Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA

Alon Goren[1–4,7], Fatih Ozsolak[5,7], Noam Shoresh[1,7], Manching Ku[1–4], Mazhar Adli[1–4], Chris Hart[5], Melissa Gymrek[1–4], Or Zuk[1], Aviv Regev[1,2,6], Patrice M Milos[5] & Bradley E Bernstein[1–4]

Chromatin structure and transcription factor localization can be assayed genome-wide by sequencing genomic DNA fractionated by protein occupancy or other properties, but 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 pg of DNA and should thus facilitate charting chromatin maps from limited cell populations.

The distinct cellular phenotypes in multicellular 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, DNase I sensitivity, chromatin solubility or DNA methylation. The enriched DNA can be evaluated by PCR, microarrays or deep sequencing. Approaches that leverage second-generation sequencing technologies (SGSTs) 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 methylation[7].

Nonetheless, SGSTs 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 owing to these inefficiencies, ~5 ng 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, as 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, a technology that enables direct sequencing of single DNA molecules at high throughput has been introduced[8]. The HeliScope Genetic Analysis platform, based on this technology, has since been used to sequence a variety of genomic templates including a complete human genome[9]. This method avoids many of the steps associated with SGST library preparation, such as adaptor 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 a 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 (~100–700 bp) 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 and H3K36Me3) or a DNA-binding protein (CCCTC-binding factor or CTCF) in mouse embryonic stem cells. Then we poly(A)-tailed ChIP DNA samples, loaded them into individual channels on the HeliScope instrument and sequenced them by synthesis.

For each channel, we generated 20–23 million quality filtered reads, which we then aligned to the mouse genome. We could uniquely align 35–45% of reads, less than typically seen with SGST on the Illumina Genome Analyzer (~40–60%; **Supplementary Tables 1** and **2**). This may reflect somewhat higher error rates and shorter read lengths (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 data[3].

We compared the results from direct sequencing to data acquired using the Illumina Genome Analyzer. To facilitate direct comparisons, we truncated matched Helicos and Illumina datasets to have the same number of reads (**Supplementary Tables 1** and **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 exhibited H3K4me3 peaks coincident both in location and size. Illumina and Helicos data were also in agreement for H3K36me3, which typically covers

[1]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Howard Hughes Medical Institute. [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]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [7]These authors contributed equally to this work. Correspondence should be addressed to B.E.B. (bernstein.bradley@mgh.harvard.edu).

RECEIVED 5 AUGUST; ACCEPTED 3 NOVEMBER; PUBLISHED ONLINE 29 NOVEMBER 2009; DOI:10.1038/NMETH.1404

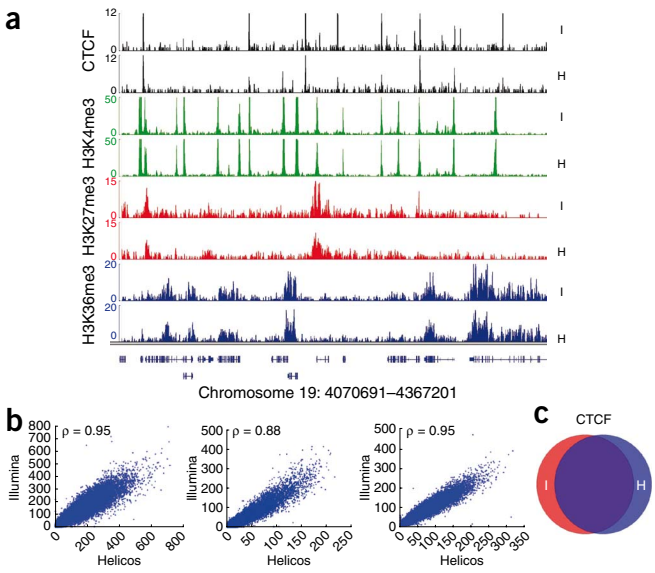

Figure 1 | Comparison of ChIP-seq data acquired by Illumina or Helicos sequencing. (**a**) Genomic tracks display ChIP-seq data (I, Illumina; H, Helicos) for CTCF, H3K4me3, H3K27me3 and H3K36me3 across a 300-kb region in mouse embryonic stem cells. The RefSeq genes track is shown below the profiles. (**b**) Quantitative comparison of histone modifications from ChIP data sequenced by Illumina or Helicos. Scatter plots show signals for H3K4me3 (1-kb bins; left), H3K27me3 (5-kb bins; middle) and H3K36me3 (5-kb bins; right) across the genome. Pearson correlation coefficients, ρ, are indicated. (**c**) Overlap between the top 20,000 genomic locations (1-kb windows) bound by CTCF as determined using Illumina (I) or Helicos (H) data was 15,327 genomic locations.

gene bodies, and for H3K27me3, which marks many inactive promoters[3]. Furthermore, CTCF data acquired with both platforms revealed comparable distributions of peaks, consistent with prior knowledge of CTCF localization[10].

