Systematic evaluation of factors influencing ChIP-seq fidelity

Yiwen Chen[1,12], Nicolas Negre[2,11,12], Qunhua Li[3], Joanna O Mieczkowska[4], Matthew Slattery[2], Tao Liu[1], Yong Zhang[5], Tae-Kyung Kim[6,11], Housheng Hansen He[1], Jennifer Zieba[2], Yijun Ruan[7], Peter J Bickel[8], Richard M Myers[9], Barbara J Wold[10], Kevin P White[2], Jason D Lieb[4] & X Shirley Liu[1]

We evaluated how variations in sequencing depth and other parameters influence interpretation of chromatin immunoprecipitation–sequencing (ChIP-seq) experiments. Using *Drosophila melanogaster* S2 cells, we generated ChIP-seq data sets for a site-specific transcription factor (Suppressor of Hairy-wing) and a histone modification (H3K36me3). We detected a chromatin-state bias: open chromatin regions yielded higher coverage, which led to false positives if not corrected. This bias had a greater effect on detection specificity than any base-composition bias. Paired-end sequencing revealed that single-end data underestimated ChIP-library complexity at high coverage. Removal of reads originating at the same base reduced false-positives but had little effect on detection sensitivity. Even at mappable-genome coverage depth of ~1 read per base pair, ~1% of the narrow peaks detected on a tiling array were missed by ChIP-seq. Evaluation of widely used ChIP-seq analysis tools suggests that adjustments or algorithm improvements are required to handle data sets with deep coverage.

ChIP-seq has become the predominant technique for profiling *in vivo* DNA-protein interactions[1,2] and histone marks[3,4] genome-wide. Multiple factors in the experimental design and data analysis influence the final interpretation of a ChIP-seq experiment. One important factor is the potential bias in the genomic coverage of sequencing reads, which can confound the true signal of interest. A second factor is whether the DNA libraries are prepared for paired-end or single-end sequencing. Paired-end libraries are well suited to characterize genomic rearrangements and identify new chimeric transcripts or alternative splice isoforms. However, the benefits of paired-end libraries for a standard ChIP-seq experiment are unclear. A third factor is the absolute and relative sequencing depth of the ChIP and chromatin input samples used as control for background signal. Chromatin input samples are generated by fragmentation or enzymatic digestion of chromatin extracts (**Supplementary Notes**). ChIP-seq is presumed to have many advantages over ChIP followed by array hybridization (ChIP-chip)[5]; some advantages, such as greater resolution and better genome coverage, are proven[6,7], but others, such as higher sensitivity and larger dynamic range, remain to be tested in a direct comparison between ChIP-chip and ChIP-seq data at a deep coverage from the same samples. A fourth factor is the computational algorithm that is used for ChIP-seq peak calling. In an earlier systematic study of ChIP-chip performance, the choice of the analysis algorithm and parameters had a larger effect on the accuracy of the final results than any other single experimental factor[5]. The most popular ChIP-seq peak callers had been developed and evaluated based on early low-coverage ChIP-seq[8,9] or simulated data sets (http://seqanswers.com/forums/showthread.php?t=1039 and http://sourceforge.net/projects/useq/files/CommunityChIPSeqChallenge/).

To evaluate the aforementioned factors, we generated high-quality ChIP-seq data sets (**Supplementary Notes**) from *Drosophila* S2 cells with a depth of ~1 read per base pair (bp) of mappable fly genome (corresponding to ~2.4 billion reads in the human)[10] enriching for the site-specific transcription factor Suppressor of Hairy-wing (Su(Hw))[11], which yielded narrow peaks and the broadly distributed histone mark H3K36me3 (refs. 12–14).

RESULTS

The effect of DNA base composition and chromatin state

In a ChIP-seq experiment, biases could be introduced during processing, for example, PCR amplification and library preparation, and sequencing of DNA fragments. Consistent with earlier results[15,16], sequencing reads from our genomic DNA (gDNA)

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts, USA. [2]Institute for Genomics and Systems Biology, Department of Human Genetics, The University of Chicago, Chicago, Illinois, USA. [3]Department of Statistics, Penn State University, University Park, Pennsylvania, USA. [4]Department of Biology, Carolina Center for the Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina, USA. [5]Department of Bioinformatics, School of Life Science and Technology, Tongji University, Shanghai, China. [6]Department of Neurobiology, Harvard Medical School, Boston, Massachusetts, USA. [7]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore. [8]Department of Statistics, University of California, Berkeley, California, USA. [9]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA. [10]Division of Biology, California Institute of Technology, Pasadena, California, USA. [11]Present addresses: Institut National de la Recherche Agronomique–Université de Montpellier II, Unité Mixte de Recherche 1333, Montpellier, France (N.N.) and Department of Neuroscience, University of Texas Southwestern Medical Center, Dallas, Texas, USA (T.-K.K.). [12]These authors contributed equally to this work. Correspondence should be addressed to X.S.L. (xsliu@jimmy.harvard.edu), J.D.L. (jlieb@bio.unc.edu) or K.P.W. (kpwhite@uchicago.edu).

