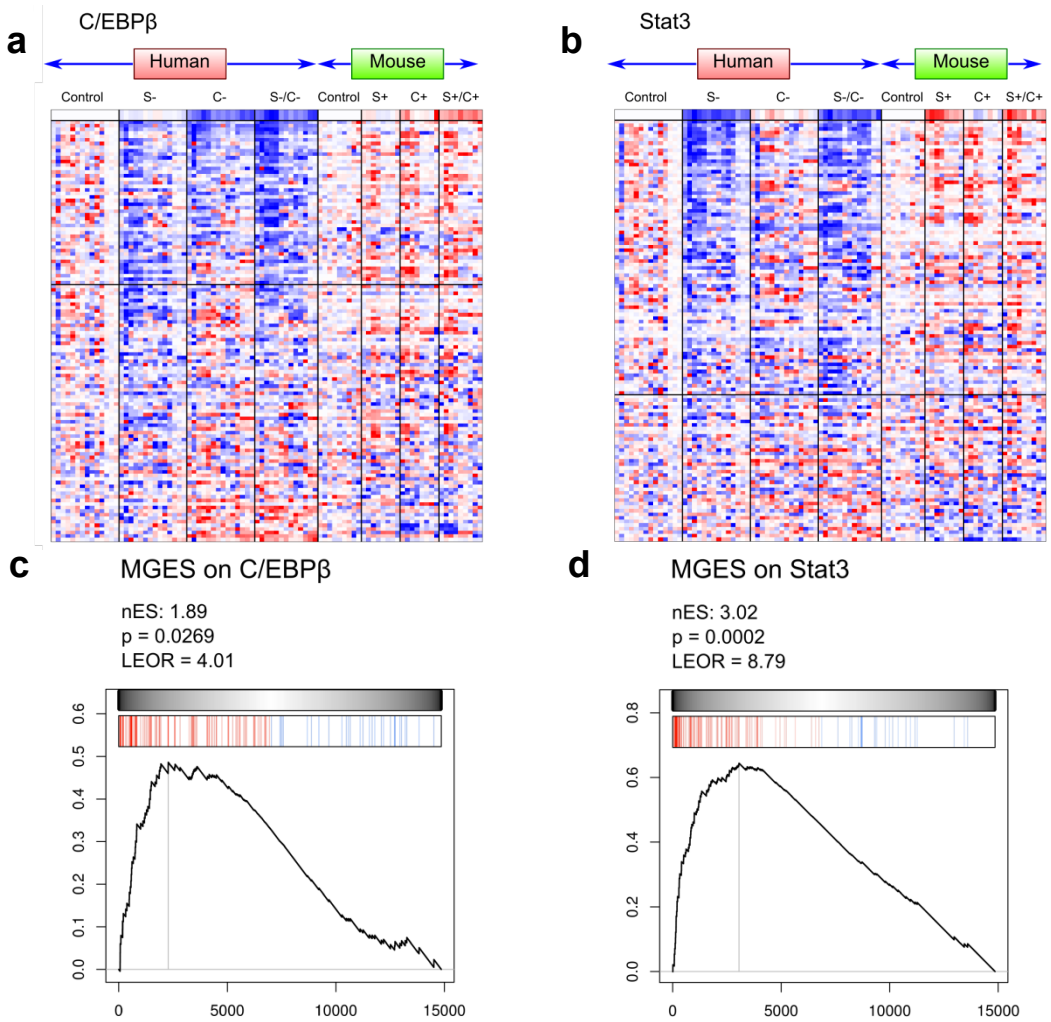
Supplementary Figure 2. Altered MGES gene expression does not result from copy number changes. Genes are shown in order of increasing mean expression and error bars indicate standard deviation of the normalized copy number. No correlation was seen between the MGES gene expression and DNA copy number for the proneural, mesenchymal, proliferative groups or the total cohort (p=0.09430, 0.1058, 0.09430, 0.1014, respectively; Spearman's rho).



