SigSeeker: An Ensemble for Analysis of Epigenetic Data

Jens Lichtenberg, Elisabeth F. Heuston, David M. Bodine
National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
Project and Sourcecode: http://sigseeker.org (GNU General Public License, version 3.0)

Epigenetics is the study of the proteins associated with (e.g. DNA binding proteins) and modifications to (e.g. DNA methylation) the primary DNA sequence. The epigenetic landscape of a nucleus determines which genes will be expressed or silenced and regulates the different genetic program of cells in an organism. There are numerous approaches for the analysis of next generation sequencing data that can be applied to define the epigenetic landscape. We believe that the quality of epigenetic analyses can be increased through an integrative framework that correlates the user-generated output of multiple prediction tools with existing biological data. We found that few sequence analysis tools integrate multiple approaches and consequently the existing techniques suffer in their prediction quality.

To address this problem we developed SigSeeker, a computational framework designed for: 1) the mapping of sequencing reads against a reference genome, 2) the detection of epigenetic mark enrichment within a set of mapped reads, and 3) the correlation of these sites with previously annotated expression and epigenetic data. By considering the complete set of established expression and epigenetic data during the analysis process, SigSeeker overcomes the shortcomings of other single technique approaches. SigSeeker incorporates commonly applied epigenetic tools using ensembles for each analysis stage. SigSeeker allows comparisons of user-generated data, as well as correlations of these data to publicly available epigenetic and expression data. The predictions made by each of the modules in the SigSeeker framework are evaluated for their statistical significance during the each stage of the analysis process as well as in a final report. SigSeeker is validated using benchmarks for ChIPSeq and HistoneSeq benchmark. These comparisons indicate that our ensemble technique exceeds single approaches (300% sensitivity increase) and is highly relevant for epigenetic data analysis.

We applied SigSeeker to study genome wide patterns of DNA methylation and gene expression in primary mouse blood cells. We found that DNA methylation was most abundant in the most primitive hematopoietic stem cells (HSC), declined in more differentiated common myeloid progenitors and declined further in nucleated red blood cells. In contrast to nucleated red blood cells, DNA methylation was increased in platelet forming cells to levels similar to HSC. The adjacent regulatory regions of the genes transcribing RNA were found to be hypomethylated in all cell types, while DNA methylation in the gene itself positively correlated with gene expression. In genes transcribing non-coding (regulatory) RNAs, DNA methylation in the gene itself was not found. We expanded our analysis to include the DNA binding proteins GATA1 and NFE2. GATA1 and NFE2 occupancy was cell-type specific. In platelet forming cells DNA methylation and NFE2 binding in the gene itself was associated with RNA-producing genes, while in nucleated red blood cell genes, DNA methylation and GATA1 binding in the gene itself was associated with inactive genes.

In summary, using our novel SigSeeker ensemble, we demonstrate that epigenetic modifications change dramatically during hematopoiesis and we have identified critical regions of the genome that regulate hematopoietic cell fate. Our ensemble technique exceeds single technique approaches in prediction quality and is ideal for identifying high confidence epigenetic profiles.