

Cite this: *Analyst*, 2011, **136**, 3060

www.rsc.org/analyst

## CRITICAL REVIEW

# To understand the whole, you must know the parts: unraveling the roles of protein–DNA interactions in genome regulation†

Lloyd M. Smith,<sup>\*a</sup> Michael R. Shortreed<sup>a</sup> and Michael Olivier<sup>b</sup>

Received 14th January 2011, Accepted 28th April 2011

DOI: 10.1039/c1an15037e

The regulation of gene transcription is fundamental to the existence of complex multicellular organisms such as humans. This process dictates which genes are expressed in which tissues, and controls how various cell types grow, differentiate, and respond to their environments. Although the deciphering of the human genome sequence has given us the “source code” for life, we still know far too little about the mechanisms that control which sets of genes are active in which tissues, and how their expression is regulated. It is clear, however, that much of this system depends upon the sequence-specific interactions of regulatory proteins with particular genetic loci. To be able to unravel the details of these interactions on a genome-wide basis, it is necessary to know what proteins are bound to the DNA where in the genome, and to be able to monitor how those proteins change over time and in response to external stimuli. Developing a new technology to provide this information constitutes a “Grand Challenge” for Analytical Chemistry. In this brief article we outline the nature of this challenge, and propose one strategy to address it.

### Introduction

One of the most important biological functions of a cell is the regulation of gene transcription to translate the information encoded in the genome into biological function. Gene expression is primarily controlled by the availabilities and activities of specific transcription factors and other regulatory proteins and

<sup>a</sup>Department of Chemistry, University of Wisconsin, 1101 University Avenue, Madison, WI, 53706, USA. E-mail: smith@chem.wisc.edu; Fax: +1 608 265 6780; Tel: +1 608 262 9207

<sup>b</sup>Biotechnology and Bioengineering Center, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI, 53226, USA

† This article is part of a themed issue on Grand Challenges.



Lloyd Smith

Lloyd Smith is the W. L. Hubbell Professor of Chemistry at the University of Wisconsin in Madison, Director of the Genome Center of Wisconsin, and Co-Director of the Wisconsin Center of Excellence in Genomics Science (CEGS). His research focuses on the development of powerful new technologies to drive biological research, with particular interest in surface chemistries, mass spectrometry, and the emerging area of synthetic biology. His work has been recognized by the Eli

Lilly Analytical Chemistry Award, ABRF Award for development of automated DNA sequencing, American Chemical Society Award in Chemical Instrumentation, and the Pittsburgh Award for Analytical Chemistry.



Michael Shortreed

Michael Shortreed is a Senior Scientist in the Department of Chemistry at the University of Wisconsin-Madison. His research is focused on the development of array and mass spectrometry-based technologies for the elucidation of biomolecular interactions. He received his PhD in Analytical Chemistry from the University of Michigan for the development of fluorescent optical biosensors and light-harvesting supermolecules under the guidance of Professor Raoul Kopelman. His postdoctoral

training at Iowa State University with Professor Ed Yeung concerned the development of high-throughput methods for single-molecule imaging.

by the physical accessibility of specific genomic regions to the transcriptional machinery.<sup>1</sup> These DNA-binding proteins influence genetic expression by interacting with promoters, enhancers, silencers, insulators and locus control regions, both proximal and distal to a gene.<sup>2</sup> Despite the availability of the entire sequence of many genomes, our empirical knowledge of the DNA sequences that are targeted for binding by regulatory proteins is limited, and the prediction of these sites and sequences computationally from DNA sequence information continues to prove challenging.<sup>3</sup> This situation was underscored by the Encyclopedia of DNA Elements (ENCODE) Consortium,<sup>2</sup> an international effort to analyze and annotate the human genome and its DNA sequence. To quote from their findings, “Consensus sequences of transcription factor binding sites (typically 6 to 10 bases) have relatively little information content and are present numerous times in the genome, with the great majority of these not participating in transcriptional regulation. Does chromatin structure then determine whether such a sequence has a regulatory role? Are there complex inter-factor interactions that integrate the signals from multiple sites? How are signals from different distal regulatory elements coupled without affecting all neighboring genes?”

