



Figure 3 (A) A heatmap and bi-clustering of a set of important genes in NP development. Colors indicate the normalized expression levels. Data are gene-wise standardization (zero-mean and unit variance) of \log_2 FPKM (same in G). (B) A heatmap showing the pairwise similarity among samples before and after differentiations as measure by genome-wide Pearson correlation coefficients. The samples were also clustered by hierarchical clustering based on the similarity matrix. (C) A violin and box-plot showing the increased genome-wide similarity to a set of four *in vivo* NP samples taken from surgical samples of the human intervertebral discs. The Student's *t* test p-value is 1.29×10^{-07} . (D) A scatter-plot of all genes before (grey, averaged over three cell-lines) and after differentiations (light-green, cyan and pink; averaged over 12 replicates). The most influential genes, whose removal significantly reduce the similarities, were highlighted by an arrow pointing from its expression level in the hESC/iPSCs to its expression in the *in vitro* differentiated NPCs. Left arrows indicate negative influential genes: those contributing to increased similarities by having lower expression in the *in vitro* differentiated NPCs; right arrows indicate positive influential genes: those contributing to increased similarities by having higher expressions in the *in vitro* differentiated NPCs. (E) A Venn diagram showing the overlaps among differentially expressed genes (DEGs) between the *in vivo* NP and hESC/iPSCs, and those between *in vitro* differentiated NPCs and hESC/iPSCs, and the positive influential genes identified in (D). (F) A gene-ontology (GO) analysis of the 148 positive influential genes in overlap shows that they are enriched for various types of ECMs and their regulatory pathways, key for normal NP functioning. (G) motif enrichment analysis of the 148 genes identifies motifs, and their corresponding transcription factors. (H) Same comparison as in (C), but now using the influential genes only ($p=3.9 \times 10^{-27}$). (I) A heatmap and bi-clustering of the 148 genes in overlap.