Quantitative analyses confirmed strong concordance between the platforms: correlation coefficients for the histone modification data (**Fig. 1b** and **Supplementary Fig. 2**) were high (0.95 for H3K4me3 and H3K36me3, and 0.88 for H3K27me3) and were similar to correlations between ChIP-seq repeats done with the Illumina Genome Analyzer (**Supplementary Fig. 3**). For the more localized DNA-binding protein CTCF, for which the signal distribution was less continuous, we instead assessed coincidence of statistically significant peaks. We compared the top 20,000 nonoverlapping genomic locations at which we determined CTCF to be present by each of the technologies (**Fig. 1c** and **Supplementary Fig. 4**). Here, too, the agreement was high, 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 in 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 was much higher in the Illumina data (**Supplementary Tables 1** and **2**). In addition to creating redundant copies, PCR tends to amplify certain templates more efficiently than others.

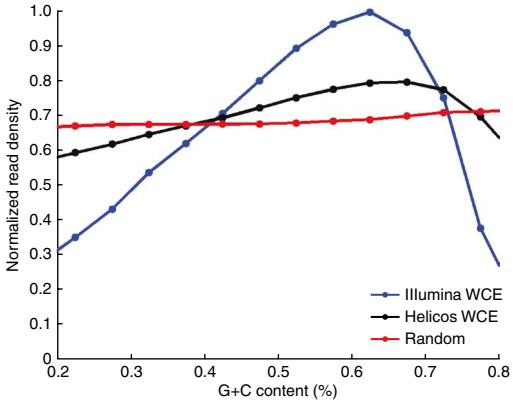

Figure 2 | Experimental bias associated with SGST procedures. Normalized read density obtained from unenriched control 'whole cell extract' (WCE) samples analyzed by Illumina or Helicos platforms is plotted as a function of G+C percentage for 100-bp windows. A theoretical 'expected' distribution obtained by computational simulation is also shown (random).

One of the known issues associated with shotgun sequencing by SGST is that the representation of sequencing reads can be biased by G+C content[11]. To investigate whether this might affect ChIP-seq experiments, we used the respective technologies to sequence unenriched 'control' ChIP DNA samples. These samples should have a relatively uniform representation of genomic sequence and, indeed, the enrichment profiles were largely consistent with this expectation (**Supplementary Fig. 5**). To explicitly evaluate G+C bias in the data, we plotted average sequencing coverage as a function of the G+C content of underlying genomic regions (**Fig. 2** and **Supplementary Fig. 6**). We observed a modest overrepresentation of reads from regions with a G+C content of ~40–65% in the Illumina data, possibly owing to bias introduced by PCR or cluster amplification. In contrast, the Helicos sequencing data had a relatively even distribution across 20–80% G+C content.

The sequenced reads in a ChIP-seq experiment also contain 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 reads[12]. With the Helicos data, such an analysis suggested 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 were similar for both targets (Online Methods). Together, these comparisons suggested that 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 we could directly sequence small quantities of ChIP DNA, thereby addressing a major shortcoming of current methods. In the experiments above, we directly sequenced several-nanogram samples. This was a major improvement over prior direct sequencing reports, was comparable to the minimum SGST sample requirements and was much lower than the 4.5 μg used in a recently described amplification-free SGST procedure[13]. Still, an optimal method would be compatible with much less starting DNA. In our experience, a typical histone