RECEIVED 23 NOVEMBER 2011; ACCEPTED 26 MARCH 2012; PUBLISHED ONLINE 22 APRIL 2012; DOI:10.1038/NMETH.1985

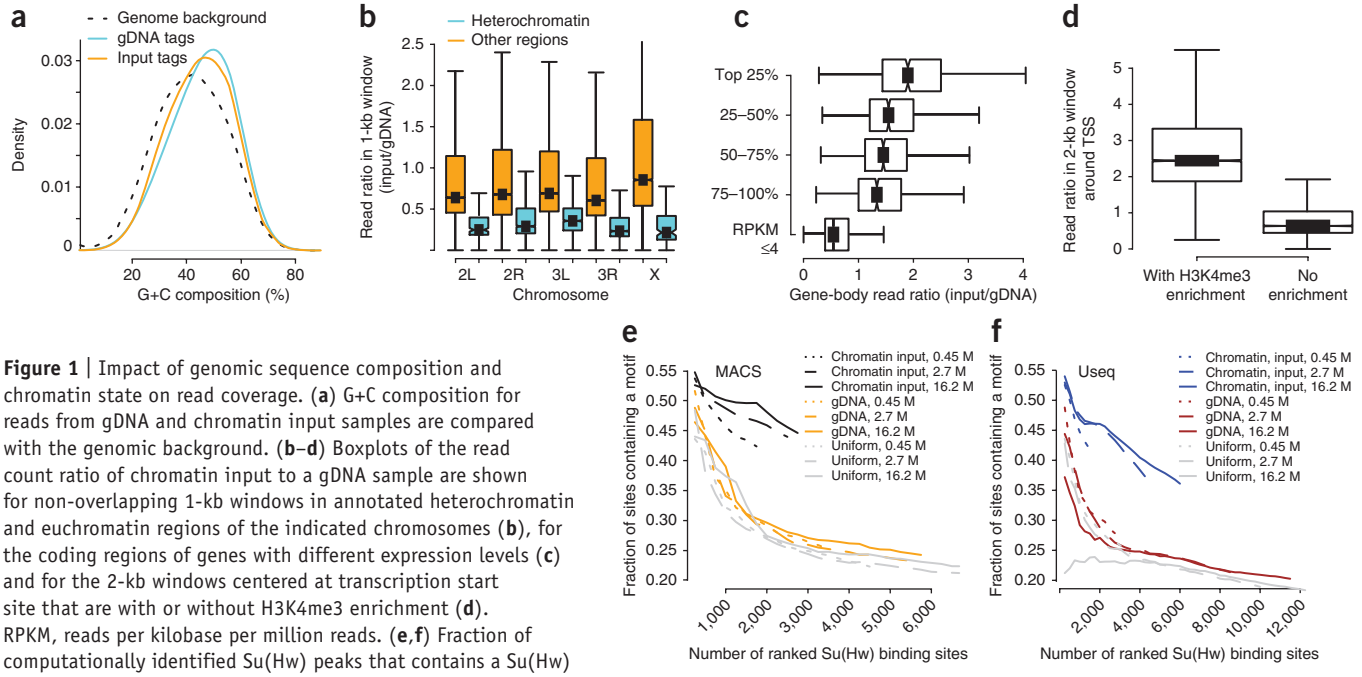

Figure 1 | Impact of genomic sequence composition and chromatin state on read coverage. (**a**) G+C composition for reads from gDNA and chromatin input samples are compared with the genomic background. (**b**–**d**) Boxplots of the read count ratio of chromatin input to a gDNA sample are shown for non-overlapping 1-kb windows in annotated heterochromatin and euchromatin regions of the indicated chromosomes (**b**), for the coding regions of genes with different expression levels (**c**) and for the 2-kb windows centered at transcription start site that are with or without H3K4me3 enrichment (**d**). RPKM, reads per kilobase per million reads. (**e**,**f**) Fraction of computationally identified Su(Hw) peaks that contains a Su(Hw) binding motif plotted as a function of the number of top-ranked binding sites for different types of controls (chromatin input, gDNA and a uniform background) and for MACS (**e**) and Useq (**f**). The ranking is based on the significance of each peak that was assigned by individual algorithms.

samples had a higher G+C content than those from the whole-genome background (Online Methods and **Fig. 1a**). Sequencing reads from the chromatin input and gDNA samples had different G+C composition distributions (median, 44% and 47%, respectively; Mann-Whitney test, $P < 2.2 \times 10^{-16}$; **Fig. 1a**), suggesting that chromatin may affect sequencing coverage.

We compared the gDNA read count–normalized coverage of the chromatin input sample in different genomic regions using read ratios of the chromatin input to the gDNA sample in non-overlapping 1-kilobase (kb) windows. We first compared heterochromatin and euchromatin based on the annotation from University of California Santa Cruz (UCSC) *Drosophila melanogaster* genome assembly (dm3) (Online Methods). Read ratios of the chromatin input to the gDNA sample in heterochromatin regions were significantly lower than those in euchromatin (Mann-Whitney test, $P < 2.2 \times 10^{-16}$, **Fig. 1b**). Comparison of sequencing coverage in enriched and depleted regions of 15 histone marks[17–19] (Online Methods) confirmed that normalized chromatin input coverage correlated positively with active histone marks and negatively with repressive histone marks (**Supplementary Fig. 1**). We also observed higher coverage in euchromatin on the X chromosome than euchromatin of autosomes in the male-derived S2 lines (**Fig. 1b**). This is consistent with the dosage-compensation mechanism in *Drosophila*[20].