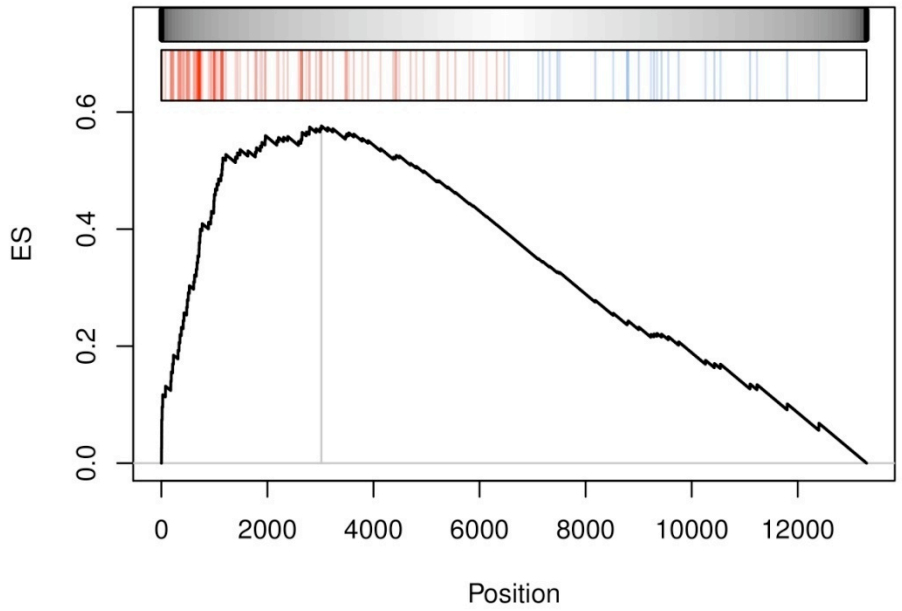
Supplementary Figure 3. Chromatin immunoprecipitation of mesenchymal TFs. Genomic regions of genes containing putative binding sites for specific TFs were specifically immunoprecipitated in the SNB75 cell line by antibodies specific for **a**, FosL2 and **b,** bHLH-B2. Total chromatin before immunoprecipitation (input DNA) was used as positive control for PCR. The OLR1 gene was used as a negative control. **c**, Chromatin immunoprecipitation for C/EBPβ and **d**, Stat3 from primary GBM tumor samples.



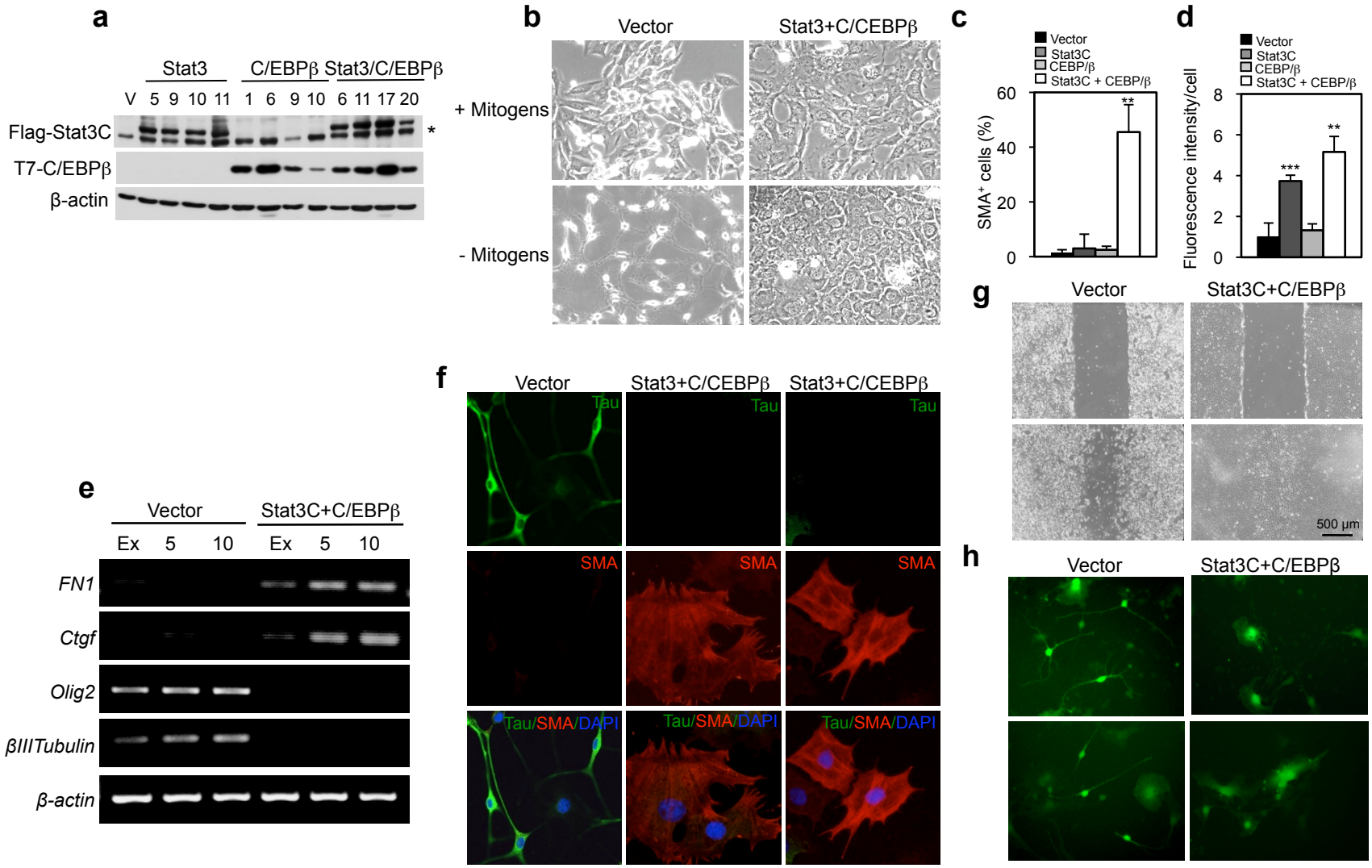
Supplementary Figure 4. Correlation between microarray and qRT-PCR measures for **a**, Stat3 and **b**, C/EBPβ mRNAs. Shown is the ratio of mRNA levels for Stat3 and C/EBPβ between silencing or over-expression and the corresponding non-targeting shRNA or vector controls, respectively. QRT-PCR estimates (x-axis) are in $\log_{10}$ scale, and microarray estimates (y-axis) are in $\log_2$ scale.



