The regulated transcription of genes in time and space is a necessity for life, and underlies the diversity of cell types as well as the transition between them. In other words, different cell types are defined by the set of genes they transcribe, as well as by the level of expression of those genes. An intricate program of regulatory events controls this variation in gene transcription levels. Its focal points are where transcription of genes is initiated: the transcription start sites (TSS). Here, RNA polymerase II and associated proteins are recruited and the decision to start transcription is made. This process is influenced by more distal regulatory events: proteins called transcription factors can bind in the region around the TSS, called the promoter, and also to more distal activating regulatory elements called enhancers. A common model suggests that the enhancers, when active, loop into the region around TSSs of the genes they regulate (Figure 1). Although several models have been proposed to explain the mechanisms by which enhancers regulate transcription, it is generally perceived that enhancers act to increase the transcriptional output of target genes by supplying factors needed for gene transcription, thereby enabling transcription or fine-tuning the exact level of gene expression. This regulation is in particular important during development when the timely control of gene expression is vital for proper

AN ATLAS OF ENHANCER USAGE ACROSS THE HUMAN BODY

ROBIN ANDERSSON, AND ALBIN SANDELIN, THE BIOINFORMATICS CENTRE, DEPARTMENT OF BIOLOGY, UNIVERSITY OF COPENHAGEN, DENMARK. ROBIN@BINFU.DK, ALBIN@BINFU.DK

Understanding the specialization of cell types

The regulated transcription of genes in time and space is a necessity for life, and underlies the diversity of cell types as well as the transition between them. In other words, different cell types are defined by the set of genes that they transcribe, as well as by the level of expression of those genes. An intricate program of regulatory events controls this variation in gene transcription levels. Its focal points are where transcription of genes is initiated: the transcription start sites (TSS). Here, RNA polymerase II and associated proteins are recruited and the decision to start transcription is made. This process is influenced by more distal regulatory events: proteins called transcription factors can bind in the region around the TSS, called the promoter, and also to more distal activating regulatory elements called enhancers. A common model suggests that the enhancers, when active, loop into the region around TSSs of the genes they regulate (Figure 1). Although several models have been proposed to explain the mechanisms by which enhancers regulate transcription, it is generally perceived that enhancers act to increase the transcriptional output of target genes by supplying factors needed for gene transcription, thereby enabling transcription or fine-tuning the exact level of gene expression. This regulation is in particular important during development when the timely control of gene expression is vital for proper...
MEASURING CELLULAR DIVERSITY IN GENE USAGE

A direct way of measuring cell-type-variation in gene expression is to quantify the abundances of produced RNAs. Genome-wide approaches utilize techniques to sequence RNAs at massively parallel sequencers and map the sequenced reads back to a reference genome. Localization and quantification of mapped reads provide a direct estimate of RNA abundances and may explain cell type specialization. In one genome-wide technique, called Cap Analysis of Gene Expression (Kamori-Katayama et al. 2011) (CAGE, genome-scale 5’ RACE), the first 20-30 nucleotides of capped RNAs are sequenced. When mapped to a reference genome, closely positioned CAGE reads indicate the locations of TSSs and their frequencies directly quantify expression levels. Hence, using CAGE one may accurately identify the locations and usage levels of TSSs, and thereby promoters, genome-wide.

In an ambitious effort to profile promoter usage and what defines cellularity, the international Functional Annotation Of Mammals (FANTOM) consortium (http://fantom.gsc.riken.jp) has in the first phase of its fifth project (FANTOM5) applied CAGE on the majority of cell types and tissues in human and mouse (FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014). The result is a very useful resource for molecular biologists, as it comprises a broad atlas not only of expression but also of the genomic locations of TSSs. Moreover, CAGE is a “blind” technology not based on gene annotation. This means that it enables the identification of rare TSSs, which are typically only used in some cell types, representing for instance the 5’ ends of novel genes or alternative promoters to known protein-coding genes. One of the key findings of the consortium was that there are very few genes that are used uniformly across cells, so called “house-keeping genes”. By tracing the evolution of human promoter sequences across species, the FANTOM5 project further identified a clear difference in the age of promoter sequences utilized in specific human cell types, among which immune cells stand out showing the least evolutionary conservation with distant species.

While the variability in promoter usage explains cellular differences, it provides little information on how these differences are regulated. To better understand the variability in regulatory programs, one needs to focus also on the cell-type-specific usage of enhancers.