Genes with higher expression had higher read ratios in gene bodies (Mann-Whitney test, $P < 7.2 \times 10^{-7}$, Online Methods and **Fig. 1c**), and promoter regions with H3K4me3 enrichment had higher read ratios than those without H3K4me3 (**Fig. 1d**, Mann-Whitney test, $P < 2.2 \times 10^{-16}$). These observations agree with results in *Saccharomyces cerevisiae*[21] and indicate that coverage was higher in regions with more open chromatin states both chromosome-wide and for individual genes.

To characterize the impact of G+C bias and chromatin-state bias on the identification of ChIP-enriched regions, we identified Su(Hw) peaks using two different algorithms, the same ChIP data but with 'control' data from either chromatin input, gDNA or generated from a uniform background model across the genome that ignores G+C bias and chromatin-state bias. The gDNA data did not contain chromatin-state information and served only to correct the G+C bias. The chromatin input control corrected for both the G+C bias and the chromatin-state bias. Peaks identified using chromatin input as a control showed much better enrichment of the Su(Hw) binding motif than those identified by other controls (**Fig. 1e**,**f**).

If we consider the fraction of peaks that did not contain a motif as a crude proxy of false discovery rate for peak calling, then at a fixed false discovery rate, using the chromatin input control resulted in more discovered binding sites than using other controls (**Fig. 1e**,**f**). We missed 4–10% of ChIP-enriched regions identified using chromatin inputs by using other controls, indicating that ignoring the G+C bias and the chromatin-state bias also had a negative effect on detection sensitivity.

Single-end versus paired-end reads for ChIP-seq

Paired-end sequencing has been widely used in DNA- and RNA-seq experiments to uncover fusion transcripts, genomic structural variations, rearrangements and new splice junctions, but the benefits of paired-end sequencing for regular ChIP-seq experiments are less clear. We first compared the percentage of the uniquely mapped paired-end reads that were also uniquely mapped when the paired-end reads were treated as if they were independent single-end reads at different read lengths. At 18-bp read length, we observed <10% uniquely mapped single-end reads and over 80% when the read length exceeded 22 bp (**Supplementary Fig. 2a** and **Supplementary Notes**).

The difference in sequencing coverage of repeat regions by uniquely mapped paired-end reads when they were mapped as either paired-end or single-end reads (36 bp) at a sequencing depth of