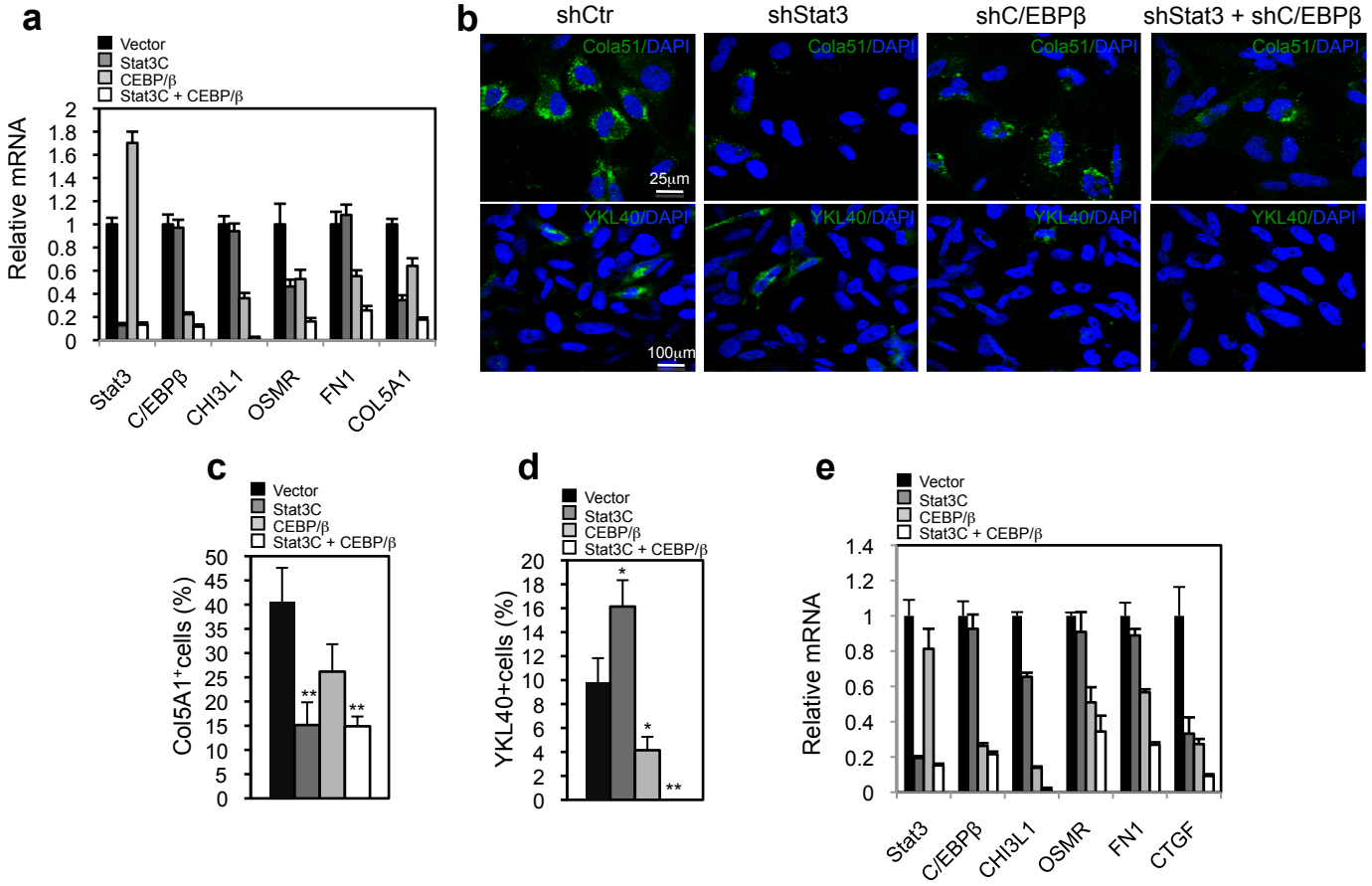
Supplementary Figure 5. Mesenchymal genes are coordinately regulated by C/EBPβ and Stat3. Gene expression integrative analysis of mouse and human cells carrying perturbations of **a**, C/EBPβ and **b**, Stat3. Heatmaps represent mRNA levels for MGES genes. Genes are in rows and samples in columns. The 89 profiled samples were grouped according to species and treatment: control shRNA or empty vector (Control), Stat3 knock-down (S-), Stat3 overexpression (S+), C/EBPβ knock-down (C-), C/EBPβ overexpression (C+), simultaneous knockdown or over-expression of both TFs (S-/C- and S+/C+). The first row of each heatmap shows the mRNA levels of C/EBPβ and Stat3 as assessed by qRT-PCR. Genes were sorted according to the Spearman correlation with the mRNA levels of the specific TF being tested. Blue and red intensity indicate lower and higher expression levels than the gene expression median, respectively. Leading edge mesenchymal genes are above the horizontal black line. **c**, GSEA analysis of the MGES on the gene expression profile rank-sorted according to the correlation with C/EBPβ and **d,** Stat3. The bar-code plot indicates the position of the MGES genes, red and blue colors indicate positive and negative correlation, respectively. The gray scale bar indicates the spearman rho coefficient, used as weighting score for GSEA analysis. nES, normalized enrichment score; p, sample-permutation-based p-value.



