International Journal of Computational Bioscience, Vol. 1, No. 1, 2010

THE DNA-PROTEOME: RECENT ADVANCES TOWARDS ESTABLISHING THE PROTEIN-DNA INTERACTION SPACE

E. Grotewold* and H. Auer**

The regulation of gene expression is an essential aspect of the biology of all organisms. The nucleus of eukaryotic cells contains similar mass quantities of DNA and proteins. Many of these proteins, such as e.g., the histones, participate in organizing the DNA into higher-order structures. Others, such as e.g., the transcription factors, are involved in regulating the expression of genes by interacting with specific DNA sequences, the *cis*-regulatory elements, often proximal to the genes that they regulate, and/or with components of the transcriptional machinery. Recent technological advances, which include the ability to cross-link proteins to DNA and immunoprecipitate protein–DNA complexes (ChIP), coupled with the hybridization of tiling/promoter arrays (ChIP-chip) or with the sequencing of the ChIPed DNA (ChIP-Seq), are permitting to obtain a first genome-wide level picture of the interaction sites of particular proteins with DNA, in normal *in vivo* environments under physiological conditions. Combined with the exponential increase in DNA sequence information for a number of organisms, and the development of sensitive statistical tools to identify conserved DNA motifs across long evolutionary distances, a picture is emerging on how cells integrate transcriptional regulatory motifs with particular histone modifications, and how these motifs have evolved.

The DNA-Proteome Barcelona BioMed Conference brought together 150 scientists from Austria, Australia, Argentina, France, Germany, Italy, the Netherlands, UK, USA, Spain, Switzerland and Sweden for 3 days (April 20–22, 2009) to the beautiful Institut d'Estudis Catalans in Barcelona, supported by the Institute for Research in Biomedicine (IRB Barcelona) and Fundación BBVA, in the first gathering to specifically explore advances in the establishing the DNA–protein space in eukaryotic organisms, from humans to yeast and plants.

In the first session, advances in elucidating the *cis*-regulatory element landscape of eukaryotic genomes were presented. Gary Stormo (Washington University, St. Louis, MO, USA) discussed the DNA-recognition code for C_2H_2 zinc-finger and homeodomain transcription factors with the objective of being able to predict the DNA-sequence motifs that other members from these families of regulatory proteins would be able to recognize. With the goal to understanding the evolution of regulatory motifs that happened as part of the emergence of humans, Greg Wray (Duke University, Durham, NC, USA) interrogated positive selection of DNA elements during the evolutionary separation of humans from great apes. He showed that non-coding elements, like promoters, are far more prone to positive selection than coding regions. This selection affected preferentially promoters of neurogenesis and carbohydrate metabolism. He showed examples of specific genes that are expressed at different levels in humans and chimpanzees, potentially contributing to the evolution of brain size. Roderic Guigó (Center for Genomic Regulation, Barcelona, Spain) used Next Generation Sequencing technology to interrogate the complexity of the human transcriptome and to understand the influence of chromatin structure on splicing. He provided evidence that approximately 2% of transcripts result from inter-chromosmal splicing events. He also showed that nucleosome occupancy is denser in the centers of exons and in weak splicing sites, but lower in pseudogenes. Jan Karlseder (Salk Institute, La Jolla, CA, USA) described two different telomere regulation mechanisms in nematodes, one of them telomerase based (TEL) and the other recombination based (ALT) [1]. TEL generates G rich 3′ overhangs while ALT generates C rich 5′ overhangs. These two mechanisms are controlled by two different proteins, namely CeOB1 and CeOB2. This is in contrast to mammals, where both mechanisms are regulated by the same protein. Gerhard Mittler (Max Planck Institute of Immunology, Freiburg, Germany) described recent advances of a proteomic approach revealing sequence specific transcription factor binding to a motif of interest (SILAC DNA protein interaction screen [2]). This screen uses quantitative mass spectroscopy to identify sequence specific DNA binding proteins in the presence of unspecific binders. This methodology was applied to discover proteins binding to an evolutionary conserved promoter element present in 12 mammalian genomes. The transcription factors identified by this approach to specifically interact (directly or indirectly) with the promoter element were confirmed by ChIP and functional assays.

* Department of Plant Cellular & Molecular Biology and Plant Biotechnology Center, The Ohio State University, Columbus, OH 43210

** Institute for Research in Biomedicine, Barcelona, Spain

(paper no. 210-1008)

