Published in final edited form as:

Nat Methods. 2010 January ; 7(1): 47–49. doi:10.1038/nmeth.1404.

Chromatin Profiling by Directly Sequencing Small Quantities of Immunoprecipitated DNA

Alon Goren^{1,2,3,4,7}, Fatih Oszolak^{5,7}, Noam Shores^{1,7}, Manching Ku^{1,2,3,4}, Mazhar Adli^{1,2,3,4}, Chris Hart⁵, Melissa Gymrek^{1,2,3,4}, Or Zuk¹, Aviv Regev^{1,2,6}, Patrice M. Milos⁵, and Bradley E. Bernstein^{1,2,3,4}

¹ Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

² Howard Hughes Medical Institute, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

³ Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

⁴ Center for Systems Biology and Center for Cancer Research, Massachusetts General Hospital, Boston, Massachusetts, USA

⁵ Helicos BioSciences Corporation, Cambridge, Massachusetts, USA

⁶ Massachusetts Institute of Technology Department of Biology, Cambridge, Massachusetts, USA

Abstract

Chromatin structure and transcription factor localization can be assayed genome-wide by sequencing genomic DNA fractionated by protein occupancy or other properties. However, current technologies involve multiple steps that introduce bias and inefficiency. Here we apply a single-molecule approach to directly sequence chromatin immunoprecipitated DNA with minimal sample manipulation. This method is compatible with just 50 picograms of DNA and should thus facilitate charting chromatin maps from limited cell populations.

The distinct cellular phenotypes in multi-cellular organisms are predicated on varied expression programs, determined and stabilized by proteins and chromatin structures that regulate genome function. Methods for analyzing these features typically involve fractionation of genomic DNA based on criteria such as protein occupancy, DNaseI sensitivity, chromatin solubility, or DNA

Correspondences to Bernstein.Bradley@mgh.harvard.edu.

⁷These authors contributed equally to this work.

AUTHOR CONTRIBUTIONS

A.G., N.S. and B.E.B. processed and analyzed the data, wrote the paper and made the figures; F.O. and P.M. developed the method for sequencing ChIP-DNA and performed the sequencing; M.K. performed the chromatin experiments; M.A., M.G., O.Z. and A.R. helped with data analysis.

Data accession

All data are publically available at NCBI (Illumina data at [GSE12241](#) and [GSE18699](#); Helicos data at [SRA009954](#)), and available online as IGV browser tracks at:

<http://www.broadinstitute.org/igv/dynsession/igv.jnlp?sessionURL=http://www.broadinstitute.org/igvdata/Alon/>

Goren_et al.xml&user=goren_et al or as links to the UCSC browser (<http://genome.ucsc.edu/>)¹⁴ at:

www.broadinstitute.org/cgi-bin/seq_platform/chipseq/shared_portal/clone/Helicos.py

List of Competing Financial Interest

There is potential Competing Interest -

F.O. C.H. and P.M.M. are employees of Helicos BioSciences Corporation.

methylation. The enriched DNA can be evaluated by PCR, microarrays or deep sequencing. Approaches that leverage second generation sequencing technologies (SGST) have gained widespread use because they yield sufficiently high read numbers to comprehensively interrogate mammalian genomes. Such approaches have been developed for mapping transcription factors and histone modifications^{1–4}, DNA accessibility^{5,6} and DNA methylation⁷.

Nonetheless, SGST methods remain subject to certain constraints that limit their utility in these applications. Specifically, they involve multiple steps, including molecular and enzymatic manipulations, DNA purifications, size selection and PCR (Supplementary Fig. 1). In part due to these inefficiencies, ~5 nanograms of DNA are typically required for SGST library preparation. This limits enrichment assays to cell types that can be obtained in large numbers. In addition, since library construction procedures (for example, the PCR step) vary in their efficiency as a function of template, sampling bias may be introduced and obscure finer features of the genomic maps.

Recently, Harris *et al.*⁸ introduced technology that enables direct sequencing of single DNA molecules at high-throughput. The HeliScope Genetic Analysis Platform, based on this technology, has since been used to sequence a variety of genomic templates including a complete human genome⁹. This method avoids many of the steps associated with SGST library preparation, such as adapter ligation and PCR. Rather, a single poly-A tailing step yields DNA template compatible with direct sequencing (Supplementary Fig. 1). We reasoned that such an approach could have substantial advantages for interrogating enriched DNA fractions, and therefore explored its suitability for mapping chromatin structure through combination of chromatin immunoprecipitation and sequencing (ChIP-Seq).