FINDING ENHANCERS BY HISTONE MARK PROFILING

Enhancers have been a hot topic during the last 30 years because of their importance in transcriptional regulation and during development. However, their detection was historically arduous, since they can be located very far away – sometimes millions of base pairs – from the gene promoters they regulate.

The merger between chromatin immunoprecipitation (ChIP) and high-throughput DNA sequencing techniques made it possible to interrogate where transcription factors and other DNA-binding proteins were bound to DNA. Simplified, the technique is based on fixing DNA-bound proteins to DNA, shearing of DNA into small pieces, and the specific selection of those DNA pieces bound to a protein of interest by the use of specific antibodies. The ChIP technique has made it possible to profile chromatin states, by targeting nucleosomal histone proteins with specific post-translational modifications. Through genome-wide profiling of chromatin states, it was evident that some combination of histone modifications (often called “histone marks”) were predictive of enhancers. In particular, enhancers were found enriched for flanking nucleosomes marked by H3K4me1 (mono-methylation of lysine 4 in histone 3) and H3K27 (acetylation of lysine 27) with no or low enrichment of H3K4me3 (tri-methylation).

The ENCODE project profiled several cancer cell lines for these and other histone marks, and were able to predict the locations of enhancers and other types of genomic elements (The ENCODE Project Consortium 2012). However, only a subset of predicted enhancers (~25%) could be validated successfully in vitro using luciferase reporter assays (Kheradpour et al. 2013). In addition to its low predictive performance, a limitation with this technique is that it requires a large number of cells thus making it challenging to profile many cell types, and especially human tissue.

FINDING ENHANCERS BY TRANSCRIPTION START SITE PROFILING

In a seminal paper, Kim et al (2010) showed that, upon stimulation of mouse cortical neurons, many activated enhancers initiated transcription locally. In
other words, they could also function as promoters, having their own TSSs and producing a subclass of long non-coding RNAs (lncRNAs) called enhancer RNAs (eRNAs). Individual enhancers had previously been observed to initiate transcription, but were generally considered outliers and this work provided the first evidence of extensive transcription from regulatory active enhancers. We reasoned that since enhancers have transcription start sites, we could be able to detect them with the FANTOM CAGE data. If this would work, we would in effect have an atlas of enhancers and their activities across the human body.

Indeed, we found that CAGE data showed a distinct pattern on enhancers compared to that of gene promoters: while transcription is heavily biased to the sense strand at mRNA promoters, enhancers initiate more balanced bidirectional transcription (Figure 2). This signature allowed us to establish a computational strategy to find other instances of this enhancer-typical pattern across the whole FANTOM data set, and the whole human genome (Andersson et al. 2014). Encouragingly, we observed that enhancers predicted by CAGE were about 3 times as likely to be shared between cells, histone marks alone, we observed many enhancers to be shared between cells, although the majority still had a more restricted usage than mRNA promoters. Interestingly, we also observed a small set of (~200) “ubiquitous” enhancers that were shared between nearly all human cell types and tissues considered.

The enhancer location and usage atlas further allowed us to study genetic variants in better detail than what has been previously possible. For many diseases, associated genetic variants are not located in exons suggesting that a large fraction may disrupt gene-distal regulatory sequences like enhancers. However, incomplete annotations of regulatory elements have made it hard to make real sense of these events. With the FANTOM enhancer atlas at hand, we observed that disease-associated single nucleotide polymorphisms (SNPs) are over-represented in enhancers, which in turn are expressed in pathologically relevant cell types or tissues. For instance Graves’ disease-associated SNPs are enriched in enhancers that are expressed predominantly in thyroid tissue, and similarly lymphocytes for chronic lymphocytic leukemia.

With the disease-perspective in mind, understanding the effect of a regulatory genetic variant requires knowing also their target genes. Since CAGE simultaneously measures enhancer and promoter activities, the data permits assessment of regulatory targets using enhancer-promoter co-expression linking. In other words, if an enhancer and a promoter have the same expression pattern across FANTOM CAGE libraries, we predict that they interact. Using expression correlation, we were able to predict novel and identify known regulatory targets of enhancers. One example of insight that such analysis can give using broad enhancer and promoter expression data was the fact that genes often are linked to many enhancers. We showed that these enhancers

AN ATLAS OF ENHANCER USAGE ACROSS THE HUMAN BODY

These observations paved the way for a paradigm shift in the effort to assess enhancer activities. Via systematic profiling of eRNAs across the broad panel of tissues and cell types surveyed in FANTOM, we were able to determine the activity of enhancers and scrutinize their cell-type-specificity. In total, we identified more than 43,000 transcribed enhancers and estimated the usage of each enhancer across the FANTOM expression atlas (Andersson et al. 2014). In contrast to estimates based on histone marks alone, we observed many enhancers to be shared between cells, although the majority still had a more restricted usage than mRNA promoters.