The second session, which focused on higher-order structures in the protein-DNA space, was kicked-of by Tom Gingeras (Cold Spring Harbor Lab, NY, USA). A leading figure in the Encyclopedia of DNA Elements (ENCODE) project [3], Tom Gingeras described the identification of a previously unknown pathway for the processing of a broad spectrum of RNAs, resulting in the generation of short RNAs (<200 bp) with modified 5′ ends which often target promoter sequences, and hence named PASRs (promoter associated short RNAs). As part of the discovery of novel RNAs, Dr. Gingeras also described the existence of transcripts derived from genomic sequences separated by 1 Mb or more, often even in different chromosomes. Rolf Ohlsson (Upsala University, Stockholm, Sweden) described the three-dimensional crosstalk between different chromosome territories. To study this higher-order crosstalk he utilized a technique called Circular Chromosome Conformation Capture (4C). Imprinting spreads to other chromosomal domains *via* physical interaction. These interactions are very dynamic and involve different domains in different tissues as well as within the same tissue at different developmental stages. Ola Söderberg (Upsala University, Stockholm, Sweden) described the Proximity Ligation Assay [PLA] [4] as a tool to study protein–protein interactions and protein–nucleic acid interactions *in situ*. Vivian Cheung (University of Pennsylvania, Philadelphia, PA, USA) described recent advances of understanding the genetic basis of individual differences of gene expression. Dr. Cheung used radiation response as a model to identify regulators of expression levels. Nearly all regulators act in trans but only a portion of them are transcription factors. For the other trans regulators, the underlying mechanisms of regulation remain to be explained.

Establishing the architecture of gene regulatory networks was the unifying theme in the third session. Mike Snyder (Yale University, New Haven, CT, USA) described efforts by his group to identify human transcription factors involved in the differentiation of mouse embryonic cells into neurons [5]. Sarah Bray (University of Cambridge, Cambridge, UK) described the identification of regulatory motifs associated with the Notch signalling pathway in *Drosophila*, and highlighted the presence of numerous incoherent feed-forward loops [6], likely involved in providing the temporal dimension to the response. Albert Jordan (CSIC, Barcelona, Spain) highlighted the roles of different histone H1 variants in chromatin compactation and gene expression. While H1.2 depletion causes decreased nucleosomal spacing and cell cycle arrest in G1-phase, H1.4 depletion causes cell death. Specific phenotypes due to depletion of individual H1 variants points to distinct roles of the linker histone variants. Duncan Odom (Cambridge Research Institute, Cambridge, UK) described recent advances in understanding of molecular mechanisms of transcription factor binding evolution. For this purpose, he used a mouse model containing the entire human chromosome 21 to show that binding of transcription factors is almost exclusively determined by the DNA sequence [7]. He also described the conservation of binding sites of the transcription factor CEBPα in the liver across five vertebrates. From approximately 30,000 binding sites per species, only 40 fall into the same position across all aligned genomes. Highlighting the increasing identification of regulatory networks in the plants, Erich Grotewold (The Ohio State University, Columbus, OH, USA) described ChIP-chip experiments aimed at understanding the mechanisms associated with the initial differentiation of *Arabidopsis* epidermal cells into leaf hairs [8].

A key aspect of establishing gene regulatory networks is to identify the genome-wide location of transcription factor binding sites, which was the topic of the fourth session. Combining the generation of the first plant interactome by investigating the pair-wise interactions of 12,000 *Arabidopsis* ($12{,}000 \times 12{,}000$ combinations) proteins with the identification by ChIP-Seq of direct targets for ethylene response regulators was the topic covered by Joe Ecker (Salk Institute, La Jolla, CA, USA). Stressing the power of gene-centered approaches [9] to identify gene regulatory networks, Bart Deplancke (École Polytechnique Fédérale de Lausenne, Lausanne, Switzerland) described the development of a collection of novel tools and resources to map transcription factor–protein and transcription factor–DNA interactions in the mouse and *Drosophila* model organisms. Thomas Graf (Center for Genomic Regulation, Barcelona, Spain) described the trans-differentiation of B-lymphocytes into macrophages by induction of a single transcription factor, namely CEBPα. Three hours after induction, expression of 700 genes is altered and 24 h after induction, 50% of cells are irreversibly committed to the macrophage lineage. CEBPα represses the PAX5 controlled B-cell program *via* induction of CEBPβ; CEBPα directly activates the macrophage specific expression program. Eileen Furlong (EMBL, Heidelberg, Germany) described the genome control network of transcription factor binding during *Drosophila* mesoderm development. Using a time-course of ChIP-chip data, she found extensive combinatorial occupation of promoters by transcription factors in a temporal manner. Analysis of expression data showed that transcription factor occupancy is sufficient to predict enhancer activity. Peggy Farnham (UC Davis, Davis, CA, USA) described her attempts to categorize transcription factors according to their binding patterns across the genome. Using ChIP-chip and ChIP-seq data, it is possible to divide transcription factors into groups that have one or multiple binding sites per gene. Another criterion is binding in a tissue-specific manner versus targeting a similar set of genes across many cell types. A third criterion is that for some transcription factors the conservation of consensus motifs is very high while others show a wide-spread pattern of sequence binding.