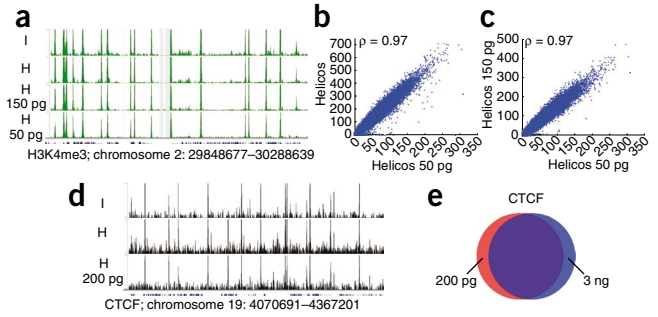

Figure 3 | Comparison of ChIP-seq data obtained for small-quantity samples. (**a**) Genomic view displays H3K4me3 ChIP-seq data obtained with either Illumina (I), standard Helicos (H) or small-sample-size Helicos methodology (H 150 pg and H 50 pg). Shown are enrichment signals across a 440-kb region in mouse embryonic stem cells. The RefSeq genes track is shown below the profiles. (**b**,**c**) ChIP-seq data derived from a 50 pg sample compared to data derived from 3 ng (**b**; Helicos) or 150 pg (**c**; Helicos 150 pg) 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 200 pg) for a 300-kb region in mouse embryonic stem cells. These datasets were not truncated and thus have different read numbers. The RefSeq gene track are shown below the profiles. (**e**) Quantitative comparison of CTCF ChIP-seq data derived by sequencing either 200 pg (red) or 3 ng (blue) of ChIP DNA by Helicos methodology. The overlap between the top 20,000 enriched genomic locations in each dataset was 16,478 genomic locations.

modification ChIP performed on 500,000 cells yields ~1 ng of DNA. Thus, ChIP-seq analysis of 50,000 cells would require the interrogation of ~100 pg 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 facilitated A-tailing of low-attomolar DNA material and reduced sample loss during the tailing and surface-capture steps (Online Methods).

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

We expected the lower numbers of aligned reads obtained with the small DNA amounts to reduce sensitivity. Indeed, some enriched regions detected in the large sample experiments did not appear in these maps. Systematic comparison of the H3K4me3 datasets suggested that the sensitivity of the 50 pg dataset was ~5% lower than the data collected from the original 3 ng sample (**Supplementary Figs. 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 had high concordance to the existing SGST standard. The direct approach offered benefits, including streamlined sample preparation and reduced representation bias. Whereas SGST bias was relatively small and appeared not to interfere with discovery of robust features, direct ChIP-seq may facilitate detection of subtle yet important effects. Conversely, although direct sequencing can be used to map the majority of the genome, applications that require greater 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.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession codes. Gene Expression Ombibus (GEO): GSE12241 and GSE18699 (Illumina data), and SRA009954 (Helicos data).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank J. Robinson and members of the IGV platform for their help with 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. This research was supported by funds from the Burroughs Wellcome Fund (to B.E.B. and A.R.), Howard Hughes Medical Institute (to B.E.B. and A.R.), Partnership for Cures Culpeper Scholarship (to B.E.B.) and the US National Human Genome Research Institute.

AUTHOR CONTRIBUTIONS

A.G., N.S. and B.E.B. processed and analyzed the data, wrote the paper and made the figures; F.O., C.H. and P.M.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.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Barski, A. *et al. Cell* **129**, 823–837 (2007).
2. Johnson, D.S. *et al. Science* **316**, 1497–1502 (2007).
3. Mikkelsen, T.S. *et al. Nature* **448**, 553–560 (2007).
4. Robertson, G. *et al. Nat. Methods* **4**, 651–657 (2007).
5. Boyle, A.P. *et al. Cell* **132**, 311–322 (2008).
6. Hesselberth, J.R. *et al. Nat. Methods* **6**, 283–289 (2009).
7. Down, T.A. *et al. Nat. Biotechnol.* **26**, 779–785 (2008).
8. Harris, T.D. *et al. Science* **320**, 106–109 (2008).
9. Pushkarev, D., Neff, N.F. & Quake, S.R. *Nat. Biotechnol.* **27**, 847–852 (2009).
10. Kim, T.H. *et al. Cell* **128**, 1231–1245 (2007).
11. Dohm, J.C. *et al. Nucleic Acids Res.* **36**, e105 (2008).
12. Tolstorukov, M.Y. *et al. Genome Res.* **19**, 967–977 (2009).
13. Kozarewa, I. *et al. Nat. Methods* **6**, 291–295 (2009).

ONLINE METHODS

Cell culture. Mouse embryonic stem (ES) cells (V6.5; male; genotype 129SvJaexC57BL/6; passages 10–15) were grown in 5% CO_2 at 37 °C on irradiated mouse embryonic fibroblasts (MEFs) in DMEM containing 15% FCS, ESGRO, penicillin-streptomycin, Glutamax (Invitrogen), nonessential amino acids and 2-mercaptoethanol. ES cells were passaged 2–3 times on 0.2% gelatin-coated plates to remove MEF contamination.

