

# Applications of high-throughput sequencing to chromatin structure and function in mammals

Ian Dunham

Address: EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Email: [dunham@ebi.ac.uk](mailto:dunham@ebi.ac.uk)

*Biology Reports* 2009, 1:32 (doi:10.3410/B1-32)

The electronic version of this article is the complete one and can be found at: <http://F1000.com/Reports/Biology/content/1/32>

## Abstract

High-throughput DNA sequencing approaches have enabled direct interrogation of chromatin samples from mammalian cells. We are beginning to develop a genome-wide description of nuclear function during development, but further data collection, refinement, and integration are needed.

## Introduction and context

Until recently, the primary motivation for DNA sequencing was a structural one. We wanted to sequence a specific molecule, perhaps a genome, to 'read' the information encoded in the DNA sequence. Determination of new genome sequences is still a major activity, and furthermore, sequencing of additional copies of the same genome to identify variation has moved on apace. However, the availability of extensive genome sequences of the major model organisms, including human [1], and the development of second-generation high-throughput (HTP) methods and instrumentation for genome sequencing [2,3] have opened up a new area of application: use of DNA sequencing as an analytic method. In this analytic mode, sequencing stands at the evolutionary pinnacle of a line of methods stretching back through DNA microarrays via Southern and Northern blotting to solution hybridisation. For any experimental sample containing nucleic acids, we want to determine the composition, preferably with accurate quantitation. Whereas prior methods depended upon nucleic acid hybridisation producing an analog signal one-step removed from the sequence of the DNA molecules, sequencing offers precise determination of the composition of a nucleic acid mixture and is not restricted to the sequences available on a microarray. Moreover, for complex nucleic acid samples, HTP sequencing generates gigabases of genome-wide sequence from each instrument run, providing

quantitative digital information on an acceptable time-scale at a cost that facilitates coverage for even large genomes.

Over the past 2 years, a number of reports have described the application of this new analytic approach to a variety of chromatin sample mixtures derived from mammalian cells. These data are beginning to give us a description of the complex regulatory processes that control the function of the genome in the cell nucleus.

## Major recent advances

Johnson *et al.* [4] described the application of HTP sequencing to chromatin immunoprecipitated with antibody to human neuron-restrictive silencer factor (NRSF) in the Jurkat T-cell line [chromatin-immunoprecipitation sequencing (ChIP-seq)]. They identified 1,946 NRSF-binding sites associated with repression of gene transcription close to 1,020 genes. These included previously unidentified gene targets of NRSF repression. In addition, a significant number of NRSF sites containing a relaxed form of the previously identified canonical NRSF-binding motif were identified. Concurrently, Robertson *et al.* [5] used ChIP-seq to identify ~41,000 and ~11,000 potential STAT1 targets in human HeLa S3 cells with or without  $\gamma$ -interferon stimulation respectively, reflecting activation of STAT1 upon cytokine stimulation. Sensitivity relative to known STAT1 targets was determined as 71%. Notably, Robertson *et al.*

deployed almost an order of magnitude more sequence reads than Johnson *et al.* In both studies, approximately one-half to two-thirds of the total sequences generated could be mapped uniquely to the genome. The success of both studies indicated that ChIP-seq using antibodies for transcription factors (TFs) could be deployed to create genome-wide maps of TF-binding sites. Future mapping of many TFs on samples from multiple tissues will result in a rich map of regulatory sites across the genome.

An alternative approach to mapping regulatory sites, including promoters, enhancers, silencers, and insulators, is to identify sites of DNase 1 hypersensitivity. HTP sequencing of libraries from DNase 1-digested nuclei of primary human CD4<sup>+</sup> T-cells was used by Boyle *et al.* [6], together with microarray data, to define ~95,000 sites on a genome-wide DNase 1 hypersensitivity map. Many of the strongest DNase 1 sites are found at transcription start sites (TSSs), but this accounts for only 16% of sites, whereas only 15.5% of sites lie outside of genes. Virtually all of the TSSs of highly expressed genes are marked by DNase 1 hypersensitive sites (HSs). Strikingly, Boyle *et al.* found that background DNase 1 sensitivity away from HSs exhibited an oscillating pattern occurring over a single nucleosome length with a frequency of one DNA double-helical turn. Furthermore, well-positioned nucleosomes could be detected close to some HSs. DNase 1 HS maps will serve as a valuable backbone on which to overlay TF-binding sites identified by ChIP-seq.

ChIP-seq has also been applied to analysis of the distributions of modified histones across the genome. Barski *et al.* [7] analysed enrichment of 20 different histone methylations as well as RNA polymerase II, CCCTC-binding factor (CTCF), and the histone variant H2AZ in CD4<sup>+</sup> T-cells. In addition to the previously described association of histone H3K4me1, H3K4me2, H3K4me3, and H3K36me3 association with gene activation, histone H3K27, H3K9, H4K20, H3K79, and H2BK5 monomethylations were all linked to active genes. In contrast, histone H3K27 and H3K9 trimethylations were associated with repressed genes and heterochromatin formation. CTCF was found to mark the boundaries of active and repressed histone modification domains [8]. Since the Barski *et al.* ChIP-seq protocol used micrococcal nuclease digestion to prepare mononucleosomes, the sequence data could be analysed on a strand-specific basis to identify nucleosome positions in promoters with extremely fine resolution [9]. Further analysis of additional histone methylations and acetylations in CD4<sup>+</sup> T-cells indicated that many of the modifications are strongly correlated [10]. For instance, a module including 17 histone modifications appears to be common at promoters. Mikkelsen *et al.* [11] also

generated genome-wide maps of chromatin modifications in embryonic stem and lineage-committed cells and observed that actively transcribed genes are marked by H3K4me3 at their promoters and H3K36me3 along their transcribed length. H3K27me3 marked either regions that are transcriptionally repressed or, when H3K27me3 occurred together with H3K4me3 at a promoter, genes poised for expression along future developmental paths. Subsequently, the H3K4me3/H3K36me3 signal at actively transcribed genes has been used to identify novel multi-exonic non-coding RNA genes [12]. A further advantage of the sequencing approach was revealed by Mikkelsen *et al.*, who were able to identify allele-specific histone modification patterns using single-nucleotide polymorphisms.