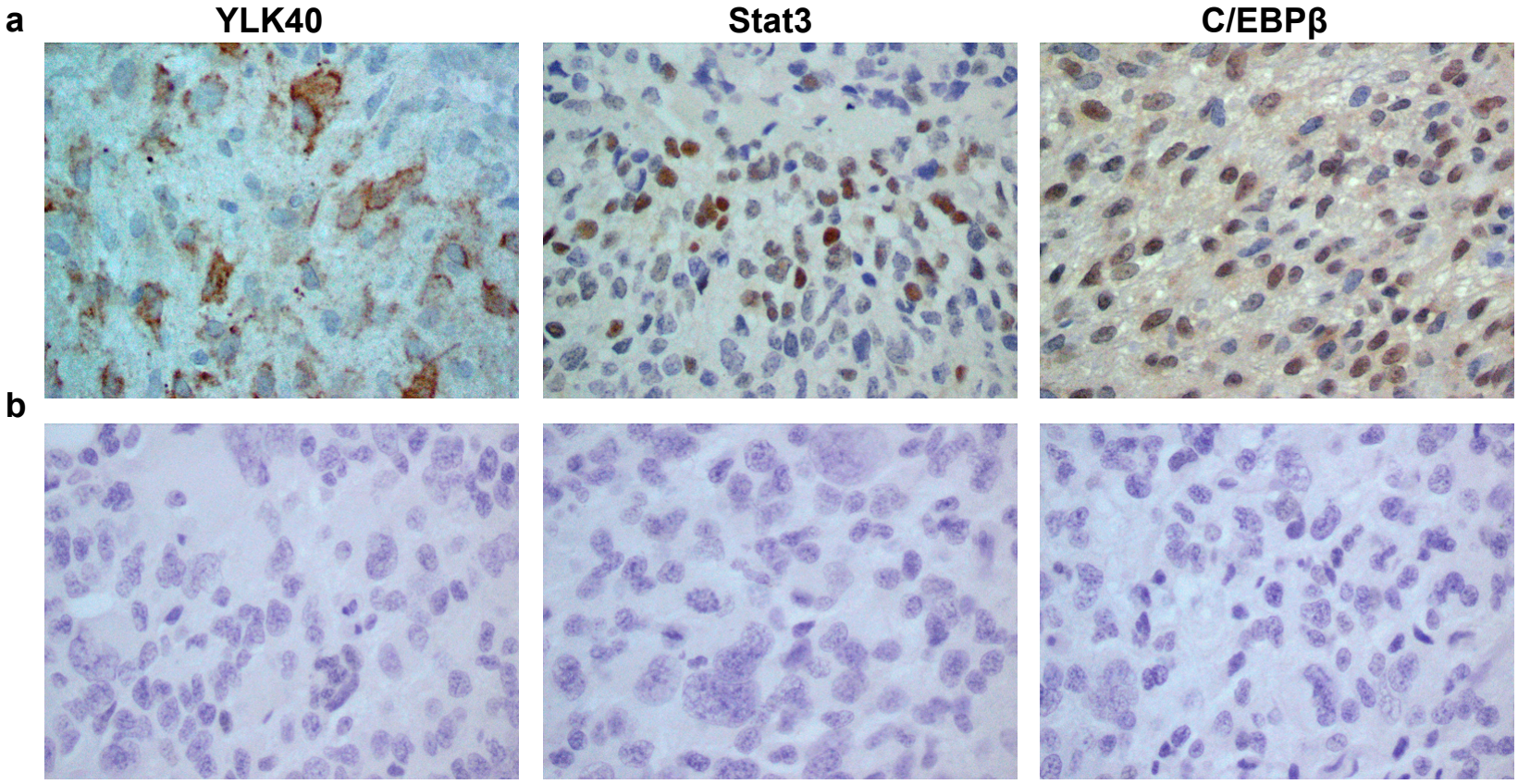
Supplementary Figure 6. **GSEA analysis of the TWPS signature**. GSEA showed that MGES genes were markedly enriched in the TWPS signature. The bar-code plot indicates the position of the MGES genes on the TCGA expression data rank-sorted by its association with bad prognosis, red and blue colors indicate over- and under-expression in the bad vs. good prognosis groups, respectively. The gray scale bar indicates the t-statistics, used as weighting score for GSEA analysis.