Key aspects to gene regulation and genome stability are chromatin structure and histone modifications, the topic of the last session. Dirk Schuebeler (Friedrich Miescher Institute, Basel, Switzerland) described the role of chromatin properties, like DNA methylation and histone modifications, on the timing of DNA replication. In contrast to yeast, where origins of replication are almost exclusively determined by DNA sequence, metazoans replication timing is controlled in a cell type specific manner and dependent on the local and chromosome-wide chromatin state. Robin Allshire (University of Edinburgh, Edinburgh,

UK) described the formation of synthetic heterochromatin in fission yeast by the targeting of the Clr4 methyltransferase [10] to euchromatic region, with significant implications for understanding centromere structure. Positioning of nucleosomes across the genome was the theme of Federica Battistini's (University of Sheffield, Sheffield, UK) presentation. A structural model was developed to predict sequence dependent local conformational changes associated with DNA bending together with the resulting strain energy. The model identifies successfully high affinity sequences that were experimentally selected. Andrew Andrews, from the Luger lab (Colorado State University, Ft. Collins, CO, USA) described the role of the histone chaperon yNap1 to prevent non-canonical histone–DNA interactions and to measure nucleosome assembly thermodynamics *in vitro*. Saadi Khochbin (Institut Albert Bonniot, Grenoble, France) described the post-meiotic male genome reorganization and compaction by stepwise replacement of histones by transition proteins and protamines, where one of the earliest steps of the genome reorganization is histone hyperacetylation. Manel Esteller (IDIBELL-ICREA, Barcelona, Spain) closed the conference by describing the influence of DNA methylation on phenotypic differences as seen in cloned animals and monozygotic twins. Healthy and diabetes- or lupus-affected twins show pronounced differences in their, NA methylation patterns and appropriate treatment of the disease restores a pattern very similar to the one of the healthy twin. DNA methylation is involved in regulation of coding RNAs as well as micro RNAs.

As a whole, the conference provided a first picture of the complexity of the protein–DNA interactions that occur inside eukaryotic cells, and highlighted the potential of emerging tools to continue to elucidate the nuclear processes associated with gene expression. The general consensus was that a similar conference in 2 years would be very valuable for the community.

References

[1] M. Raices, R.E. Verdun, S.A. Compton, C.I. Haggblom, J.D. Griffith, A. Dillin, & J. Karlseder, *C. elegans* telomeres contain G-strand and C-strand overhangs that are bound by distinct proteins, *Cell*, *132*, 2008, 745–757.

[2] G. Mittler, F. Butter, & M. Mann, A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements, *Genome Research*, *19*, 2009, 284–293.

[3] E. Birney, J.A. Stamatoyannopoulos, A. Dutta, R. Guigo, T.R. Gingeras, E.H. Margulies, Z. Weng, M. Snyder, E.T. Dermitzakis, R.E. Thurman, *et al.*, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, *447*, 2007, 799–816.

[4] S. Fredriksson, M. Gullberg, J. Jarvius, C. Olsson, K. Pietras, S.M. Gustafsdottir, A. Ostman and U. Landegren, Protein detection using proximity-dependent DNA ligation assays, *Nature Biotechnology*, *20*, 2002, 473–477.

[5] E. Theodorou, G. Dalembert, C. Heffelfinger, E. White, S. Weissman, L. Corcoran, & M. Snyder, A high throughput embryonic stem cell screen identifies Oct-2 as a bifunctional regulator of neuronal differentiation, *Genes and Development*, *23*, 2009, 575–588.

[6] S. Kaplan, A. Bren, E. Dekel, & U. Alon, The incoherent feed-forward loop can generate non-monotonic input functions for genes, *Molecular Systems Biology*, *4*, 2008, 203.

[7] M.D. Wilson, N.L.B.-Morais, D. Schmidt, C.M. Conboy, L. Vanes, V.L. Tybulewicz, E.M. Fisher, S. Tavare, & D.T. Odom, Species-specific transcription in mice carrying human chromosome 21, *Science*, *322*, 2008, 434–438.

[8] K. Morohashi & E. Grotewold, A systems approach reveals regulatory circuitry for *Arabidopsis* trichome initiation by the GL3 and GL1 selectors, *PLoS Genetics*, *5*, 2009, e1000396.

[9] B. Deplancke, A. Mukhopadhyay, W. Ao, A.M. Elewa, C.A. Grove, N.J. Martinez, R. Sequerra, L.D.-Stamm, J.S.R.-Hoyes, I.A. Hope, *et al.*, A gene-centered *C. elegans* protein–DNA interaction network, *Cell*, *125*, 2006, 1193–1205.

[10] R.C. Allshire & G.H. Karpen, Epigenetic regulation of centromeric chromatin: Old dogs, new tricks?, *Nature Reviews Genetics*, *9*, 2008, 923–937.

Biographies

Erich Grotewold obtained his Ph.D. degree in Chemistry from the University of Buenos Aires (1988). He began his research on maize genetics as a postdoc at Cold Spring Harbor Lab (1988), where he continued as an Assistant Investigator until 1998, when he joined Department of Plant Cellular and Molecular Biology at The Ohio State University, where he is currently a professor. His research focuses on the control of gene expression in plants and establishing the architecture of plant gene regulatory networks.

Herbert Auer received his M.Sc. degree in Biology from the University of Vienna, Austria (1998). He worked in various areas of molecular biology since 1982 until he switched to technology development in the area of genomics in 1998. Since then, he directed genomic core facilities in Europe and the United States. Currently he is the Director of the Functional Genomics Core at the Institute for Research in Biomedicine (IRB) in Barcelona, Spain.