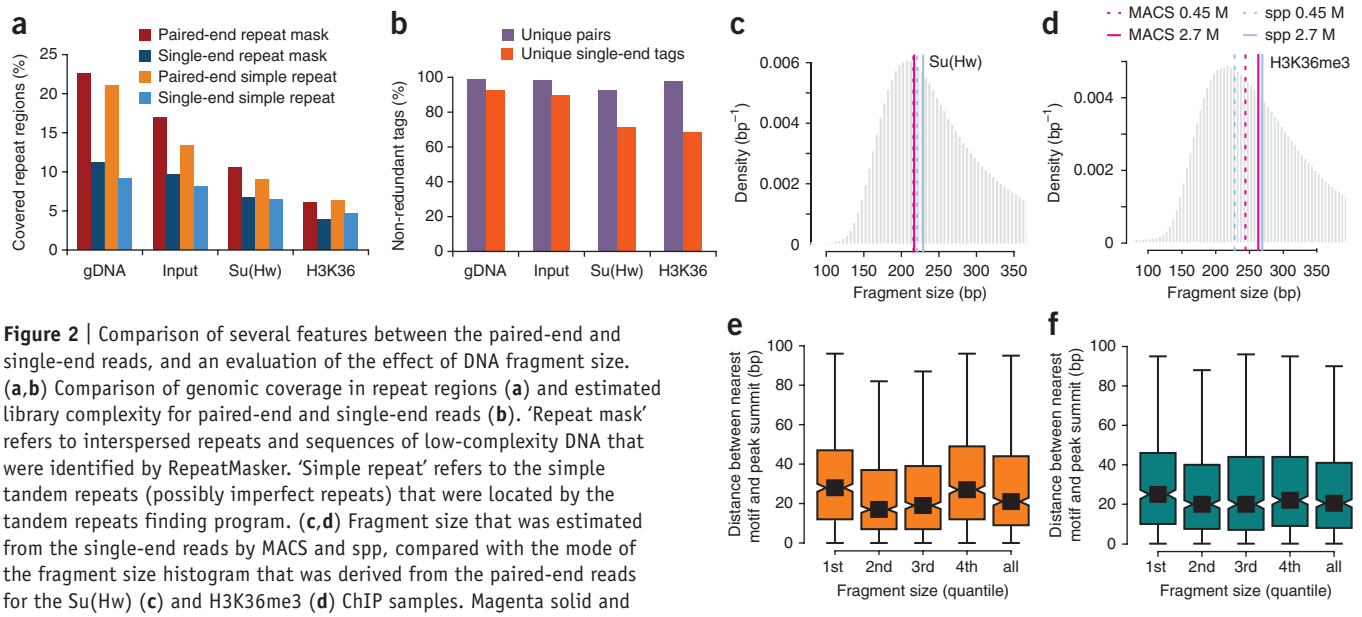

Figure 2 | Comparison of several features between the paired-end and single-end reads, and an evaluation of the effect of DNA fragment size. (**a**,**b**) Comparison of genomic coverage in repeat regions (**a**) and estimated library complexity for paired-end and single-end reads (**b**). 'Repeat mask' refers to interspersed repeats and sequences of low-complexity DNA that were identified by RepeatMasker. 'Simple repeat' refers to the simple tandem repeats (possibly imperfect repeats) that were located by the tandem repeats finding program. (**c**,**d**) Fragment size that was estimated from the single-end reads by MACS and spp, compared with the mode of the fragment size histogram that was derived from the paired-end reads for the Su(Hw) (**c**) and H3K36me3 (**d**) ChIP samples. Magenta solid and dashed lines represent the fragment size that was estimated from single-end reads by MACS (magenta) and spp (blue) at sequencing depths of 2.7 M and 0.45 M reads, respectively. (**e**,**f**) Box-plot comparisons of the summit resolution of the peaks identified by MACS (**e**) and spp (**f**) for the cases in which paired-end reads from DNA fragments with different sizes were used.

16.2 million (M) reads was approximately twice that of the single-end reads for the gDNA sample. This sequencing depth approximately corresponds to 327 M reads for the mappable human genome[10]. In contrast, for the chromatin input sample and for the ChIP samples of Su(Hw) and H3K36me3, the difference in sequencing coverage of the repeat regions between paired-end and single-end reads was less pronounced (**Fig. 2a**). The gain from paired-end data in discovering Su(Hw) or H3K36me3-enriched regions in repeat regions was typically less than 15% (**Supplementary Fig. 2b**).

A common quality measure for ChIP-seq libraries is library complexity (Online Methods). There are many factors that can lead to poor library quality, such as poor antibody quality, over-cross-linking, an insufficient amount of starting material of ChIP DNA, inappropriate sonication and overamplification by PCR. We observed a major discrepancy between the paired-end and single-end read data–based estimates of library complexity for Su(Hw) and H3K36me3 ChIP samples but not for the gDNA and chromatin input samples at a sequencing depth of 16.2 M paired-end reads (**Fig. 2b**). Therefore, caution is warranted when using single-end ChIP-seq data to model library complexity.

Choosing ChIP-seq data analysis algorithms for evaluation

There are more than 30 published algorithms for identifying peaks from ChIP-seq data, with more being published continuously[22,23]. We selected seven algorithms[7,10,24–28] that can use chromatin input data[12,13], are not restricted to analysis of only transcription factors or histone marks, directly support analyzing ChIP-seq data from *Drosophila* and are among the most highly cited. We evaluated their performance at different sequencing depths (Online Methods).

Size of sequenced fragments and peak calling

Model-based analysis of ChIP-seq (MACS)[7] and spp[26] explicitly report the estimated sizes of the DNA fragments in the library from the single-end data. For Su(Hw) ChIP-seq data, MACS and spp gave notably accurate size estimations that deviated from the paired-end data–inferred fragment size by only 10–20 bp (**Fig. 2c**). However, both algorithms were less accurate for the H3K36me3 data set (**Fig. 2d**). To characterize the influence of the fragment size on the spatial resolution of the narrow peaks, we next used paired-end reads from different size fragments in the same library for peak calling. For both MACS and spp, the larger the fragment size was, the wider the peak was (**Supplementary Fig. 3**). In contrast, peak-summit resolution did not depend strongly on the size of sequencing fragments. Thus, the use of smaller fragments did not necessarily improve the peak-summit resolution (**Fig. 2e**,**f**).

Sensitivity and specificity

We used the enrichment of the Su(Hw) binding motif within peaks to evaluate algorithm specificity. Site identification from short sequence reads (SISSRs)[25], MACS and Useq[27] were the best-performing algorithms in terms of specificity, and they showed a notable improvement with increasing sequencing depth (**Fig. 3**). At most sequencing depths, SISSRs had the best overall specificity for all of the identified peaks but identified fewer peaks than the other algorithms. To evaluate the overall sensitivity of the different algorithms for all identified peaks, we used the confidently enriched regions that were identified from ChIP-chip analysis as a proxy for true positives (Online Methods). Useq and spp had the highest overall sensitivity (**Supplementary Fig. 4**).