We briefly review below the roles that chromatin structure and accessibility, epigenetic modification to histones and DNA, and critical genomic regulatory elements play in controlling transcription. We then will make the case that the major missing component in gaining a complete understanding of gene transcription is knowledge of the identities and locations of the proteins that associate with and control expression of the genome. Developing a technology that will reveal this information constitutes a “Grand Challenge” for Analytical Chemistry.

## Chromatin changes and histone modifications

Modification of histone proteins affects the accessibility of genomic DNA in the nucleus to the protein machinery responsible for translation. Genomic DNA within the nucleus of a cell is

normally packaged into a smaller volume by forming a complex with histones and other structural proteins. This complex of protein and DNA, referred to as chromatin, provides multiple layers of structural organization. The base unit of chromatin, the nucleosome core particle, comprises 147 base pairs of DNA wrapped 1.6 times around a core histone octamer. This octamer consists of two molecules each of the four core histone proteins: H2A, H2B, H3, and H4. A short span of “linker” DNA connects adjacent core particles and is capped in mammalian cells by a molecule of linker histone H1. Further levels of compaction, and the addition of scaffold proteins, produce complex arrangements creating chromatin fibers of varying thickness and with different transcriptional activity. DNA in highly compacted fibers is usually transcriptionally inactive, and genes in those regions are not actively expressed. In contrast, DNA regions that are actively transcribed are usually opened to provide for access of transcription factors, regulatory proteins, and required enzymes to copy the DNA sequence into RNA molecules.<sup>4</sup>

Accessibility of chromatin and the packaging of DNA into condensed structures are mediated by changes in structural proteins associated with chromatin. Specifically, the *N*-terminal tails of histone proteins at the core of the nucleosome undergo extensive posttranslational modifications, including acetylation, methylation, phosphorylation, ubiquitination, ADP ribosylation, biotinylation, citrullination, and sumoylation.<sup>5</sup> As early as 1964 Allfrey *et al.* observed that increased histone acetylation levels correlate with active transcription.<sup>6</sup> Conversely, histone methylation has been linked to gene repression.<sup>7</sup> Antibodies targeting specific histone modifications have uncovered site or pattern-dependent correlations between modification and gene activity. Turner *et al.*, for example, examined *Drosophila* polytene chromosomes to revisit Allfrey’s findings.<sup>8</sup> With antibodies specific to each of the four acetylation sites of H4, they found strikingly divergent staining patterns: acetylated lysine 5 or lysine 8 (acK5 or acK8) stained throughout euchromatin (extended chromatin) regions, while acK12 was overrepresented in heterochromatin (condensed chromatin) and acK16 localized to the male  $\times$  chromosome. These and subsequent studies<sup>9</sup> led to the proposal of the histone code hypothesis by Strahl and Allis: “multiple histone modifications, acting in a combinatorial or sequential fashion on one or multiple histone tails, specify unique downstream functions”.<sup>10</sup> This hypothesis essentially proposes that the location of histone cores along the DNA is not static. Rather, different types of histone modification(s) modulate and alter the binding of DNA to nucleosome cores in chromatin structures, and thus affect transcriptional activation or repression.

## DNA methylation

In addition to the modification of histone proteins, chemical modification of the DNA molecule itself has also been shown to affect the level of gene expression. DNA methylation<sup>11,12</sup> involves the addition of a methyl group, commonly to the number 5 carbon of the cytosine pyrimidine ring. In adult somatic tissues, DNA methylation typically occurs in a CpG dinucleotide. Clusters of CpG dinucleotides (CpG islands) are often found in regulatory regions surrounding genes, and increased methylation of these regions is correlated with suppression of transcription.



Michael Olivier

*Michael Olivier is the Principal Investigator and Co-Director of the Wisconsin Center of Excellence in Genomics Science (CEGS). He is a Professor in the Department of Physiology at the Biotechnology and Bioengineering Center of the Medical College of Wisconsin. Dr Olivier’s research focuses on the analysis of genetic function in human disease, with a focus on common human disorders such as obesity and diabetes. As part of his research efforts, Dr Olivier has a long standing record of*

*technology development and implementation in proteomic mass spectrometry and in functional genomic analysis. He coordinates the CEGS research and technology development efforts, and its application to biological systems.*