In ChIP-Seq^{1,3}, living cells are treated with formaldehyde to fix *in vivo* protein-DNA interactions. Chromatin is then sheared to small fragments (~100–700bp), and immunoprecipitated with antibodies that specifically recognize a modified histone or other DNA-associated protein. The isolated DNA is sequenced, and a discrete representation of enrichment is derived from the distribution of aligned reads. Here, we used a standard ChIP protocol to enrich genomic DNA associated with specific histone modifications (H3K4Me3, H3K27Me3, H3K36Me3) or a DNA-binding protein (CCCTC-binding factor or ‘CTCF’) in murine embryonic stem cells. ChIP DNA samples were then poly-A tailed, loaded into individual channels on the HeliScope instrument, and sequenced-by-synthesis.

For each channel, we generated 20 to 23 million quality filtered reads, which were subsequently aligned to the mouse genome. We could uniquely align 35 to 45% of reads, a lower percentage than typically seen with the Illumina Genome Analyzer (~40–60%; Supplementary Tables 1, 2). This may reflect somewhat higher error rates and shorter read lengths, ranging from 25–55 bases, associated with the Helicos (HeliScope) technology (Supplementary Table 3). We processed aligned Helicos reads into ChIP-Seq maps using a computational pipeline originally developed for SGST data³.

We compared the results from direct sequencing to data acquired using the Illumina Genome Analyzer. In order to facilitate direct comparisons, we truncated matched Helicos and Illumina datasets to have the same number of reads (Supplementary Tables 1, 2). Visual comparison of the maps generated by the two independent technologies suggests good agreement for all four examined epitopes (Fig. 1a). In both datasets, promoters exhibit H3K4me3 peaks coincident both in location and size. Illumina and Helicos data are also in agreement for H3K36me3, which typically covers gene bodies, and H3K27me3, which marks many inactive promoters³. Furthermore, CTCF data acquired with both platforms reveal comparable distributions of peaks, consistent with prior knowledge of CTCF localization¹⁰.

Quantitative analyses confirm strong concordance between the platforms: correlation coefficients for the histone modification data (Fig. 1b, Supplementary Fig. 2) are high (0.95 for H3K4me3 and H3K36me3, and 0.88 for H3K27me3), and are similar to correlations between ChIP-Seq repeats done with Illumina (Supplementary Fig. 3). For the more localized DNA-binding protein CTCF, where the signal distribution is less continuous, we instead assessed coincidence of statistically significant peaks. We specifically compared the top 20,000 non-overlapping genomic locations where CTCF is determined to be present by each of the technologies (Fig. 1c, Supplementary Fig. 4). Here, too, the agreement is remarkable with 75% of high confidence peaks found by one of the methods also found by the other.

Next, we considered whether the elimination of intermediate steps might yield a less biased representation of the DNA fragments within a ChIP sample. The PCR amplification in SGST is perhaps the most substantial difference between the methods. ChIP-Seq procedures typically require 18 or more PCR cycles because of the small DNA quantities obtained by immunoprecipitation. One potential consequence of the amplification would be the presence of multiple identical PCR-copied fragments in the sequencing library, and indeed, the percentage of duplicate reads is much higher in the Illumina data (Supplementary Tables 1,2). In addition to creating redundant copies, PCR tends to amplify certain templates more efficiently than others. One of the known issues associated with shotgun sequencing by SGST is that the representation of sequencing reads can be biased by GC content¹¹. To investigate whether this might affect ChIP-Seq experiments, we used the respective technologies to sequence un-enriched 'control' ChIP DNA samples. These samples should have a relatively uniform representation of genomic sequence and, indeed, the enrichment profiles are largely consistent with this expectation (Supplementary Fig. 5). To explicitly evaluate GC bias in the data, we plotted average sequencing coverage as a function of the GC-content of underlying genomic regions (Fig. 2, Supplementary Fig. 6). We observed a modest over-representation of reads from regions with a GC-content of ~40–65% in the Illumina data, possibly due to bias introduced by PCR or cluster amplification. In contrast, the Helicos sequencing data show a relatively even distribution across 20–80% GC-content.

