

Meeting report

## Sequencing the regulatory genome

Stein Aerts\* and Stefanie Butland†

Addresses: \*Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven B-3000, Belgium. †Centre for Molecular Medicine and Therapeutics, CFRI, University of British Columbia, Vancouver, V5Z 4H4, Canada.

Correspondence: Stefanie Butland. Email: butland@cmmmt.ubc.ca

Published: 19 June 2008

*Genome Biology* 2008, **9**:313 (doi:10.1186/gb-2008-9-6-313)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2008/9/6/313>

© 2008 BioMed Central Ltd

---

A report on the Cold Spring Harbor Laboratory meeting 'Systems Biology: Global Regulation of Gene Expression', Cold Spring Harbor, USA, 27-30 March 2008.

---

The line between the biological and computational research communities has disappeared in the field of gene regulation. The group of regulatory biology researchers represented at the recent meeting at Cold Spring Harbor on systems biology shares the same goal: develop and apply experimental and computational technologies to decipher the genomic regulatory code and the gene regulatory networks that are the driving forces of development and evolution. As in previous years, important gaps in solving the complex problem of gene regulation were bridged. This year featured the emerging massively parallel sequencing technologies, which are now being applied to every conceivable step in the gene regulation process, from gene annotation and alternative splicing, to transcription factor binding, and chromatin structure. Topics covered at the meeting ranged widely and in this report, we give our impressions of some highlights in two dominant themes: gene regulation in a nuclear context and transcription factor binding specificity.

### Genome geography in three dimensions

When transcription factors are reading the genomic regulatory code to determine the complement of active genes in a cell at a given time, they can be aided, guided, or obstructed by the chromatin they operate on. To catch chromatin in the regulatory act, laboratories are sequencing the sites associated with histone modifications that mark repressive, activating, and bivalent chromatin states, high-resolution DNase I hypersensitive sites (DHSs) that mark accessible chromatin, possible insulator sites, sites bound by transcription factors and RNA polymerase II, and *in vivo* cross-linked sites that represent long-range regulatory interactions in a locus. In an

increasing number of laboratories, the regulatory geography of the genome is now being assessed within the three-dimensional context of the nucleus.

Bas van Steensel (Netherlands Cancer Institute, Amsterdam, the Netherlands) provided an elegant picture of human gene regulation in three dimensions by identifying nuclear-lamina-associated domains (LADs) in interphase chromosomes. LADs range from 100 kb to 10 Mb and have sharp borders that define chromatin regions with distinctive characteristics: they tend to have fewer genes with lower expression levels compared with genes outside LADS, low RNA polymerase II occupancy, and enrichment of the repressive histone mark H3 trimethylated on lysine 27 (H3K27me3) at their borders. Thirty percent of LAD borders have at least one of three other marks: binding sites for the transcription factor CTCF (commonly held to act as insulators), a CpG island, or a promoter directing transcription away from the LAD.

Sites of chromatin accessibility across the human genome were precisely delineated by John Stamatoyannopoulos (University of Washington School of Medicine, St Louis, USA) in nine cell types by 'digital DNase I', an *in vivo* assay of DNase I hypersensitive sites identified by single-molecule sequencing. Stamatoyannopoulos has identified around 400,000 DHSs genome-wide, of which around 170,000 were highly regulated cell-type specific elements. A subset of these are organized into approximately 2,000 tissue 'regulons', each comprising a large cluster of lineage-specific elements spread out over tens or even hundreds of kilobases. Genes that were marked by these regulons in a given cell type showed striking over-representation of Gene Ontology terms for processes associated with the cell lineage in which they were observed. In addition to accessibility to DNase, active enhancers show specific histone modifications. Gary Hon (Ludwig Institute for Cancer Research, San Diego, USA) was able to distinguish cell-type specific enhancers

from promoters by their enrichment for H<sub>3</sub>K<sub>4</sub>me<sub>1</sub> over H<sub>3</sub>K<sub>4</sub>me<sub>3</sub> modifications. He proposed that these enhancers are what drive cell-type specific patterns of gene expression.

With such data it is critical to determine how different regulon elements interact with each other to elicit a response. Job Dekker (University of Massachusetts Medical School, Worcester, USA) described his 'chromosome conformation capture carbon copy' (5C) method to detect many-by-many chromatin interactions for a picture of spatial conformation of genomic regions. His analysis of a 1 Mb region around the human beta-globin locus showed that an alternative promoter 250 kb upstream physically interacts with the globin locus control region. Dekker pointed out that "simple models are insufficient" for gene regulation, as CTCF sites, usually considered as insulators, at the beta-globin locus actually facilitate long-range interactions between promoters and enhancers.

