ELSEVIER

Experimental Hematology

Experimental Hematology 2013;41:354–366

Gene set control analysis predicts hematopoietic control mechanisms from genome-wide transcription factor binding data

Anagha Joshi, Rebecca Hannah, Evangelia Diamanti, and Berthold Göttgens

Department of Hematology, Cambridge Institute for Medical Research and Wellcome Trust and MRC Cambridge Stem Cell Institute, Cambridge University, Hills Road, Cambridge, UK

(Received 18 October 2012; revised 29 November 2012; accepted 29 November 2012)

Transcription factors are key regulators of both normal and malignant hematopoiesis. Chromatin immunoprecipitation (ChIP) coupled with high-throughput sequencing (ChIP-Seq) has become the method of choice to interrogate the genome-wide effect of transcription factors. We have collected and integrated 142 publicly available ChIP-Seq datasets for both normal and leukemic murine blood cell types. In addition, we introduce the new bioinformatic tool Gene Set Control Analysis (GSCA). GSCA predicts likely upstream regulators for lists of genes based on statistical significance of binding event enrichment within the gene loci of a user-supplied gene set. We show that GSCA analysis of lineage-restricted gene sets reveals expected and previously unrecognized candidate upstream regulators. Moreover, application of GSCA to leukemic gene sets allowed us to predict the reactivation of blood stem cell control mechanisms as a likely contributor to LMO2 driven leukemia. It also allowed us to clarify the recent debate on the role of Myc in leukemia stem cell transcriptional programs. As a result, GSCA provides a valuable new addition to analyzing gene sets of interest, complementary to Gene Ontology and Gene Set Enrichment analyses. To facilitate access to the wider research community, we have implemented GSCA as a freely accessible web tool (http://bioinformatics.cscr.cam.ac.uk/GSCA/GSCA.html).

Cell type–specific gene expression is an inherent property of all multicellular organisms and indeed represents a major determinant that underlies the generation of differentiated cell types with distinct functionality. Elucidating the molecular mechanisms controlling cell type–specific expression has the power to reveal fundamental insights into the regulatory circuitry controlling both human and model organism development. Moreover, identification of control mechanisms in normal cells provides potential avenues for manipulating cellular fates, as exemplified by the recent explosion in cellular reprogramming studies [1]. It also enables the rational design of new therapies aiming to revert abnormal pathological cellular states back to their normal condition [1].

The blood or hematopoietic system has long been recognized as a powerful model system for studying cell type–specific gene expression [2]. Within the blood system, more than 10 distinct mature hematopoietic lineages (e.g., red blood cells, T cells, B cells) are generated from pluripotent hematopoietic stem cells (HSCs) via a sequence of intermediate progenitors, often represented as a lineage differentiation tree. Both the mature lineages as well as the various immature blood stem and progenitor populations can be purified based on the expression of combinations of specific cell surface markers, thus enabling powerful studies of cellular differentiation.

Transcription factors have long been recognized as major regulators of hematopoietic cell type specification [3–6]. To understand the mechanisms underlying cell type specification by transcription factors, it will be essential to identify their transcriptional targets. An important advancement in this research area was provided by the introduction of chromatin immunoprecipitation (ChIP) coupled to massively parallel sequencing (ChIP-Seq),

Offprint requests to: Anagha Joshi, Department of Hematology, Cambridge Institute for Medical Research, Cambridge University, Hills Road, Cambridge, CB2 0XY, UK; E-mail: aj379@cam.ac.uk and Berthold Göttgens, Department of Hematology, Cambridge Institute for Medical Research, Cambridge University, Hills Road, Cambridge, CB2 0XY, UK; E-mail: bg200@cam.ac.uk

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.exphem.2012.11.008.

0301-472X/$ - see front matter.
http://dx.doi.org/10.1016/j.exphem.2012.11.008

which allows genome scale identification of all DNA sequences (regions) bound by a given transcription factor (TF) in a given cell type [7]. The technique has been rapidly adopted with over 100 individual studies now deposited in public databases for the murine hematopoietic system alone. This wealth of new data represents unprecedented opportunities to unravel the transcriptional control mechanisms that mediate expression of specific sets of genes within the various hematopoietic cell lineages [8].