Chromatin immunoprecipitation (ChIP). ChIP experiments were carried out as described previously[3]. Briefly, cells were fixed using formaldehyde and chromatin was fragmented to a size range of 200–700 bp with a Branson 250 Sonifier. Solubilized chromatin was immunoprecipitated with antibody to H3K4me3 (Abcam 8580), H3K27me3 (Upstate 07-449), H3K36me3 (Abcam 9050) or CTCF (Upstate 07-729). Protein A–sepharose was used to pull down the antibody-chromatin complexes which were then washed and eluted. After cross-link reversal (5 h, 65 °C) and proteinase K digestion, immunoprecipitated DNA was extracted with phenol-chloroform, ethanol-precipitated and treated with RNase.

Single-molecule sequencing. ChIP DNA was quantified using PicoGreen (Invitrogen) for sequencing trials using differing amounts of starting materials. Initial sequencing method development trials were carried out starting from 3 ng of ChIP DNA as follows. Samples were spiked with fluorescently labeled DNA oligonucleotides (5′-GCGGTGACACGGGAGATCTGAACTCGT ACT-3′) that could then be used to monitor and calibrate tailing and blocking conditions across a range of DNA quantities (100 amol–5 fmol) of ChIP DNA. Spiked samples were denatured at 95 °C for 5 min and then snap-cooled on ice. After addition of terminal transferase (New England Biolabs), BSA and dATP, samples were incubated at 37 °C for 1 h, followed by enzyme inactivation at 70 °C for 10 min. The blocking step was performed by another round of terminal transferase addition in the presence of 100 pmol ddTTP, incubating at 37 °C for 1 h, followed by enzyme inactivation. Tailing efficiency was then evaluated by using capillary electrophoresis (ABI 3730; Applied Biosystems) to determine the fraction of oligonucleotides that were tailed and to evaluate the distribution of tail lengths. In this way, dATP:enzyme: template ratios that resulted in optimal tailing and blocking of ChIP DNA samples could be identified. The most effective condition used 4 U terminal transferase and 200 pmol dATP, yielding 80–150-nucleotide poly(A) tails.

Using the methods described above, which confirmed our ability to effectively tail and sequence small quantities of materials, 3–9 ng of ChIP DNA was tailed using the optimal tailing conditions described above. Tailed DNA samples were supplemented with 3′ dideoxy–blocked oligonucleotide (5′-TCACTATTGTTG AGAACGTTGGCCTATAGTTCGTATTACGCGC GGTGACACG GGAGATCTGAACTCGTACTCACGddC-3′) to minimize sample loss during hybridization. Samples were hybridized to the flowcells and single-molecule sequencing-by-synthesis with reversible terminators[14] was carried out using standard Heliscope procedures (Helicos BioSciences Corporation).

We next explored other conditions and implementations to improve the procedure and to reduce sample requirements even more. First, we experimented with using additional carrier molecules. We used RNA oligoribonucleotides for this purpose as they are not a substrate for terminal transferase under the conditions used here and therefore do not attach to the oligo(dT) sequencing surface. We found that the addition of 2 pmol of oligoribonucleotide (5′-CGUCAGGGCAGAGGAUGGAUGCAA GGAUAAGUGGA-3′) stabilized the tailing reaction and allowed for addition of higher concentrations of enzyme (20 U) and dATP (400 pmol). We also used a second carrier approach involving DNA oligonucleotides covalently attached to a solid surface (magnetic dynabeads; Invitrogen). In this approach, the carrier DNA oligonucleotide is A-tailed along with the sample DNA, but can then be removed from the reaction before hybridization. We added 5 µl of oligonucleotide-coated beads to the tailing reaction (after three washes in 1× terminal transferase buffer) that included the following: heat-denatured ChIP DNA, 400 pmol dATP and 20 U of terminal transferase in a 20 µl final volume. The blocking step was performed by adding 200 pmol ddTTP and 20 units of terminal transferase. Incubation steps were performed as described above. The beads were removed by placing the reaction on a magnetic stand. These implementations enabled the analysis of as little as 50 pg of ChIP DNA sample (as measured before A-tailing), which roughly corresponds to a two orders of magnitude reduction in sample requirements compared to existing standards; this was the lowest DNA amount that could be A-tailed and still give sufficient aligned read numbers for ChIP.

Processing of sequencing reads. We used sequencing reads acquired by direct sequencing on the HeliScope instrument ranging from 25 to 75 bases in length (**Supplementary Fig. 12** and **Supplementary Table 4**). These reads were aligned to the *Mus musculus* February 2006 assembly (mm8) assembly of the mouse genome using IndexDP Genomic algorithm[15] aligner. Sequences generated by the Illumina platform were aligned using ARACHNE[16], as described previously[3]. Sequences with more than a single best match in the reference genome were discarded. We also created a map of 'unalignable' genomic positions to which no unique 36-base read could be uniquely aligned owing to inherent sequence redundancy for consideration in downstream processing.