The set of sequenced reads in a ChIP-Seq experiment also contains other information that may be relevant to the underlying biology. For example, insight into the sizes of genomic regions protected by the ChIP target can be inferred from cross-correlations between positively- and negatively-oriented aligned reads¹². With the Helicos data, such an analysis suggests protection of ~200 bases by H3K4me3 and ~100 bases by CTCF, consistent with the structural distinction between nucleosomal histone and DNA binding protein (Supplementary Fig. 7). In contrast, protected regions inferred from Illumina data are similar for both targets (see Online Methods). Together, these comparisons suggest that Helicos' direct sequencing provides a more faithful readout of enriched genomic fractions, and may thus offer unique insights into the nature of protein-DNA interactions in chromatin.

Finally, we explored whether direct sequencing could be applied to small quantities of ChIP DNA, thereby addressing a major shortcoming of current methods. In the experiments above, we performed direct sequencing on samples of several nanograms. This represents a major improvement over prior direct sequencing reports, is comparable to the minimum SGST sample requirements, and is significantly lower than the 4.5 micrograms used in a recently described amplification-free SGST procedure¹³. Still, an optimal method would be compatible with significantly less starting DNA. In our experience, a typical histone modification ChIP performed on 500,000 cells yields ~1 nanogram of DNA. Thus, ChIP-Seq analysis of 50,000 cells would require the interrogation of ~100 picograms of DNA.

We therefore sought to develop a direct sequencing protocol that would be compatible with small quantities of ChIP DNA. We found that carriers such as oligoribonucleotides and

oligonucleotides covalently-attached to solid surfaces facilitate A-tailing of low-attomole level DNA material and reduce sample loss during the tailing and surface-capture steps (see Online Methods).

We successfully tailed and sequenced 50 and 150 picogram samples of H3K4Me3 ChIP DNA, obtained by dilution, as well as a 200 picogram sample of CTCF ChIP DNA. These experiments yielded between 3.6 and 5 million aligned reads, somewhat lower than the numbers achieved in the initial experiments (Supplementary Table 1). Encouragingly, enrichment profiles derived from these data appear to offer robust and accurate signals. Despite having fewer reads, the ChIP-Seq maps for the small quantity samples show exquisite correlation with the datasets acquired from 3 nanograms of ChIP DNA (Fig. 3, Supplementary Fig. 8–11).

The lower numbers of aligned reads obtained with the small DNA amounts are expected to reduce sensitivity. Indeed, some enriched regions detected in the large sample experiments do not appear in these maps. Systematic comparison of the H3K4me3 datasets suggests that the sensitivity of the 50 picogram dataset is ~5% lower than the data collected from the original 3 nanogram sample (Supplementary Fig. 9–11). Accordingly, it may be necessary to perform additional replicates when analyzing small quantity ChIP samples.

In conclusion, we combined direct sequencing with ChIP for genome-wide analysis of chromatin structure and transcription factor localization. Data collected with this method show high concordance to the existing SGST standard. The direct approach also offers certain benefits, including streamlined sample preparation and reduced representation bias. While SGST bias is relatively small and appears not to interfere with discovery of robust features, direct ChIP-Seq may facilitate detection of subtle yet important effects. Conversely, although direct sequencing can map a significant majority of the genome, applications that require higher genome coverage or detailed information on repetitive regions may benefit from longer read lengths and paired-end information offered by SGST platforms. Finally, we demonstrate that direct sequencing can be applied to very small quantities of ChIP DNA. This relaxed sample requirement should enable charting of genome-wide chromatin maps from previously inaccessible cell populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We wish to thank J. Robinson and IGV platform (<http://www.broad.mit.edu/igv>) for their help in data presentation. A.G. is supported by an EMBO long term postdoctoral fellowship. M.K. is supported by a Croucher Foundation fellowship. A.R. is an investigator of the Merkin Foundation for Stem Cell Research at the Broad Institute. B.E.B. is a Charles E. Culpeper Medical Scholar. A.R. and B.E.B. are Early Career Scientist of the Howard Hughes Medical Institute. This research was supported by funds from the Burroughs Wellcome Fund (to B.E.B. and A.R.), HHMI and the National Human Genome Research Institute.