DNA methylation may affect transcription of genes by physically hindering the binding of transcriptional proteins to the gene, or by preferentially binding to proteins known as methyl-CpG-binding domain proteins (MBDs). MBDs bind to histone deacetylases and other chromatin remodeling proteins that can modify histones, thereby forming compact, inactive chromatin, which has been associated with a variety of human disorders. The loss of methyl-CpG-binding protein 2 (MeCP2) has been implicated in Rett Syndrome<sup>13</sup> and methyl-CpG binding domain protein 2 (MBD2) mediates the transcriptional silencing of hypermethylated genes in cancer.<sup>14</sup>

## Identification of gene regulatory elements and DNA-binding proteins

Experimental methods to identify DNA-binding proteins and their specific position of binding are beginning to emerge<sup>15–19</sup> and will undoubtedly become more powerful as comprehensive databases and informatics tools for transcription factor interaction networks become available.<sup>1</sup> However, no methodologies exist to probe protein–DNA interactions on a genome-wide scale. Traditionally, individual DNA sequences are investigated for binding of nuclear proteins using electrophoretic mobility shift assays (EMSA).<sup>20</sup> In this approach, protein binding is assessed by examining the mobility of the DNA fragment on the gel, which is retarded if the DNA is bound by proteins. While this method is often used to test whether a particular DNA sequence has the ability to bind nuclear proteins, it does not allow routine identification of the bound proteins, and is only suitable for the analysis of individual DNA sequences. Similar to EMSA, DNaseI footprinting<sup>21</sup> has been used to analyze protein–DNA interactions *in vitro* whereby proteins of interest (*e.g.* nuclear extracts) are allowed to bind to a DNA fragment. The fragment is then cut with DNaseI and/or other agents, generating a series of smaller fragments of various sizes. The part of the DNA bound by a protein cannot be cut, resulting in a gap in the ladder of small fragments produced. This approach can be applied genome-wide by using DNA tiling arrays.<sup>22</sup> It is important to note, however, these methods do not provide information on the nature of the protein–DNA interaction that is present *in vivo*, which is critical to understanding the regulatory mechanism and dynamics.

Alternatively, numerous methods have been developed that allow the investigation of individual proteins and their interaction with DNA. Chromatin immunoprecipitation (ChIP) is probably the most widely accepted method for studying protein–DNA interactions.<sup>23,24</sup> Cross-linking by formaldehyde, followed by sonication to generate DNA fragments of a few hundred bp with physically attached proteins, is common for ChIP analysis of non-histone proteins. Other crosslinking agents have also been employed,<sup>25</sup> although this is less widespread. Antibodies specific to a protein of interest, including specific PTMs, are used to immunoprecipitate the protein–DNA complex. The DNA fragments are then dissociated from the protein and analyzed by PCR, real-time PCR, DNA microarray (ChIP-Chip), or sequencing (ChIP-seq).<sup>26</sup> The resolution and coverage of ChIP-Chip ultimately depends on the composition of the DNA chip. Efficient sequencing of short DNA fragments following ChIP (ChIP-seq),<sup>27</sup> on the other hand, may reveal unexpected protein

binding sites within DNA sequences that have not been pre-selected.

Both DNaseI and ChIP-Chip approaches have been used extensively by the ENCODE Consortium to reveal transcription factor binding sites using antibodies against a growing panel of sequence-specific transcription factors, components of the general transcription machinery, and modified histone proteins. In addition, the Consortium tested more than 600 potential promoter fragments for transcriptional activity by transient-transfection reporter assays in different human cell lines. This functional test is the primary validation of a biological role of the identified DNA sequences. However, such validation is elaborate and difficult, and negative results do not prove that a particular sequence is without biological relevance in transcriptional regulation (it may simply not play a significant role in the cell system tested).

New methods for mass-spectrometric identification of proteins binding to specific genomic loci are also beginning to emerge.<sup>17–19,28–31</sup> Early attempts at accomplishing this involved exposure of synthetic dsDNA as bait to trap specific DNA-binding proteins from nuclear extract.<sup>28–30</sup> The technique of SILAC (Stable Isotope Labeling by Amino acids in Cell culture) has been used to improve the confidence of such methods.<sup>17</sup> These *ex vivo* approaches have an advantage in that large amounts of DNA and extract can be used to isolate sufficient material for MS identification. In contrast, *in vivo* approaches are considerably more challenging because the DNA sequence of interest may be present at a level of as few as one copy per cell. Butala *et al.*<sup>18</sup> were able to achieve successful identification of proteins from protein–DNA complexes formed *in vivo* in bacteria by increasing the abundance of the DNA through clever use of a low copy number plasmid containing the sequence of interest and LacI to facilitate extraction. Déjardin and Kingston<sup>19</sup> used locked nucleic acid (LNA) probes to isolate genomic DNA with its associated proteins. There, they captured telomeric sequences, which are highly repetitive regions at the end of chromosomes, to obtain sufficient material for protein identification. It remains to be seen if any of these methods can be multiplexed for parallel analysis of multiple gene sequences. Furthermore, none have yet demonstrated sensitivity for identification of *in vivo* bound DNA-binding proteins when the sequence of interest is present at only a single copy per cell.