Integration of enrichment profiles. The short sequence reads acquired by these technologies correspond to the ends of the DNA fragments in the library (Illumina) or the ChIP sample itself (Helicos). Since the average size of the ChIP fragments is in the range of 200 bp, we extended the aligned reads *in silico* to a total length of 200 bases. Digital maps of sequence coverage or 'enrichment profiles' were then computed by counting the number of simulated ChIP fragments that overlap a given genomic position (calculated for a 25-bp sliding window). The signal in a given window was calculated as the number of extended reads that it overlapped (entirely or in part). Genomic windows for which more than 10% of included bases were 'unalignable' were excluded from further analysis to avoid spurious peaks in the ChIP-seq maps often associated with repetitive genomic sequence. Genomic tracks were visualized using the IGV (http://www.broad.mit.edu/igv).

Comparative analysis of Helicos and Illumina data. The Helicos datasets were compared to maps derived for biological replicate ChIP samples sequenced on the Illumina Genome Analyzer,

doi:10.1038/nmeth.1404

as described previously[3]. To facilitate direct comparisons, we trimmed corresponding Illumina and Helicos datasets for each modification so they had the same numbers of aligned reads. This was done by randomly discarding aligned reads from the larger of the two datasets until it had the same number of aligned reads as the other. These matched datasets were used to generate ChIP-seq signal tracks as above, which we used for all comparative analysis, including those shown in **Figures 1a–c** and **3a–c**. Scatter plots (also displayed as $\log_{10}$-transformed values and two-dimensional histograms; **Supplementary Figs. 2**,**6** and **13**), Venn diagrams and correlation coefficients were based on summed signals across bins of 1 kb (H3K4me3 and CTCF) or 5 kb (H3K27me3 and H3K36me3). Correlations are Pearson correlation coefficients between the vectors of genome-wide signal values in all non-overlapping bins of the sizes described above.

The G+C bias plot in **Figure 2** was computed for 100-bp windows. The total number of aligned 'whole cell extract' (WCE) reads was computed for each window. Percentage G+C content was calculated based on fraction of guanine and cysosine nucleotides in corresponding windows of the mm8 build. We then plotted the average read number as a function of percentage G+C content. To generate a theoretical 'expected' distribution, 35-base intervals were randomly selected from either the forward or reverse strands of the mouse genome (mm8), realigned back to the reference genome using the same uniqueness criteria and plotted as above.

Partial insight into the size distribution of sequenced fragments may be gained by measuring the cross-correlation between positively and negatively oriented aligned reads[12]. The size of the sequenced library fragments may contribute to a peak in the cross-correlation plot when multiple PCR copies are read from both ends. Such a peak could also reflect a tendency for the ends of ChIP fragments to lie on either side of a genomic region protected by the protein target. Strand cross-correlations were calculated using aligned reads from corresponding Illumina and Helicos datasets, trimmed to have identical read numbers. Aligned reads were parsed into a set of positive reads and a set of negative reads. Each read was counted as 'hitting' just its first base to yield two vectors that represent the signal originating from the two DNA strands, at single-base resolution. If $P(x)$ is the number of positively oriented reads that have x as their starting position and $N(x)$ is the corresponding number for negatively oriented reads, then the cross-correlation function at shift Δ was calculated as

$$\rho(\Delta) = \frac{\mathrm{COV}(P(x+\Delta), N(x))}{\sqrt{\sigma_P^2 \sigma_N^2}},$$

with

$$\mathrm{COV}(P(x+\Delta), N(x)) = E(P(x+\Delta) \times N(x)) - E(P) \times E(N),$$

and

$$\sigma_P^2 = E(P^2) - E(P)^2$$

(similarly for σ_N^2). Here $E(.)$ denotes the expectation value.

Data visualization. All data can be visualized online as Integrative Genomics Viewer (IGV) browser tracks at http://www.broadinstitute.org/igv/dynsession/igv.jnlp?sessionURL=http://www.broadinstitute.org/igvdata/Alon/Goren_etal.xml&user=goren_etal/ or as links to the University of California Santa Cruz (UCSC) genome browser at http://www.broadinstitute.org/cgi-bin/seq_platform/chipseq/shared_portal/clone/Helicos.py/.

14. Bowers, J. *et al. Nat. Methods* **6**, 593–595 (2009).
15. Lipson, D. *et al. Nat. Biotechnol.* **27**, 652–658 (2009).
16. Batzoglou, S. *et al. Genome Res.* **12**, 177–189 (2002).