Supplementary Figure 7. **Ectopic Stat3C and C/EBPβ in NSCs induce a mesenchymal phenotype and inhibit neuronal differentiation. a,** Western blot analysis of C17.2 clones expressing Flag-Stat3C, T7-C/EBPβ or Flag-Stat3C and T7-C/EBPβ. V, control vector; *, non-specific band. **b,** Morphology of vector-transduced and Stat3 plus C/EBPβ-expressing C17.2 clones grown in the presence or in the absence of mitogens. **c,** Quantification of SMA positive and **d**, Quantification of the fluorescence intensity for fibronectin in C17.2. Bars indicate Mean±SD. n=3 for each group. **e**, Semi-quantitative RT-PCR analysis of mesenchymal and neural markers in C17.2 expressing Stat3C plus C/EBPβ or control vector cultured in growth medium (Ex) or in the absence of mitogens for 5 or 10 days. **f,** Immunofluorescence for Tau and SMA in two C17.2 clones expressing Stat3C and C/EBPβ or control vector cultured in absence of mitogens. **g**, Microphotographs of C17.2 expressing Stat3C and C/EBPβ or the empty vector. 1 mm scratch was made on confluent cultures (upper panels). The ability of the cells to cover the scratch was evaluated after three days (lower panels). **h,** Microphotographs of primary mouse NSCs expressing Stat3C and C/EBPβ or control vector grown in absence of growth factors. $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.