## A “Grand Challenge” for Analytical Chemistry

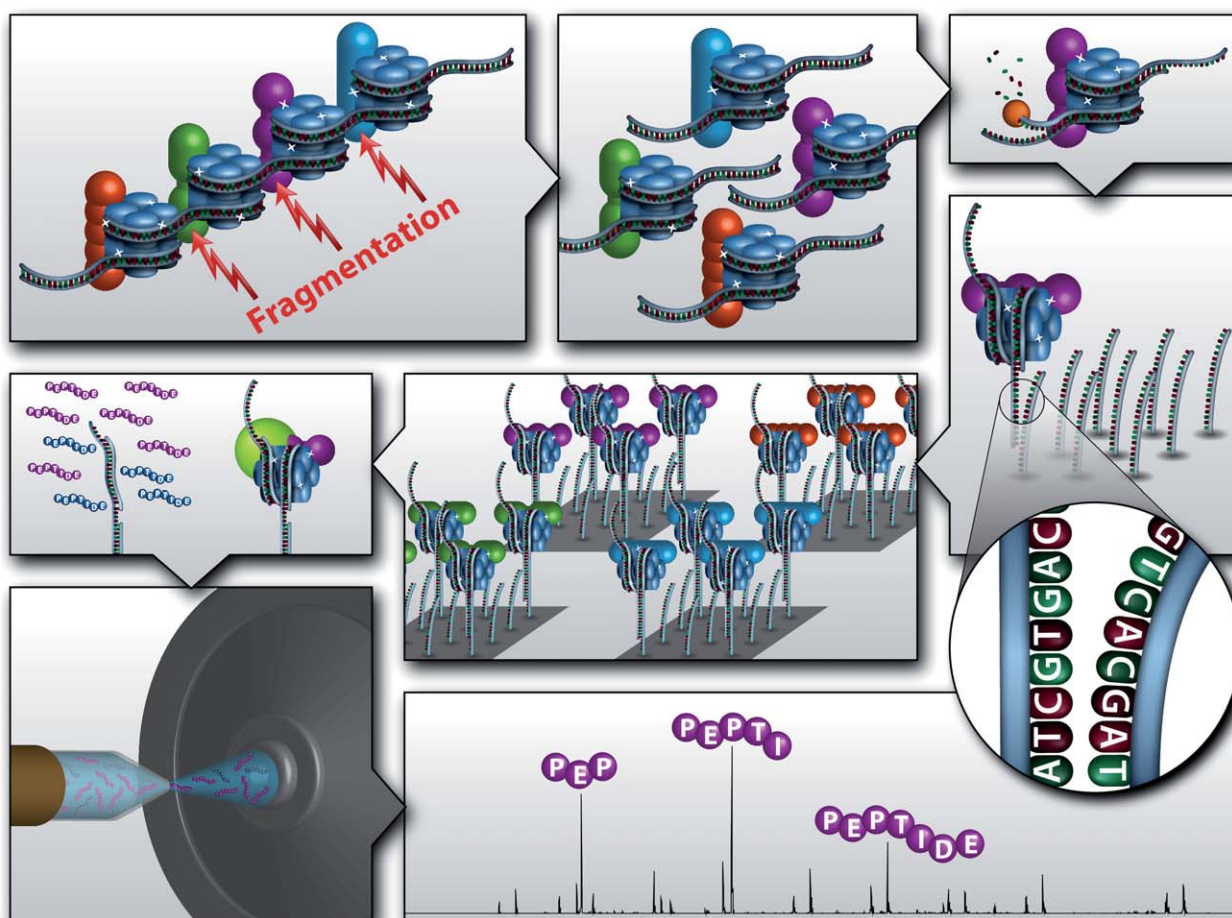
Based on this brief overview, it is clear that a complete analysis of transcriptional regulation and the identification of mechanisms underlying changes in gene transcription observed in physiological systems or disease will require a comprehensive analysis of global gene expression, DNA methylation, protein–DNA interactions and histone modifications, including the resulting changes in nucleosome positioning and DNA binding. While technologies already exist for the genome-wide analysis of gene transcription and DNA methylation, there is a desperate need for new technologies that enable the *comprehensive parallel analysis of all protein–DNA interactions (including histones) without prior knowledge or assumptions.*