Effect of imbalanced coverage between ChIP and input

We evaluated how imbalanced sequencing coverage between ChIP and chromatin input samples influenced peak calling by MACS and Useq. We excluded SISSRs from this evaluation because the number of peaks it identified differed substantially between replicates when sequencing coverage was unbalanced. For the same sequencing depth (both small and large)

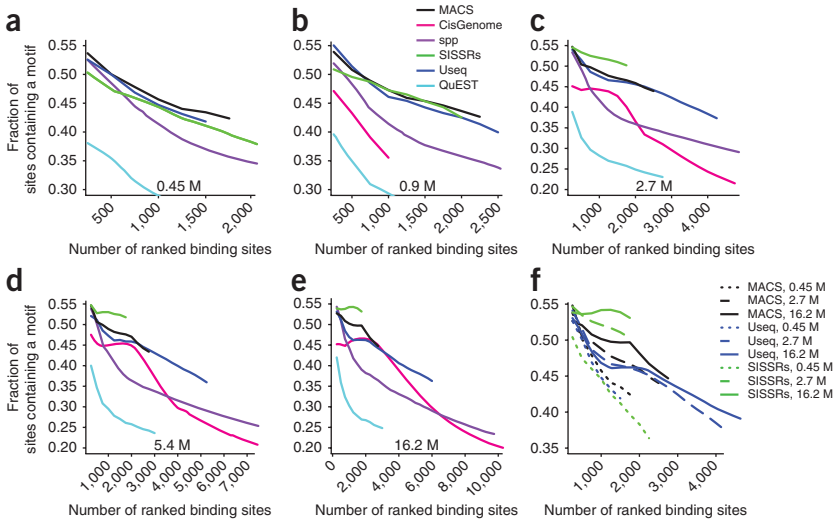

Figure 3 | Quality of Su(Hw) peaks. (**a**–**e**) Fraction Su(Hw) peaks, identified by peak callers indicated in **b**, that contains a Su(Hw) binding motif, plotted as a function of the number of top-ranked binding sites at the sequencing depths of 0.45 M (**a**), 0.9 M (**b**) 2.7 M (**c**), 5.4 M (**d**) and 16.2 M (**e**) reads. The ranking is based on the significance of each peak that was assigned by an individual algorithm. (**f**) Evaluation results for the top three best-performing peak-callers at sequencing depths of 0.45 M, 2.7 M and 16.2 M reads.

of the ChIP sample, deeper sequencing of the chromatin input sample gave rise to better detection specificity (**Supplementary Fig. 5a**,**b**). Therefore it is beneficial to sequence the chromatin input sample to a depth at least equal to the ChIP sample, if not deeper.

Effect of redundant reads on narrow peak calling

Redundant reads in ChIP-seq data sets often indicate poor library complexity, and as a result, many peak callers remove redundant reads, that is, reads with the same 5′ genomic location, during peak calling. However, with very deep sequencing, redundant single-end reads from ChIP samples may also result from ChIP-enrichment signal. We therefore evaluated the effect of retaining or removing redundant reads on both the sensitivity and specificity at large sequencing depths. We first compared enrichment of the Su(Hw) motif in peaks that were identified by SISSRs, MACS and Useq between two conditions: one in which we kept only one read at each genomic location and another in which we retained the redundant reads completely (SISSRs and Useq) or partially (MACS). In general, removing redundant reads improved the specificity of identified peaks for MACS and Useq. For SISSRs, removing redundant reads only improved the specificity at a high sequencing depth (**Supplementary Fig. 5c**–**e**).

The paired-end data allowed us to differentiate the source of redundant reads in peak regions because redundant reads that result from experimental artifacts, such as PCR-amplification bias, should be identical at both ends. The data indicated that redundant reads from duplicate fragments represented fewer than 10% of all reads, whereas in most peak regions, the proportion of redundant reads was 20–40% (**Supplementary Fig. 6**). Thus, most redundant reads in peak regions represented true signal. Nonetheless, for MACS and Useq removal of redundant reads had little or no effect on sensitivity (**Supplementary Fig. 7**). Overall, removal of redundant reads was usually beneficial because it removed noise in the non-enriched regions but had little effect on detection sensitivity.

Dependence of detection on sequencing depth

One indication that a sufficient sequencing depth has been reached is when the number of binding sites plateaus with an increasing read count. Different algorithms had distinct saturation profiles. The number of binding sites identified by MACS, SISSRs and QuEST[28] started to plateau or plateaued at ~16.2 M reads (corresponding to ~327 M reads in human), whereas the number of peaks identified by CisGenome[24], spp and Useq steadily increased with sequencing depth (**Supplementary Figs. 8** and **9a**). We then compared the enriched regions discovered at a given depth to those identified from the complete set of 120 M reads. Most algorithms detected more than half of the Su(Hw)-enriched regions identified from the complete data at a sequencing depth of 5.4 M reads (corresponding to ~110 M reads in human). MACS and Useq identified more than 60% of the Su(Hw)-enriched regions (threefold enrichment or greater) identified from the complete data at a depth of 2.7 M reads (corresponding to ~55 M reads in human; **Supplementary Fig. 10a**).