Supplementary Figure 8. C/EBPβ and Stat3 are essential to maintain the mesenchymal phenotype of human glioma cells. a, QRT-PCR of mesenchymal genes in BTSC-3408 infected with lentiviruses expressing Stat3, C/EBPβ, or Stat3 plus C/EBPβ shRNA. n = 3; Bars indicate Mean±SD. **b,** Immunofluorescence for Col5A1 and YKL40 in SNB19 cells infected with lentiviruses expressing Stat3, C/EBPβ, or Stat3 plus C/EBPβ shRNA. **c**, Quantification of Col5A1 and **d**, YKL40 positive cells in experiments in **(b)**. n = 3 from three independent experiments. Bars indicate Mean±SD. **e,** QRT-PCR of mesenchymal genes in SNB19 infected with lentiviruses expressing Stat3, C/EBPβ, or Stat3 plus C/EBPβ shRNA. n = 3; Bars indicate Mean±SD. *$p \leq 0.05$, **$p \leq 0.01$. qRT-PCR data are 18S ribosomal RNA normalized fold changes.



Supplementary Figure 9. **YKL40 expression correlates with C/EBPβ and Stat3 expression in primary tumors.** Immunohistochemical analysis of YKL40, C/EBPβ and Stat3 expression in tumors from patients with newly diagnosed GBM. **a,** Representative YKL40/Stat3C/EBPβ-triple positive tumor. **b,** Representative YKL40/Stat3/C/EBPβ-triple negative tumor.

Supplementary Notes

Supplementary Note 1: Detailed description of the ARACNe, MRA, and SLR analyses

ARACNe was first used to infer the regulon (transcriptional target set) of 928 TFs represented in the gene expression profile. Threshold for MI was set at $p \leq 0.05$ (Bonferroni corrected for multiple TF and multiple target testing) and 0% tolerance was implemented for the DPI analysis. To improve MI estimation, 100 bootstrapping steps were used, as recommended in [1]. The final network was determined by consensus voting, based on interactions that were inferred by a statistically significant number of bootstrapping steps. Statistical significance was determined by expression profile shuffling, as also discussed in [1, 2]. The Fisher Exact Test (FET) was then used to determine the statistical significance of the overlap between each TF's regulon and the MGES. The method identified 53 TFs whose regulon was significantly enriched in MGES genes ($p \leq 0.05$, Bonferroni corrected). These were selected as candidates Master Regulators of the MGES signature and used to determine a regulatory program for each MGES gene, see below.

Then, promoter sequences of MGES genes (foreground) and proneural signature genes (background) were selected to identify TF DNA-binding motifs enriched in the foreground against the background. Motifs were taken from vertebrate subset of the experimentally verified motifs in TRANSFAC Professional and Jaspar database [3, 4]. These motifs were first scored at every position in each promoter, i.e. the [-2kb, + 2kb] region centered on the transcription start site (TSS). The highest score attained on each promoter was tested to determine statistical significance using a threshold optimized to produce the lowest relative error rate (average of the false-positive rate and false-negative rate) in the foreground vs. background promoter set classification [5]. Null distribution of the statistical significance for the relative error rate was computed by random reassignment of members in the foreground and background sets. The method identified 52 additional TFs whose DNA binding motif were significantly enriched in the proximal promoter of MGES genes, compared to the same regions in proneural signature genes, and whose expression had sufficient dynamics across samples, based on the coefficient of variation ($CV \geq 0.5$). These were selected as additional candidate Master Regulators of the MGES signature and used to determine a regulatory program for each MGES gene.