How might this be achieved? One powerful strategy that we are pursuing as part of the Wisconsin Center of Excellence in

Genomics Science is called *GENECAPP*, for *Global ExoNuclease-based Enrichment of Chromatin-Associated Proteins for Proteomics*, and is illustrated in Fig. 1. In this approach, a specific DNA fragment is captured in a sequence-specific manner, allowing the isolation and subsequent characterization of all proteins bound to that region. As for chromatin immunoprecipitation, formaldehyde may be used to crosslink proteins and DNA *in vivo*, locking into place the protein–DNA interactions which are present at that time. The chromatin is then fragmented, either by a physical means such as sonication, or by restriction enzyme digestion. An exonuclease then removes one of the two strands of the duplexes protruding from the complex, leaving behind a free single-stranded region suitable for DNA hybridization. Incubation of this material with a solid support modified with complementary single-stranded DNA capture probes results in specific binding of the chromatin fragments of interest along with associated proteins. Subsequent characterization of these bound proteins by standard proteomic mass spectrometry techniques provides a comprehensive identification of all proteins that are bound to the targeted DNA region, and potentially the characterization of posttranslational protein

modifications. Parallel analysis of many regions across the genome can be achieved by using the multiplexing capabilities of either array-based or bead-based platforms with multiple capture oligonucleotide probes that are complementary to targeted DNA regions of interest.

Although this would appear to be a conceptually straightforward task, it is in fact an extremely challenging proposition (and hence worthy of the term “Grand Challenge”!). The biggest obstacle to the successful implementation of this approach involves detection sensitivity. In the ChIP approach, the captured nucleic acid is amplified by PCR, to convert trace amounts of captured DNA into the much larger quantities needed for subsequent identification and analysis. However, since no such capability is available to amplify trace amounts of proteins, one must make do with what one is able to capture. A specific discovery proteomics experiment, for example, may require as much as a picomole of a protein of interest to be captured and available for mass spectrometry analysis. If the protein of interest is present at a very low abundance (*e.g.* a single copy of the protein binds to a specific target sequence), as is often the case for “master regulators” of gene expression such as



**Fig. 1** The *GENECAPP* (*Global ExoNuclease-based Enrichment of Chromatin-Associated Proteins for Proteomics*) strategy. The parallel identification of proteins bound at specific genomic loci begins with fragmentation of cross-linked chromatin from fixed cells or tissues. Single-stranded regions of DNA are created on each fragment using an exonuclease to digest one of the two DNA strands. Individual fragments are captured on DNA arrays by hybridization to the free single-stranded portions of DNA on each fragment prior to proteolytic digestion and identification of the proteins using mass spectrometry.

transcription factors, the approach described above, at best, would allow the isolation of a single protein copy per cell. Obtaining a picomole of protein, accordingly, would require at least a picomole of cells, that is  $6 \times 10^{11}$  cells. In tissue culture of human cells,  $10^6$  cells per ml of culture media is a reasonable standard cell density, which leads one to the somewhat less reasonable projection of  $10^5$  ml or  $10^2$  litres to capture and isolate the necessary amount of protein for the mass spectral analysis. These 100 litres of cell culture medium would need to end up eventually in a volume of a few microlitres for the mass spectral analysis, a concentration factor of well over a million.

