Motif analysis in DNAse hypersensitivity regions uncovers distal cis elements associated with gene expression

Mark Ziemann¹,²*, Antony Kaspi¹,², Ross Lazarus⁵ & Assam El-Osta¹,²,³,⁴

¹Epigenetics in Human Health and Disease Laboratory; ²Epigenomics Profiling Facility, Baker IDI Heart & Diabetes Institute, The Alfred Medical Research and Education Precinct, Melbourne, Victoria 3004, Australia; ³Department of Pathology, The University of Melbourne, Victoria 3010, Australia; ⁴Central Clinical School, Department of Medicine, Monash University, Victoria 3800, Australia; ⁵Medical Bioinformatics, Baker IDI Heart & Diabetes Institute; Mark Ziemann – Email: Mark.Ziemann@bakeridi.edu.au; *Corresponding author

Received January 22, 2013; Accepted January 29, 2013; Published February 21, 2013

Abstract: Reliable identification of cis regulatory elements influencing transcription remains a challenging problem in molecular bioinformatics. This is especially true for enhancer elements which are often located hundreds of kilobases from the gene promoter. High resolution DNAse hypersensitivity and connectivity profiling by the ENCODE consortium provides evidence of millions of interacting cis-acting elements in the human genome. This prior knowledge can be incorporated into genome-wide expression analyses, in the form of gene sets sharing regulatory sequence motifs in known DNAse hypersensitivity peak regions. High proportions of enrichment among the most extreme differentially transcribed genes from controlled biological experiments may suggest novel hypotheses about signalling pathways. The utility of this approach is demonstrated with the reanalysis of a microarray-derived gene expression data set through the Gene Set Enrichment Analysis pipeline, uncovering new putative distal cis elements in the context of innate immunity. The DNAse Hypersensitivity Connectivity informed Motif Enrichment in Gene Expression (DHC-MEGE) method described here has the advantage of identifying distal elements such as enhancers, which are often overlooked with standard promoter motif analysis.

Availability: The DHC-MEGE shell script can be obtained from Sourceforge (https://sourceforge.net/projects/dhcmege/) and the generated GMT file is attached as supplementary data.

Key words: DNAse hypersensitivity, motif enrichment, gene expression, enhancer, gene set enrichment analysis.
are over-represented, in up- and down-regulated genes. These motifs can serve as a proxy for DNA-binding transcription factors (TFs) and other sequence-specific elements that regulate gene expression. Restricting MEGE to the immediate vicinity of known transcriptional start sites (TSS) ignores cis regulatory elements outside of the promoter, while other approaches using sequence conservation overlook cis elements which may be unique to a species [3, 4]. The MEGE method described here limits motif discovery to DH regions and takes advantage of the DH connectivity map provided by ENCODE, enabling the identification of long-range distal cis elements which potentially drive changes in expression [5]. Genes that share an exposed motif are collected into individual gene sets and can be appended to existing motif gene collections [6] and tested using conventional Gene Set Enrichment Analysis [7].

Utilities to the biological community:
Gene sets provide a convenient way to systematically summarise complex empirical biological information such as GSEA analysis modules for genome-wide profiling experiments. Motifs identified as significant from GSEA could represent the activation or repression of DNA binding factors that drive gene expression changes. The most significant advantage of this approach over current methods is the ability to identify long-range distal cis elements, which can only be achieved by incorporating chromatin connectivity maps. To demonstrate the utility DHC-MEGE, we have re-analysed a microarray expression data set investigating the effect of 2 hr lipopolysaccharide (LPS) treatment on the THP-1 immortal monocyte cell line (GEO: GSE32141) [9]. The publically available dataset was examined for differential expression with GEO2R. There were 590 up-regulated and 714 down-regulated probes assigned to genes (nominal p≤0.01). We used the ENCODE human DH connectivity map with a correlation threshold of 0.8, a sequence similarity threshold of 10 and a maximum gene size set of 1000.