### Transcription factor binding specificity

We were reminded by Kevin Struhl (Harvard Medical School, Boston, USA) that the epigenetic states of chromatin cannot explain the specificity of gene expression, but are rather instructed by the sequence-specific transcription factors that translate the regulatory code and recruit chromatin-modifying activities. A cornerstone of our understanding of the regulatory language is the knowledge of a transcription factor's DNA-binding specificity. Significant progress has been achieved in deriving high-quality DNA-binding profiles through a variety of approaches with a large dose of collaboration, particularly with Martha Bulyk (Brigham & Women's Hospital and Harvard Medical School, Boston, USA) for protein-binding microarrays (PBMs). Scot Wolfe (University of Massachusetts Medical School, Worcester, USA) reported new binding profiles for 84 homeodomain transcription factors for *Drosophila melanogaster* through a bacterial one-hybrid system, while Gong-Hong Wei (University of Helsinki, Finland) described binding profiles for all 27 human and 26 mouse ETS family members using a microwell-based high-throughput assay and PBMs, and Christian Grove (University of Massachusetts Medical School, Worcester, USA) reported profiles for most of the basic helix-loop-helix (bHLH) dimers in *Caenorhabditis elegans* using a novel version of assay by PBMs. Timothy Hughes (University of Toronto, Canada) described profiles for 300 human and mouse transcription factors across 23 structural classes, including 168 profiles (of 175 total) for homeodomain transcription factors using PBMs.

All these profiles are highly conserved across species and can be ported between orthologous transcription factors when the DNA-contacting amino acids are conserved. This implies that a full compendium of transcription factor binding specificities across all animals can be accomplished in the near future, with about one third being finished and released

by these groups very soon. A question that remains is precisely how other contributors to specificity, such as transcription factor cooperativity, cell-type specific expression, variant or 'weak' recognition sites, and chromatin state together distinguish between correct target sites of related transcription factors that have virtually identical position weight matrices (PWMs).

While the relationship between transcription factors and their binding profiles is well conserved, independent data from various speakers showed yet again that the locations of *bona fide* regulatory elements are not always conserved in an alignment between orthologous regions. This plasticity of transcription factor recognition sites between functionally conserved regulatory regions is still posing a challenge for their computational prediction. Pouya Kheradpour (Massachusetts Institute of Technology, Cambridge, USA) presented a pragmatic solution by allowing for movement of a predicted site in an alignment, which for many motifs resulted in increased recovery of conserved sites (sensitivity) at a given specificity. Furthermore, many nonconserved sites are located in transposable elements that are generally not under selection and are usually masked before sequence analysis. For example, Guillaume Bourque (Genome Institute of Singapore, Singapore) found that 43% of non-conserved p53-binding sites are repeat-associated. Ting Wang (University of California, Santa Cruz, USA) identified a similar proportion of p53-binding sites in human endogenous retrovirus long terminal repeats, and Stamatoyannopoulos noted that around 10% of his DHSs map to transposable elements. Mobile elements thus provide an additional substrate for evolution of species-specific gene regulation.

### What's next in transcriptional regulation?

The key challenge will be to combine the two topics highlighted in this report, namely determination of the specific binding sites for multiple transcription factors and the genome-scale characterization of chromatin states, and to link these with spatial and temporal differences in gene expression. Advances in measuring cell-type specific gene expression were shown by Bob Waterston (University of Washington, St Louis, USA), who is using automated image-processing tools to analyze three-dimensional movies of fluorescent-marker tagged transcription factors in *C. elegans* embryos. Comparing massive numbers of images, they can make direct quantitative comparisons of expression patterns of different transcription factors "cell-by-cell, minute-by-minute". On the same topic, Philip Benfey (Duke University, Durham, USA) has leveraged a compendium of gene-expression data at cell-type specific resolution for an entire organ. His group performed microarray experiments on diverse cell lineages across the radial and longitudinal axes of the *Arabidopsis* root. A complementary set of experiments on six different cell types showed that specific cell types respond uniquely to high-salt or low-iron stress

conditions in terms of which genes are up- or down-regulated.

Robert Kingston (Harvard Medical School, Boston, USA) is developing technologies for locus-specific chromatin isolation to get the complete list of players that bind *in vivo* to a regulatory locus. He presented a convincing proof of principle by isolating 95% of known telomere interactors and identifying new biologically relevant ones. A more classical way of determining the input of multiple transcription factors to a specific locus is by genetic screens. Results of a high-throughput assay were presented by Pinay Kainth (University of Toronto, Canada), who tested the input contributions of all nonessential yeast transcription factors and their potential regulators on 27 cell-cycle-specific promoters using quantitative fluorescence measurements. Although genetic perturbations that alter a promoter's output are not limited to the transcription factors that physically bind to the promoter, such data can approximate direct interactions, especially when combined with PWM-based motif predictions.

Once the transcription-factor-specific regulatory sites, chromatin accessibility, and long-range interactions are determined for a given cell state, one must still determine the *cis*-regulatory logic and the rate of transcription initiation that it produces. This is still a difficult problem addressed by only few groups, including that of Jason Gertz (Washington University School of Medicine, St Louis, USA), who reported the use of libraries of synthetic regulatory regions to examine putative roles of combinations of *cis*-elements even before they have been discovered in real enhancers. This approach provides a possible solution to the sparse sampling of sets of *in vivo* validated regulatory regions that produce a similar output.

This high-quality meeting of regulatory biology researchers indicates that we are taking important steps toward the construction of a powerful toolkit to identify and model *in vivo* regulatory interactions and networks. The strong proofs of principle demonstrated at this meeting, together with increased access to massively parallel sequencing platforms, anticipate an era in which systems geneticists will collaborate to perform gene-regulation experiments in unprecedented detail and scale to characterize their pet 'regulome', and niche biologists will apply these technologies to address specific hypotheses about development, health and disease.