# Regulatory Network Inference Method

## Overview

We propose here a general method for identifying regulatory players during cell differentiation process by detecting expression (Figure 1a) and ChIP-Seq transcription factor binding site (Figure 1c) changes between pairs of cells: before and after a commitment steps (ie. MPP vs. CDP) or between committed cell types (i.e. cDC vs. pDC). The ChIP-Seq data is based on the master regulator PU.1, which works by recruiting co-binding partners to define cell specific expression partners (Lee, 2013 and Heinz, 2010). Therefore, we further complement the analysis by inclusion of sequence based motif search to detect transcription factors binding sites that co-localize with PU.1 ChIP-Seq peaks (Figure 1d). The selection of relevant transcription factors out of the hundreds of factors with known motif is based on their expression (Figure 1a), presence of their motif in public databases (Figure 1b) and an enrichment analysis (Figure 1e), i.e. we only use factors which have high number of binding sites near PU.1 peaks close to differentially expressed genes in a cell type. Finally, we construct a cell specific regulatory network by drawing links from TF with binding sites close to PU.1 peaks to putative target genes in the vicinity of the PU.1 peak (Figure 1f).

## Differential Genes and Transcription Factors

For a pair of cell types, the first step is to define the set of differentially expressed genes (DE) between the two types as previously explained in Section XXX. For example, Spib is under-expressed comparing pDC to cDC and Bcl6 is over-expressed comparing pDC to cDC transition (Figure 1a). Next, we list all transcription factors in the list of differential genes and retrieve their motifs from the databases Jaspar version Vertebrates 2009 (Brayne, 2008) and Uniprobe Mouse 2011 (Newburger, 2009) (Figure 1b). We also include motifs from relevant factors, which were not present in the previous databases. AICS Irf8 motif was obtained by running the tools Meme-ChIP with default parameters (Machanic, 2010) in the Irf8 ChIP-Seq peaks provided by Marquis, 2011. The Irf8 ISRE and IEACS motifs were obtained with the sequences provided in Kanno, 2005. Lastly, we have used the Klf4 and E2A motifs derived from ChIP-Seq experiments (Lin, 2010 and Lister, 2009) available under the homer (<http://biowhat.ucsd.edu/homer/>). We ignored secondary motifs from the Uniprobe database, as they mostly had low information content. From the original 533 motifs, 53 were kept after the filtering for differential expression. The position frequency matrixes used in this study are available in our online resources webpage (XXX).

## Differential PU.1 Peaks

The next step is to search for differential PU.1 peaks between two cell types. Note that this procedure aims at detecting peaks that have a significant increase or decrease in size during the cell commitment step. This cannot be performed trivially by usual ChIP-Seq peak calling methods as MACS (Zhang, 2008), as they concentrate on detecting the presence or absence of the peak. To detect differential peaks (DP), we first run MACS with default parameters for each cell type using the input signal as control. We obtain the positions of the peak summits and extend the regions to have a total length of 500 bps. This extension is performed for the latter detection of factors co-binding with Pu1. Next, we make a union of all peaks of all cell types. If peaks from distinct cell types overlap, they will originate a single peak covering the region of both original peaks. We call this the complete set of PU.1 peaks. Next we estimate the coverage of PU.1 reads for each cell type within each of the peaks in the complete set. The coverage values for each cell type are used as input for a digital differential analysis. For such, we use the exact empirical test from EdgeR (Robinson, 2010) to compare the peaks between pairs of cell types (MPP vs. CDP and cDC vs. pDC). A peak is called differential if they have a significant fold change; p-value of 0.01 after Benjamini-Hochberg multiple test correction (Benjamini, 1995). We filter peaks with low read support (logCPM >3.0).  Finally, peaks are assigned to genes by a proximal and distal rule. The proximal rule assigns a peak to a gene if it lies inside the gene body or on its proximal promoter (1Kbps upstream of the TSS). The distal rule assigns a peak to a gene if it lies 50Kbps away from the TSS and there is no other gene’s TSS in between. Examples of differential peaks are seen in Figure 1c. For Spib, a PU.1 peak near the TSS decreases from pDC to cDC, while for Bcl6 a peak near the TSS increases from the pDC to cDC.