After motif identification and genome-wide screening, there were 105 de novo motif gene sets and 150 known motif gene sets (Figure 1B). Probes with a signal above the detection threshold and annotated with a valid RefSeq gene identifier were submitted to GSEA with the newly generated GMT files. GSEA identified 20 motifs to be enriched in the up-regulated genes (FDR adj p-values≤0.05), including known NF-kB motifs, but there were also several high-ranking de novo motifs such as TATGACAATC (Figure 2). The ADORA2A gene is the most highly upregulated gene associated to the TATGACAATC motif, with the motif occurring in a DH region 274 kbp upstream of the ADORA2A promoter DH site within a CABIN1 intron. This distal DH site (chr22:24549440-24549590) is bound by FOXA1 and USF-1 transcription factors, while USF-1 is also found at the ADORA2A promoter according to ChIP-seq profiling, suggesting that USF-1 might be an adapter protein in a chromatin loop [10]. The ADORA2A example also highlights the combinatorial contribution of distal elements, with the two ADORA2A promoters associated with regions containing FOXA2, GATA-IR3 and CTTCAGTAAGTT elements that are significantly associated with up-regulated genes (FDR-adjusted p-values of 0.013, 0.22 and 0.047 respectively). This example highlights the biological complexity of long-range chromatin interactions that are overlooked by current promoter motif analysis tools.

Figure 1: (A) Schematic diagram of DHC-MEGE methodology of generating custom motif-gene association sets with the file format shown in brackets; (B) Statistics from an example analysis using a publicly-available microarray expression data set with a correlation threshold of 0.8, minimum sequence similarity score of 10 and maximum gene size of 1000.

Methodology:
A schematic workflow of the DHC-MEGE program is shown in (Figure 1A). The DHC-MEGE program utilizes lists of up- and down-regulated gene symbols for motif analysis. It also requires a DH connectivity map in BED8 format, which describes the interaction between a promoter and distal elements together with the connection correlation coefficient. The human ENCODE DH connectivity map is accessible from EBI [5]. The program extracts the DH regions associated to the ‘up’ and ‘down’ gene expression sets. These BED regions are then searched for significantly enriched 20 nt sequence motifs with the Hypergeometric Optimization of Motif EnRichment program (HOMER) compared to the background genome sequence. HOMER initially utilises a library of known motifs, then performs de novo enrichment analysis for further enriched motifs and outputs these as a position weight matrix (PWM). Instances of these motifs are then counted in DH regions genome-wide, scored and sorted based upon their degree of sequence similarity to the HOMER PWM [8]. Users are able to fine-tune the analysis by setting a DH site correlation threshold, gene set size as well as motif sequence similarity thresholds. Paths to input files and parameter settings are provided in a properly formatted configuration file. The output from DHC-MEGE is a list of gene sets in the GMT format. Because initial motif searching is restricted to sets of differentially expressed genes, the motifs identified will be unique for each experiment. Gene lists generated in the GMT format are accepted by the freely available Gene Set Enrichment Analysis (GSEA) software [7].
Figure 2: Example GSEA using DHC motif gene sets for gene expression analysis. The LPS-stimulated THP-1 cell microarray data is publically available [10]. (A) The top 10 ranked motif gene sets in up-regulated genes contains known and novel motifs (ranked by FDR-adj p-value). Abbreviations: SIZE, number of genes in the set detected in the experiment; ES, enrichment score; NES, normalised enrichment score; NOM p-val, nominal p-value; (B) Enrichment plot of the newly identified TATGACAATC motif gene set, showing the majority of genes associated with this motif are highly up-regulated; (C) List of the top 10 up-regulated genes associated with the TATGACAATC motif; (D) Example of long-range cis elements interacting with the ADORA2A promoter. The TATGACAATC motif is positioned in a distal DH peak 274 kbp upstream of the ADORA2A promoter DH peak (red line) in a CABIN1 intron. ADORA2A is also associated with GATA-IR3 (orange), CTTACGTAAGTT (blue), FOXA1 (pink) motifs that were significant in GSEA analysis (FDR<0.05).

Conclusion:
The contribution of long-range chromatin interactions to the control of gene expression remains poorly understood, but this will improve as higher-resolution maps from chromatin conformation profiling experiments are integrated. Tools such as DHC-MEGE may also assist researchers understand the complex interplay between gene expression and epigenetic marks such as DNA methylation in disease.

Acknowledgement:
The authors acknowledge support from the Juvenile Diabetes Research Foundation International (JDRF), the National Health and Medical Research Council (NHMRC), and the National Heart Foundation of Australia (NHF). AE-O is a Senior Research Fellow supported by the NHMRC. Supported in part by the Victorian Government’s Operational Infrastructure Support Program.

References:
[1] Crawford GE et al. Genome Res. 2006 16: 123 [PMID: 16344561]
[2] Thurman RE et al. Nature. 2012 489: 75 [PMID: 22955617]
[3] Xi H et al. Genome Res. 2007 17: 798 [PMID: 17567998]
[4] Xie X et al. Nature. 2005 434: 338 [PMID: 15735639]