Gene ontology [9] overrepresentation analysis provides information on various types of functional categories enriched within a given gene set of interest [10] and GSEA determines whether a gene set of interest shows statistically significant expression differences between two or more cell types [11]. However, neither of these approaches explicitly links a gene set to transcriptional control mechanisms. In this study, we report a new computational framework for linking gene sets with transcriptional control, called Gene Set Control Analysis (GSCA). Unlike previous algorithms developed to provide functional enrichment [10], GSCA links gene sets to likely upstream regulators responsible for coordinated expression. By exploiting multiple transcription factor binding patterns from genome-wide ChIP-Seq studies, GSCA can provide previously unattainable insights into possible transcriptional control mechanisms operating in both normal and malignant cells. To gain insights into combinatorial control mechanisms (i.e. multiple transcription factors occupying the same binding site in a gene locus), we further developed a novel tool called combinatorial-GSCA (C-GSCA). Through integrated analysis of 142 blood-specific ChIP-Seq binding datasets, C-GSCA identifies likely combinatorial transcriptional control mechanisms by revealing TF cooccupancy patterns specifically associated with gene regulatory elements from a given gene set. A web-based implementation of GSCA and C-GSCA allows user-friendly access for the wider research community, and thus provides a substantial new addition to the bioinformatic toolbox for hematopoietic gene set analysis.

Methods

ChIP-seq compendium

Binding events for 35 transcription factors in seven major hematopoietic lineages were obtained from Hannah et al. [8]. Sixty new ChIP datasets from 18 publications and ENCODE murine datasets were analyzed, starting from the raw data set in each case, and peaks were identified in each sample using the protocol described previously [8]. A supplementary website (http://bioinformatics.cscr.cam.ac.uk/BLOOD_compendium_PUBLISHED.html) lists the number of peaks, reference, and peak calling method for each of the ChIP dataset. All binding events were mapped to genes using the same protocol described previously [12]. Binding events in the promoter and gene body were associated to the corresponding gene, whereas intergene peaks were associated to the nearest gene on either side within 50 kb, such that each peak is assigned to at most two genes.

Tissue-specific enhancer elements in mouse were downloaded from [13] and p value was calculated for overlap between each of the 61 tissue-specific enhancer regions and blood-specific regulatory regions [8] using a hyper-geometric test (Supplementary Table 1, online only, available at www.exphem).

GSCA method

Of 270,261 genomic regions bound by at least one TF (N), for a set of user-defined genes, we calculate the number of genomic regions mapped to the genes (n). For each ChIP-Seq ChIP dataset, the number of peaks (m) near user defined genes (k) is calculated. The p value is calculated using a hypergeometric test (Fischer exact test).

cGSCA method

A matrix of binding events with 270,261 genomic regions as rows and overrepresented ChIP-seq data sets (K) from GSCA step as columns is generated. The ChIP-seq data sets (K columns) are then clustered using a hierarchic clustering with Pearson's correlation coefficient as a distance measure.

Reference data set

Gene sets for 80 clusters of tightly coexpressed genes (their induction patterns) in 38 hematopoietic cell types were obtained from Novershtern et al. [14]. Human genes were mapped to orthologous mouse genes using MGI mammalian orthology (http://www.informatics.jax.org/orthology.shtml). We calculated the p value for each gene set with respect to each signature cluster using a hypergeometric test. We used the number of Novershtern clusters significantly overrepresented (Bonferroni corrected $p < 0.001$) for one or more transcription factor targets as a measure to evaluate performance while comparing different methods.

Gene expression datasets

Nine gene expression signatures (d-erythroid, differentiated, d-lymphoid, d-myeloid, r-myelolymphoid, s-erythroid, s-mpp, s-myelolymphoid, and stem) were obtained from [15]. Differentially expressed genes in various leukemia datasets were downloaded from their respective publications. Gene lists were then interrogated against the ChIP-seq compendium using both GSCA and C-GSCA.

GSCA web tool

The GSCA output was produced using R, and the web user interface of the application was done using Perl/CGI/HTML. R commands are executed through the perl–cgi script to produce the image. The web tool can be accessed at the following URL: http://bioinformatics.cscr.cam.ac.uk/GSCA/GSCA.html.

Results

Definition of a candidate regulatory genome in mouse hematopoiesis

We recently reported a compendium of more than 50 TF ChIP-Seq experiments in mouse blood cells collected from publicly available datasets [8]. We have doubled the compendium by adding 60 new ChIP datasets from 18 recently published studies [16–33] and ENCODE murine unpublished datasets to obtain genome-wide binding