### Future directions

The application of HTP sequencing in ChIP-seq and DNase 1-seq is now extending to other TFs and other cells and tissues. The next step will be to integrate the different datasets with each other and with other approaches using HTP sequencing including DNA methylation analysis [13] and sequencing of cellular transcriptomes [14,15] to provide a comprehensive view of the regulation of transcription at a chromatin-wide level. Further on, it may be possible, along with HTP data on the three-dimensional interactions within chromatin [16], to model the regulatory structure of the nucleus. Comparison of these comprehensive chromatin-state maps between tissues and along developmental pathways should facilitate our understanding of regulation in development and disease.

The application of HTP sequencing to these complex samples is not without its caveats. Substantial algorithmic development to map the sequence reads efficiently and to score the sites of enrichment compared with an untreated control (non-immunoprecipitated chromatin or 'input' for ChIP-seq) is still ongoing. Indeed, even for the most comprehensively sequenced genome, not all sequences are represented in the reference, including unknown amounts of certain repetitive sequences, therefore careful filtering of the data is required. Improvements in the sequencing technologies, including higher throughput and longer sequence reads, will help to address these issues as well as to increase productivity and cost efficiency. However, sample input for ChIP-seq is also limited by the availability of suitable antibodies. In addition, future technologies that sequence single molecules without amplification may allow certain applications to work with very few cells, thereby overcoming the limitation restricting current analyses to the averages of broad populations of cells. In any case, it is

not hard to imagine a time when an HTP sequencer will be located in every functional genomics laboratory.

### Abbreviations

ChIP-seq, chromatin immunoprecipitation sequencing; CTCF, CCCTC-binding factor; HS, hypersensitive site; HTP, high-throughput; NRSF, neuron-restrictive silencer factor; TF, transcription factor; TSS, transcription start site.

### Competing interests

The author declares that he has no competing interests.

### Acknowledgements

The author thanks John Collins and Stephan Graf for their comments on the draft of the manuscript.

### References

- International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431**:931-45.  
F1000 Factor 9.0 *Exceptional*  
Evaluated by Joachim Messing 27 Oct 2004
- Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Masinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, et al.: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456**:53-9.
- Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452**:872-6.
- Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of *in vivo* protein-DNA interactions.** *Science* 2007, **316**:1497-502.  
F1000 Factor 8.6 *Exceptional*  
Evaluated by John Jaenike 21 Jun 2007, Deyou Zheng 29 Jun 2007, Ulf Pettersson 17 Jul 2007, Gabriele Varani 14 Aug 2007, Magdalena Zernicka-Goetz 15 Jan 2008
- Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.** *Nat Methods* 2007, **4**:651-7.
- Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE: **High-resolution mapping and characterization of open chromatin across the genome.** *Cell* 2008, **132**:311-22.
- Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129**:823-37.  
F1000 Factor 7.0 *Must Read*  
Evaluated by Steven Henikoff 22 May 2007, Xing Wang Deng 05 Jun 2007, Michael Meisterernst 19 Jun 2007, Deyou Zheng 29 Jun 2007, Magdalena Zernicka-Goetz 15 Jan 2008
- Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K: **Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains.** *Genome Res* 2009, **19**:24-32.  
F1000 Factor 6.0 *Must Read*  
Evaluated by I King Jordan 10 Feb 2009
- Schmid CD, Bucher P: **ChIP-Seq data reveal nucleosome architecture of human promoters.** *Cell* 2007, **131**:831-2.
- Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K: **Combinatorial patterns of histone acetylations and methylations in the human genome.** *Nat Genet* 2008, **40**:897-903.  
F1000 Factor 6.4 *Must Read*  
Evaluated by Isaac Kohane 23 Jun 2008, Ian Dunham 18 Jul 2008
- Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nussbaum C, Lander ES, Bernstein BE: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448**:553-60.  
F1000 Factor 4.8 *Recommended*  
Evaluated by Ian Dunham 07 Mar 2008, Hans de Jong 13 Sep 2007
- Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regue A, Rinn JL, Lander ES: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.** *Nature* 2009, **458**:223-7.  
F1000 Factor 6.4 *Must Read*  
Evaluated by Oliver Rando 13 Mar 2009, E John Wherry 01 Apr 2009
- Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nussbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature* 2008, **454**:766-70.
- Wang ET, Sandberg R, Luo S, Khrebukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456**:470-6.  
F1000 Factor 6.4 *Must Read*  
Evaluated by Ken Irvine 20 Nov 2008, Donald Rio 01 Dec 2008
- Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nat Genet* 2008, **40**:1413-5.
- Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nussbaum C, Green RD, Dekker J: **Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements.** *Genome Res* 2006, **16**:1299-309.