So, is this at all plausible? Sure, but it clearly is not easy. Continuous advances in instrument sensitivity for mass spectrometers suggest we may need well less than a picomole of a protein of interest for mass spectral analysis. Furthermore, proteins of interest will be present at much higher levels than one copy per cell (e.g. histone proteins on a DNA fragment with multiple nucleosomes, or transcription factors that bind as dimers or multimers or contain multiple binding sites for cooperative binding). With such revised assumptions, the total required culture volume for cells may be significantly reduced. For individual studies of high importance, it is certainly plausible that an investigator would be willing and able to produce the requisite tens to hundreds of litres of cell culture. However, for the approach described above to have widespread impact, it would be most advantageous to develop very efficient strategies for isolating and concentrating molecules of interest away from the complex cellular background, and to push the state of the art in mass spectrometric analysis as far as possible, thereby reducing the requirements for such voluminous and expensive cell culture work at the front end. Much work will also need to be done to implement the multiplexed technology necessary to probe many sites across the genome in parallel, and the technology will have to be rapid, efficient, and inexpensive, in order for it to be feasible to study the critically important dynamic and spatial variations occurring in cells and tissues during processes such as development, differentiation, and cellular communication and signaling.

The challenge does not end there. Other noteworthy issues and questions to be addressed include:

- Will the kinetics of the formaldehyde cross-linking be fast enough to capture transient protein–protein interactions, which, although weak, may nonetheless play critical roles in gene regulation?
- Will there be enough DNA present and accessible in the fragmented chromatin for robust DNA hybridization to solid supports?
- Will it be possible to identify posttranslational modifications of the bound proteins and monitor how they vary with time?
- Will it be possible to obtain not just qualitative, but also quantitative information on all forms of the proteins?
- Will it be possible to integrate the results that are obtained with information from other studies (e.g. genome sequence variation; gene expression analysis) to develop an integrated view of the system-wide regulation of gene expression?

These and many other issues will have to be tackled as the development of this novel and important new technology proceeds. The reward for this effort will be critical new insights into the workings of the genome, leading eventually to the much

fuller understanding of normal and disease biology that is the ultimate goal of biological research worldwide.

## Conclusions

The successful sequencing of the human and many other genomes has ushered in a new age in Biology—the “Genome Age”. Armed with this “Blueprint of Life”, we must now learn the mechanisms that control which sets of genes are active in which tissues, and how their expression is regulated. New technologies are needed to provide information on how organisms control and express their “source code”. In this brief article we have described the need to know what proteins are bound to the DNA where in the genome, and to be able to monitor how those proteins change in time and in response to external stimuli. We have outlined one possible strategy for obtaining this information. Addressing this problem clearly constitutes a “Grand Challenge” in Analytical Chemistry.

## Acknowledgements

This work was funded by the Wisconsin Center for Excellence in Genomics Science through NIH/NHGRI grant 1P50HG004952. We gratefully acknowledge A. J. Bureta for figure preparation.

## Notes and references

- 1 P. J. Farnham, *Nat. Rev. Genet.*, 2009, **10**, 605–616.
- 2 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, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhani, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Heri, S. C. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetric, M. Weaver, S. Wilcox, M. Yu, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sckinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermuller, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W. K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C. L. Wei, Y. Ruan, K. Struhl, M. Gerstein, S. E. Antonarakis, Y. Fu, E. D. Green, U. Karaoz, A. Siepel, J. Taylor, L. A. Liefer, K. A. Wetterstrand, P. J. Good, E. A. Feingold, M. S. Guyer, G. M. Cooper, G. Asimenos, C. N. Dewey, M. Hou, S. Nikolaev, J. I. Montoya-Burgos, A. Loytynoja, S. Whelan, F. Pardi, T. Massingham, H. Huang, N. R. Zhang, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, A. Sidow, N. D. Trinklein, Z. D. Zhang, L. Barrera, R. Stuart, D. C. King,