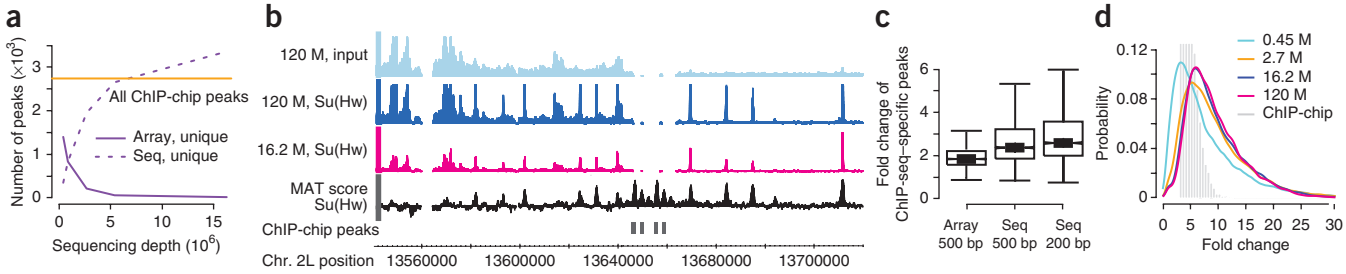

Figure 4 | Comparison of identified narrow peaks and the dynamic range between the sequencing and the tiling array platform. (**a**) Number of identified peaks on the sequencing platform (seq) and the tiling array platform (array). (**b**) Examples of ChIP-chip peaks that were missed in the sequencing platform, the MAT score for ChIP-chip data, and the ChIP-seq signal coverage at the sequencing depths of 16.2 M and 120 M reads. (**c**) Fold change difference between sequencing and tiling arrays in 200-bp and 500-bp windows centered on the peaks that were unique to the sequencing platform at a sequencing depth of 16.2 M. (**d**) Dynamic range of the signal (ChIP versus the chromatin input fold change) for the sequencing and the tiling array platform.

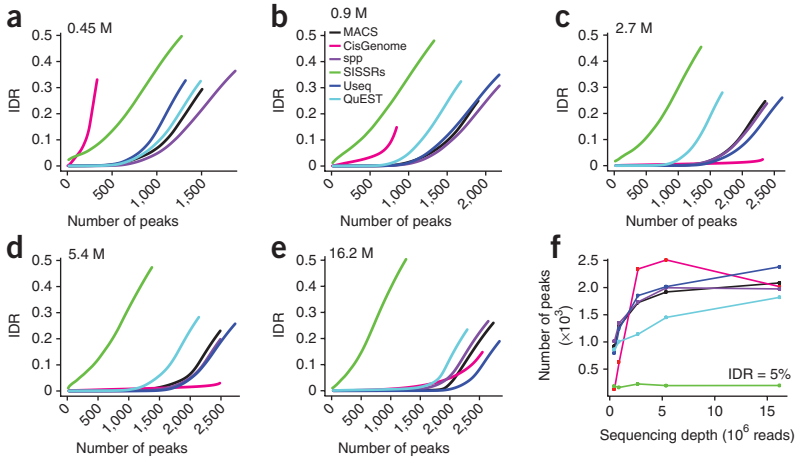

Figure 5 | Evaluation of reproducibility across replicates for six peak callers. (**a**–**e**) The number of reproducible peaks at various IDR levels is plotted for sequencing depths of 0.45 M (**a**), 0.9 M (**b**), 2.7 M (**c**), 5.4 M (**d**) and 16.2 M (**e**) reads. (**f**) Number of reproducible peaks identified at an IDR of 5% is plotted as a function of sequencing depth.

Narrow peak differences between sequencing and array data

We compared ChIP-enriched regions identified by tiling arrays (Affymetrix) and sequencing platforms on the same set of Su(Hw) samples using Useq and model-based analysis of tiling arrays (MAT)[29], respectively. At low sequencing depths (≤0.90 M reads, corresponding to ≤~18 M reads in human), ~30–50% of the ChIP-chip peaks were missed by ChIP-seq. When the sequencing depth reached 2.7 M reads (corresponding to ~55 M in human), over 90% of ChIP-chip peaks were identified by ChIP-seq (**Fig. 4a**). Notably, even when the sequencing depth reached 16.2 M reads (corresponding to ~327 M reads in human), ~1% of the ChIP-chip peaks were not detected in the sequencing data. These peaks had sparse or no sequencing coverage, even using all reads in our data set (**Fig. 4b**), mostly owing to low mappability (**Supplementary Notes**)[10,23]. Su(Hw) peaks specific to ChIP-chip data were enriched of the Su(Hw) binding motif, suggesting that they were genuine Su(Hw) binding sites. We performed ChIP–quantitative (q)PCR experiments and validated seven randomly selected ChIP-chip peaks that were missed in the sequencing data (**Supplementary Methods** and **Supplementary Fig. 11**). Either a lack of probe coverage on the array or higher sensitivity of ChIP-seq relative to array[5] (**Fig. 4c**) contributed to those ChIP-seq–specific peaks. The sequencing platform showed a larger dynamic range of fold change than the array, and increased sequencing depth improved both sensitivity and quantification of regions with low fold enrichment (**Fig. 4d**).

Algorithm reproducibility for narrow peaks across replicates

We quantified the reproducibility of peak calling across replicates using the irreproducible discovery rate (IDR)[30], which assesses the consistency between the ranks of the peaks that were commonly identified on a pair of replicates. We found that the relative reproducibility of different algorithms depended on sequencing depth. Whereas MACS and spp were more reproducible across replicates than any other algorithms at shallow sequencing depths (fewer than 0.90 M reads, corresponding to fewer than 18 M reads in human), CisGenome and Useq became the most 'reproducible' across replicates at or above 2.7 M reads (corresponding to ~55 M reads in human) (**Fig. 5**).