The enhancer location and usage atlas further allowed us to study genetic variants in better detail than what has been previously possible. For many diseases, associated genetic variants are not located in exons suggesting that a large fraction may disrupt gene-distal regulatory sequences like enhancers. However, incomplete annotations of regulatory elements have made it hard to make real sense of these events. With the FANTOM enhancer atlas at hand, we observed that disease-associated single nucleotide polymorphisms (SNPs) are over-represented in enhancers, which in turn are expressed in pathologically relevant cell types or tissues. For instance Graves’ disease-associated SNPs are enriched in enhancers that are expressed predominantly in thyroid tissue, and similarly lymphocytes for chronic lymphocytic leukemia.

With the disease-perspective in mind, understanding the effect of a regulatory genetic variant requires knowing also their target genes. Since CAGE simultaneously measures enhancer and promoter activities, the data permits assessment of regulatory targets using enhancer-promoter co-expression linking. In other words, if an enhancer and a promoter have the same expression pattern across FANTOM CAGE libraries, we predict that they interact. Using expression correlation, we were able to predict novel and identify known regulatory targets of enhancers. One example of insight that such analysis can give using broad enhancer and promoter expression data was the fact that genes often are linked to many enhancers. We showed that these enhancers
were often very similar in terms of their expression pattern, so that in terms of expression information, they were “redundant”. However, their combined usage seems to have a function in enabling higher levels of gene expression in an additive fashion. In other words, the more enhancers with similar expression profiles connected with a gene, the higher the expression level of that gene. This modeling highlights the importance of deciphering the properties and composition of regulatory architectures in order to understand the variability in gene expression and potential effect of genetic variation.

CAPTURING DYNAMICS IN REGULATORY ACTIVITIES

The first phase of FANTOM5 allowed us to measure differences in promoter and enhancer usage across the majority of human cell types and tissues. It should be noted that these maps are snapshots of single stages of human cells. Although this provides a means to study what determines the nature of specific cells, it provides no insight into the variability of regulatory events within a certain cell type or during cellular differentiation from one cell type to another.

In the second phase of FANTOM5, efforts were made to measure the dynamics of promoter and enhancer usage across mouse and human cellular differentiation and in response to stimuli (Arner et al. 2015). The addition of additional cell states allowed us to extend the FANTOM5 enhancer atlas with roughly 20,000 enhancers and to study their dynamics, alongside with that of gene promoters, in detail. Surprisingly, when focused on early response dynamics, we observed a highly generalized regulatory program across time courses. Regardless of cell type in focus or stimulus, and although few enhancers were shared between systems, enhancers generally responded very rapidly (at 15-45 minutes after stimulus), shortly followed by increased transcription of transcription factor genes. Non-transcription factor genes were, in general, activated later in the time course. Notably, regulatory targets of enhancers often showed a lagged response in transcriptional activation. These results suggest a general cascade of regulatory events, in which enhancers, although the agents differ between cellular systems, are the key players that first respond to changes in cellular environment.

OUTLOOK

The ability to locate enhancers from their produced RNA has important implications: because RNAs are much more abundant than DNA in cells, it is possible to locate enhancers in typical biopsies from living patients, with obvious importance for medical research. Secondly, the large array of molecular biology methods that work on RNAs can be applied. For instance, it is possible to use PCR-based methods to measure the usage of enhancers once located.

Why do enhancers produce eRNAs? Although functional roles for eRNAs have been suggested (Lam et al. 2014), these claims are based on anecdotal observations and a general functional attribution is debatable. While individual exceptions may exist, eRNAs as a group are rapidly degraded by the exosome, a enzymatic complex that degrades RNAs from their 3’ ends (Andersson et al. 2014). At the same time, eRNAs are typically not conserved across species. Thus, if eRNAs as a group are functional, their function must not rely on copy number or specific sequence. We find it equally likely that eRNAs are a consequence of enhancer action, which could be called biological noise.

Regardless of their potential function, it is clear that eRNAs are highly useful in understanding enhancer action and transcriptional regulation in general.

REFERENCES