In total, 99 TFs were identified as candidate MRs of the MGES, either by regulon enrichment or by DNA binding motif analysis (six were overlapping in the two lists). These were used to define a regulatory program for each MGES, as a linear combination of the expression level of its candidate MRs, using a stepwise linear regression approach. The model considers the log2-expression level of the target MGES gene as the response variable and the log2-expression level of the candidate

MRs as the explanatory variables. TFs were iteratively added to the model, by choosing each time the one producing the smallest relative error, $E = \sum |x_i - x_{i0}|/x_{i0}$, between predicted and observed target expression. This was repeated until the decrease in relative error was no longer statistically significant ($p \leq 0.05$, Bonferroni corrected). P-value for the non-parametric test was computed by repeating the error calculation 10,000 times using permuted expression values of the selected TF.

Supplementary Note 2: Comparative Analysis

To determine whether the analysis of ARACNe inferred networks using the Master Regulator Analysis (MRA) algorithm could outperform more conventional statistical association methods, we tested whether the transcriptional control module identified by the analysis, and the two master regulators (C/EBP and Stat3) in particular, could also be inferred by these methods.

Differential Expression Analysis: we first tested whether the two MRs and the other TFs in the module could be identified by differential expression analysis. We note, first of all, that C/EBPβ, AND Stat3 are not included among the MGES genes, indicating that their differential expression is not among the most significant in the mesenchymal signature, thus preventing their identification by the statistical association method in [6]. This may be perhaps expected for regulators that are upstream of the signature. Genome-wide differential expression analysis shows that while the MRs C/EBPδ, C/EBPβ and Stat3 are differentially expressed in the mesenchymal samples ($p \leq 0.05$ Bonferroni corrected) they are ranked 123rd, 356th, and 1496th respectively by significance. Thus, the master regulators of the mesenchymal signature and the other TFs in the module could not have been effectively identified by this analysis.

TF DNA-Binding site analysis: We also tried using alternative association-based reverse engineering methods. Specifically, the only method that was proposed in the literature to identify master regulators of cancer signatures is [7], which was not experimentally validated. Rather than inferring transcriptional interactions by molecular-profile analysis, they suggested using the TFs' DNA-binding profiles in TRANSFAC to identify candidate targets. This has three drawbacks. First, only a relatively small number of TFs have low-entropy signatures that can be used to identify high-quality, non-degenerate matches. Second, DNA-binding and regulation are not equivalent, thus resulting in many false positives. Third, TF-binding is obviously non context specific. Thus the inferred interactions would not be high-grade glioma specific. When this method was applied to the MGES analysis, 52 TFs had statistically significant enrichment of their putative DNA-binding sites in the proximal [-2kbp, +2kbp] promoter of the MGES genes ($p \leq 0.05$, Bonferroni corrected). Of the 5-gene transcriptional module, only Stat3 and BHLHB2 were included in this list, ranked in 4th and 35th respectively by statistical significance. Indeed, of the top 20 TFs identified by MRA analysis, only Stat3 and bHLH-B2 could be identified by DNA-binding enrichment

analysis (Suppl. Table 3a, b). Thus, also this method could not reconstruct the module and only one of the two master regulators, the one with the least impact on the phenotype, could be recovered.

Network Analysis by statistical association: Finally, replacing ARACNe with relevance network analysis [8], which uses statistical association to infer regulatory interactions, did not identify C/EBPβ and Stat3 within the 100 most significant genes, based on the MGES-enrichment of their interacting genes.

References

1. Margolin, A.A., et al., *Reverse engineering cellular networks.* Nat Protoc, 2006. **1**(2): p. 662-71.
2. Margolin, A.A., et al., *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics, 2006. **7 Suppl 1**: p. S7.
3. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles.* Nucleic Acids Res, 2003. **31**(1): p. 374-8.
4. Vlieghe, D., et al., *A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.* Nucleic Acids Res, 2006. **34**(Database issue): p. D95-7.
5. Smith, A.D., et al., *DNA motifs in human and mouse proximal promoters predict tissue-specific expression.* Proc Natl Acad Sci U S A, 2006. **103**(16): p. 6275-80.
6. Phillips, H.S., et al., *Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis.* Cancer Cell, 2006. **9**(3): p. 157-73.
7. Rhodes, D.R., et al., *Mining for regulatory programs in the cancer transcriptome.* Nat Genet, 2005. **37**(6): p. 579-83.
8. Butte, A.J. and I.S. Kohane, *Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.* Pac Symp Biocomput, 2000: p. 418-29.