References

1. Barski A, Cuddapah S, Cui K, et al. *Cell* 2007;129(4):823. [PubMed: 17512414]
2. Johnson DS, Mortazavi A, Myers RM, et al. *Science* 2007;316(5830):1497. [PubMed: 17540862]
3. Mikkelsen TS, Ku M, Jaffe DB, et al. *Nature* 2007;448(7153):553. [PubMed: 17603471]
4. Robertson G, Hirst M, Bainbridge M, et al. *Nat Methods* 2007;4(8):651. [PubMed: 17558387]
5. Boyle AP, Davis S, Shulha HP, et al. *Cell* 2008;132(2):311. [PubMed: 18243105]
6. Hesselberth JR, Chen X, Zhang Z, et al. *Nat Methods* 2009;6(4):283. [PubMed: 19305407]
7. Down TA, Rakyan VK, Turner DJ, et al. *Nat Biotechnol* 2008;26(7):779. [PubMed: 18612301]

8. Harris TD, Buzby PR, Babcock H, et al. Science 2008;320(5872):106. [PubMed: 18388294]
9. Pushkarev D, Neff NF, Quake SR. Nat Biotechnol 2009;27(9):847. [PubMed: 19668243]
10. Kim TH, Abdullaev ZK, Smith AD, et al. Cell 2007;128(6):1231. [PubMed: 17382889]
11. Dohm JC, Lottaz C, Borodina T, et al. Nucleic Acids Res 2008;36(16):e105. [PubMed: 18660515]
12. Tolstorukov MY, Kharchenko PV, Goldman JA, et al. Genome Res 2009;19(6):967. [PubMed: 19246569]
13. Kozarewa I, Ning Z, Quail MA, et al. Nat Methods 2009;6(4):291. [PubMed: 19287394]
14. Kent WJ, Sugnet CW, Furey TS, et al. Genome Res 2002;12(6):996. [PubMed: 12045153]
15. Bowers J, Mitchell J, Beer E, et al. Nat Methods. 2009
16. Lipson D, Raz T, Kieu A, et al. Nat Biotechnol 2009;27(7, 6):52.
17. Batzoglou S, Jaffe DB, Stanley K, et al. Genome Res 2002;12(1):177. [PubMed: 11779843]

Figure 1. Comparison of ChIP-Seq data acquired by Illumina or Helicos sequencing

(a) Genomic tracks display ChIP-seq data (I, Illumina; H, Helicos). Shown are enrichment profiles for CTCF (black), H3K4me3 (green), H3K27me3 (red) and H3K36me3 (blue) across a 300kb region in mouse ES cells. The UCSC known genes track is shown below the profiles. (b) Quantitative comparison of histone modifications from ChIP data sequenced by Illumina or Helicos. Scatter plots show signals for H3K4me3 (1kb bins), H3K27me3 (5kb bins) and H3K36me3 (5kb bins) across the genome. Indicated on each plot is the Pearson correlation coefficient, ρ . (c) Quantitative comparison of CTCF ChIP-Seq data. Venn diagram indicates overlap between top 20,000 genomic locations bound by CTCF as determined by Illumina (red) or Helicos (blue). The comparisons confirm strong concordance for data collected on the respective platforms.

Figure 2. Experimental bias associated with SGST procedures

Normalized read density obtained from un-enriched control ‘whole cell extract’ (WCE) samples analyzed by Illumina (blue) or Helicos (black) is plotted as a function of GC percentage for 100bp windows. A theoretical ‘expected’ distribution obtained by computational simulation is shown in red (Random). The Illumina data modestly over-represent reads from regions with a GC-content of ~40–65%, while the Helicos data show a relatively even distribution across 20–80% GC-content.

Figure 3. Comparison of ChIP-Seq data obtained for small quantity samples

(a) Genomic view displays H3K4me3 ChIP-seq data obtained with either Illumina (I), standard Helicos (H) or small sample size Helicos methodology (H 150pg and H 50pg). Shown are enrichment signals across a 440kb region in mouse ES cells. The UCSC known genes track appears below the profiles. (b–c) ChIP-Seq data derived from 50 picograms of sample compared against data derived from 3 nanogram or 150 picogram samples. Scatter plots show H3K4me3 signal for 1 kb bins across the genome. Pearson correlation, ρ , is indicated for each comparison. (d) Genomic view displays ChIP-seq data obtained with either Illumina (I), standard Helicos (H) or small sample size Helicos methodology (H 200pg) for a 300kb regions in mouse ES cells. These datasets were not truncated and thus have different read numbers. (e) Quantitative comparison of CTCF ChIP-Seq data derived by sequencing either 200 picograms or 3 nanograms of ChIP DNA by Helicos. Venn diagram indicates overlap between the top 20,000 enriched genomic locations in each dataset.