- A. Ameur, S. Enroth, M. C. Bieda, J. Kim, A. A. Bhinge, N. Jiang, J. Liu, F. Yao, V. B. Vega, C. W. Lee, P. Ng, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, J. C. Fowler, P. Couttet, A. W. Bruce, O. M. Dovey, P. D. Ellis, C. F. Langford, D. A. Nix, G. Euskirchen, S. Hartman, A. E. Urban, P. Kraus, S. Van Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, M. Xu, J. N. Haidar, Y. Yu, V. R. Iyer, R. D. Green, C. Wadelius, P. J. Farnham, B. Ren, R. A. Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, L. Armengol, C. P. Bird, P. I. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, A. Woodroffe, E. Davydov, A. Dimas, E. Eyraes, I. B. Hallgrimsdottir, J. Huppert, M. C. Zody, G. R. Abecasis, X. Estivill, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu and P. J. de Jong, *Nature*, 2007, **447**, 799–816.
- 3 V. Boeva, D. Surdez, N. Guillon, F. Tirode, A. P. Fejes, O. Delattre and E. Barillot, *Nucleic Acids Res.*, 2010, **38**, e126.
- 4 M. Chioda and P. B. Becker, *Heredity*, 2010, **105**, 71–79.
- 5 B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Molecular Biology of the Cell*, Garland Science, New York, 2002.
- 6 V. G. Allfrey, R. Faulkner and A. E. Mirsky, *Proc. Natl. Acad. Sci. U. S. A.*, 1964, **51**, 786–794.
- 7 R. Desrosiers and R. M. Tanguay, *Biochem. Biophys. Res. Commun.*, 1985, **133**, 823–829.
- 8 B. M. Turner, A. J. Birley and J. Lavender, *Cell*, 1992, **69**, 375–384.
- 9 J. E. Brownell and C. D. Allis, *Proc. Natl. Acad. Sci. U. S. A.*, 1995, **92**, 6364–6368.
- 10 B. D. Strahl and C. D. Allis, *Nature*, 2000, **403**, 41–45.
- 11 R. J. Klose and A. P. Bird, *Trends Biochem. Sci.*, 2006, **31**, 89–97.
- 12 H. D. Morgan, F. Santos, K. Green, W. Dean and W. Reik, *Hum. Mol. Genet.*, 2005, **14 Spec No 1**, R47–R58.
- 13 R. E. Amir, I. B. Van den Veyver, M. Wan, C. Q. Tran, U. Francke and H. Y. Zoghbi, *Nat. Genet.*, 1999, **23**, 185–188.
- 14 J. Berger and A. Bird, *Biochem. Soc. Trans.*, 2005, **33**, 1537–1540.
- 15 J. P. Lambert, J. Fillingham, M. Siahbazi, J. Greenblatt, K. Baetz and D. Figeys, *Mol. Syst. Biol.*, 2010, **6**, 448.
- 16 D. Jiang, H. W. Jarrett and W. E. Haskins, *J. Chromatogr., A*, 2009, **1216**, 6881–6889.
- 17 G. Mittler, F. Butter and M. Mann, *Genome Res.*, 2009, **19**, 284–293.
- 18 M. Butala, S. J. Busby and D. J. Lee, *Nucleic Acids Res.*, 2009, **37**, e37.
- 19 J. Déjardin and R. E. Kingston, *Cell*, 2009, **136**, 175–186.
- 20 P. Perez-Romero and M. J. Imperiale, *Methods Mol. Med.*, 2007, **131**, 123–139.
- 21 A. P. Boyle, L. Song, B. K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford and T. S. Furey, *Genome Res.*, 2010.
- 22 B. Shi, X. Guo, T. Wu, S. Sheng, J. Wang, G. Skogerbo, X. Zhu and R. Chen, *BMC Genomics*, 2009, **10**, 92.
- 23 L. Elnitski, V. X. Jin, P. J. Farnham and S. J. Jones, *Genome Res.*, 2006, **16**, 1455–1464.
- 24 A. J. Walhout, *Genome Res.*, 2006, **16**, 1445–1454.
- 25 G. T. Hermanson, *Bioconjugate Techniques*, Academic Press, San Diego, 1996.
- 26 P. Collas, *Mol. Biotechnol.*, 2010, **45**, 87–100.
- 27 G. Robertson, M. Hirst, M. Bainbridge, M. Bilenyk, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder and S. Jones, *Nat. Methods*, 2007, **4**, 651–657.
- 28 J. A. Stead, J. N. Keen and K. J. McDowall, *Mol. Cell. Proteomics*, 2006, **5**, 1697–1702.
- 29 E. Nordhoff, A. M. Krogsdam, H. F. Jorgensen, B. H. Kallipolitis, B. F. Clark, P. Roepstorff and K. Kristiansen, *Nat. Biotechnol.*, 1999, **17**, 884–888.
- 30 T. J. Griffin and R. Aebersold, *J. Biol. Chem.*, 2001, **276**, 45497–45500.
- 31 C. L. Himeda, J. A. Ranish, J. C. Angello, P. Maire, R. Aebersold and S. D. Hauschka, *Mol. Cell. Biol.*, 2004, **24**, 2132–2143.