## Enrichment Analysis for PU.1 Combinatorial Binding

To explore potential factors co-binding with PU.1, we search for binding sites for all motifs from transcription factors, which are differentially expressed in one of the cell types. Binding site detection was performed within PU.1 differential peaks close to differentially expressed genes, i.e. peak increase close to over-expressed genes or peak decrease close to under-expressed genes. For example, comparing pDC to cDC we use the genomic regions of increased PU.1 peaks near genes over-expressed from pDC to cDC, such as Bcl6 in Figure 1c. Conversely, we use genomic regions of peaks decreased from pDC to cDC close to genes with under-expression comparing pDC to cDC, such as Spib in Figure 1c. The motif search is performed using Biopython (Cook, 2009), which calculates a bit score based on the application of a Position Weight Matrix (PWM) in the regions mentioned previously. In order to create the PWMs, a pseudocount of 0.1 is added for all nucleotides and the default background information from Biopython is considered. To define a bit score threshold, a distribution of scores is estimated with a dynamic programming approach, with the precision set to 104. Then, the threshold corresponds to the bit score value that represents a FDR of 10-4 on the estimated distribution. We randomly sample sequences from the genome with 10 times the number of PU.1 differential peaks to obtain random binding sites. Finally, we employed a one-tailed Fisher exact test to measure if the proportion of differential PU.1 peaks close to DE genes with at least one binding site of the motif is higher then the proportion of PU.1 peaks with at least one binding site of the motif in the random regions. The test is repeated for all motifs and cell specific differential peaks. Final p-values were corrected by the Benjamini-Hochberg method (Benjamini, 1995), using an alpha value of 0.05. The corrected p-values (or enrichment scores) are visualized as a heatmap, where dark colors indicate the TFs potentially co-binding with PU.1 at a particular cell type (Figure 1 e).

## Cell Specific Regulatory Networks

As a last step, we construct a cell specific regulatory network based on the binding site prediction from the previous analysis. We include in the network (1) all factors with motifs that were enriched in at least one of the cell types, (2) a list of genes with previously implicated to be functional in dendritic cells (Felker, 2010; Satpathy, 2011; Satpathy, 2012) and (3) through the analysis of our gene expression data. See supplementary table XXX for a list of these genes, the DC development stage they are putatively active and the source of the functional annotation.

In the network, we display nodes in red, if they are TF with known motif and they are enriched in that particular cell type. Rosa nodes indicate TFs with known motif, which is not enriched in that particular cell type. Lastly, grey nodes indicate the target genes included based on the literature analysis. Note that we also include in the list of target genes transcription factors. Whenever there is no motif available (i.e. Zbtb46) or they do not bind to DNA (i.e. Id2).

The network edges for a particular cell type are constructed by the following rules. Whenever an enriched TF (red node) has a binding site on a differential PU.1 peak close to a DE gene, we draw a edge from the TF to the DE gene. For example, in the PU.1 increased peak close to Bcl6, we have PU.1, Klf4, Irf1, Ap1 and Spi1 binding sites (Figure 1 d). Of those, PU.1, Ap1 and Klf4 have a significant enrichment for the cDC (Figure 1 e), therefore we draw regulatory edges from PU.1, Ap1 and Klf4 to Bcl6 (Figure 1 e). Analogously, we have an edge from PU.1, Spib and Tcf4 to Spib for the pDC network, as we have a PU.1, Spib and Tcf4 binding sites in a peak decreased from pDC to cDC and PU.1, Spib and Tcf4 were enriched for pDC (Figure 1 e). We use Cytoscape software for visualization of the networks (Cline, 2007).