Detecting broad enriched regions at different depths

We evaluated sensitivity and specificity of different algorithms in detecting broad patterns of enrichment. As a 'gold standard' for H3K36me3-enriched (H3K36me3-positive) regions, we used all exonic regions from genes with the top 4,000 expression levels (**Supplementary Fig. 12**; results were similar with the top 1,000 or 2,000 genes, **Supplementary Figs. 13** and **14**). We called the gene bodies of unexpressed genes non-enriched H3K36me3 (H3K36me3-negative) regions (Online Methods). To control for the width differences of the enriched regions that were identified by different algorithms, we used the coverage of the 'true positives' normalized by the total width of all predicted enriched regions as a proxy for the sensitivity and the width-normalized coverage of the 'true negatives' as a proxy for the false positive rate. QuEST had the highest specificity, with PeakSeq second. Spp was among the algorithms with the lowest sensitivity and specificity at a shallower sequencing depth (fewer than 0.90 M reads, corresponding to fewer than ~18 M reads in human), but showed distinct improvement at larger sequencing depths (≥2.7 M reads, corresponding to ≥~55 M reads in human; **Supplementary Fig. 12**).

Similar to the case of Su(Hw), different algorithms showed distinct saturation profiles in the broad enrichment data (**Supplementary Figs. 9b** and **15**). MACS, QuEST, spp and Useq showed a faster saturation in identifying ChIP-enriched regions of H3K36me3 than Su(Hw) (**Supplementary Fig. 10b**). Unlike the case of Su(Hw), the number of identified H3K36me3-enriched regions did not increase monotonically with the sequencing depth for many algorithms, including MACS, spp and QuEST, because neighboring regions started to merge at high sequencing depths.

Algorithm reproducibility for broad regions across replicates

We performed IDR analysis tailored to the broad regions (Online Methods) to evaluate the algorithm reproducibility across replicates. Again, we found that this depended on sequencing depth. QuEST and Useq produced more reproducible regions across replicates than did other algorithms at or below 2.7 M reads (corresponding to ~55 M reads in human), whereas spp and Useq did so when the sequencing depth was above 2.7 M reads (**Supplementary Fig. 16**).

DISCUSSION

Sequencing depth had a profound impact on several aspects of ChIP-seq results, including some that were unexpected. Our study suggests that for such transcription factors as Su(Hw) and such histone marks as H3K36me3, the regularly adopted sequencing depth of 15–20 M reads in humans may be insufficient to identify the vast majority of enriched regions.

Our finding that the removal of redundant reads helped to reduce false positives and had little effect on the detection sensitivity is seemingly incompatible with the fact that at a high sequencing depth, most redundant reads in narrow peak regions represent true signals. The probable explanation is that most of the regions containing redundant reads were among the more highly enriched, such that even after the removal of

redundant reads, the vast majority of those regions still showed significant enrichment in ChIP versus chromatin input samples. Additionally, the removal of redundant reads in the non-enriched regions differentially reduced the amount of reads originating from experimental bias in PCR amplification and library preparation. Because removing redundant reads influenced quantitative information associated with enriched regions, for high-quality libraries, it may be most appropriate to identify peaks in the absence of redundant reads, but then to include all reads in downstream analyses.

There were notable variations in sensitivity and specificity between the algorithms we evaluated. Some algorithms behaved unexpectedly at high sequencing depths, indicating the importance of improving algorithms for use at a high sequencing depth, including a more effective handling of reads mapped to multiple genomic locations. In practice, it is beneficial to use more than one algorithm to ensure robustness of the analysis results for the deep sequencing data.

One important factor that we did not assess here is the choice of sequencing platform. We chose the Illumina platform for this study because the vast majority of publicly available ChIP-seq data sets were generated on this platform, including those from the Encyclopedia of DNA elements (ENCODE) and modENCODE projects (http://www.genome.gov/10005107). There are also important open questions, specifically regarding identifying broadly enriched regions that we did not address in this study, such as how to determine the boundaries of broadly enriched regions and how sequencing depth influences boundary determination. We anticipate that our data set will be a valuable resource for the ChIP-seq community to address these and other technical questions related to deep sequencing.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Gene Expression Omnibus: GSE27679.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the authors of all of the algorithms that we evaluated in this study: H. Ji, R. Jothi, P. Kharchenko, W. Li, D. Nix, J. Rozowsky and A. Valouev. We thank N. Bild, D. Roqueiro and M. Sabala for help in performing PeakSeq on the Bionimbus Cloud, D. Schmidt and D. Odom for sharing their sequencing data of the ENCODE spike-in sample, A. Kundaje for sharing his unpublished results on IDR analysis of H3K36me3 in humans, N. Rashid for sharing the mappability data of *Drosophila* genome, M. Greenberg for support in the early stage of this project, and E. Birney, M. Snyder, J. Ahringer, M. Gerstein, M. Kellis, P. Park and other members of modENCODE consortium for helpful discussions. This work was partially funded by US National Institutes of Health (HG4069 to X.S.L., 3U01 HG004270-03S1 to X.S.L. and J.D.L., and U01HG004264 to K.P.W.).

AUTHOR CONTRIBUTIONS

Y.C. performed bioinformatic analysis. N.N. performed cell culture, ChIP experiments and library preparation with help from J.Z. J.O.M. performed library preparation and sequencing experiments. Q.L. and P.J.B. contributed code for the IDR method. Q.L. participated in writing the description of IDR method and interpretation of the IDR analysis result. M.S. performed ChIP–quantitative (q)PCR validation of the selected array-specific Su(Hw) peaks and analyzed the ChIP-qPCR data. T.L., Y.Z., T.-K.K., H.H.H., Y.R., R.M.M. and B.J.W. contributed to the early development of the project. B.J.W., K.P.W., J.D.L. and X.S.L. conceived the project. T.-K.K., H.H.H., Y.R. and R.M.M. performed pilot experiments. Y.C., J.D.L. and X.S.L. wrote the manuscript with the help from other authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.1985 . Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
2. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
3. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
4. Mikkelsen, T.S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
5. Johnson, D.S. *et al.* Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome Res.* **18**, 393–403 (2008).
6. Ho, J.W. *et al.* ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics* **12**, 134 (2011).
7. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
8. Laajala, T.D. *et al.* A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics* **10**, 618 (2009).
9. Wilbanks, E.G. & Facciotti, M.T. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* **5**, e11471 (2010).
10. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27**, 66–75 (2009).
11. Negre, N. *et al.* A comprehensive map of insulator elements for the *Drosophila* genome. *PLoS Genet.* **6**, e1000814 (2010).
12. Myers, R.M. *et al.* A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
13. Pepke, S., Wold, B. & Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* **6**, S22–S32 (2009).
14. Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.* **41**, 376–381 (2009).
15. Dohm, J.C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**, e105 (2008).
16. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods* **6**, 291–295 (2009).
17. Kharchenko, P.V. *et al.* Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**, 480–485 (2011).
18. Negre, N. *et al.* A cis-regulatory map of the *Drosophila* genome. *Nature* **471**, 527–531 (2011).
19. Roy, S. *et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
20. Larschan, E. *et al.* X chromosome dosage compensation via enhanced transcriptional elongation in *Drosophila*. *Nature* **471**, 115–118 (2011).
21. Teytelman, L. *et al.* Impact of chromatin structures on DNA processing for genomic analyses. *PLoS ONE* **4**, e6700 (2009).
22. Feng, X., Grossman, R. & Stein, L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. *BMC Bioinformatics* **12**, 139 (2011).
23. Rashid, N.U., Giresi, P.G., Ibrahim, J.G., Sun, W. & Lieb, J.D. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.* **12**, R67 (2011).
24. Ji, H. *et al.* An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* **26**, 1293–1300 (2008).
25. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.* **36**, 5221–5231 (2008).
26. Kharchenko, P.V., Tolstorukov, M.Y. & Park, P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).
27. Nix, D.A., Courdy, S.J. & Boucher, K.M. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* **9**, 523 (2008).
28. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **5**, 829–834 (2008).
29. Johnson, W.E. *et al.* Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl. Acad. Sci. USA* **103**, 12457–12462 (2006).
30. Li, Q., Brown, J.B., Huang, H. & Bickel, P.J. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics* **5**, 1752–1779 (2011).

ONLINE METHODS

Overall experimental design. We amplified S2 cells from the modENCODE batch before transferring them to the plates. We put a total of 15 plates in culture until the cells reached the appropriate concentration. We collected cells from two plates to extract the gDNA. We subsequently treated the remaining 13 plates with formaldehyde at the same time and then extracted the chromatin. Next, we used three plates to produce a chromatin input sample corresponding to the chromatin DNA of the treated cells. Each plate corresponded to an Eppendorf tube of chromatin, and we performed ChIP experiments on five tubes for H3K36me3 and on five tubes for Su(Hw). We performed ChIP and DNA extraction procedures independently on each tube. After the purification of the DNA, we pooled together the tubes corresponding to the same sample (Su(Hw), H3K36me3, chromatin input and gDNA). After mixing, we again aliquoted each sample into five tubes. Next, we used one aliquot of each for ChIP-chip on Affymetrix tiling arrays for quality control, one aliquot for single-end (SE) sequencing at the high-throughput genomic analysis core of the University of Chicago, one aliquot for SE sequencing at the high-throughput sequencing facility of the University of North Carolina at Chapel Hill (UNC), 1 aliquot for paired-end (PE) sequencing at UNC and the remaining aliquot to back up the original samples.

ChIP experiment. After the expansion of the S2 cell line, we transferred cells to a plate. Once cells were confluent, we added 1.8% formaldehyde to the cell culture. We collected cells in the presence of the formaldehyde with the help of a cell scraper. After 15 min of incubation at room temperature, we quenched the cross-linking reaction with glycine for 5 min. We subsequently washed cell pellets 3 times with a lysis buffer. We performed regular chromatin extraction before sonication. We then used sonicated chromatin for ChIP experiments or directly for DNA extraction for the chromatin input samples. We performed ChIP as previously described[11,18]. The anti-H3K36me3 antibody was from Abcam (ab9050, lot 927884). The anti-Su(Hw) antibody was from P.K. Geyer's lab. Both are rabbit polyclonal antibodies.

Library preparation and sequencing. At the University of Chicago, we prepared the libraries according to Illumina's instructions accompanying the DNA Sample kit (0801-0303). Briefly, we end-repaired DNA using a combination of T4 DNA polymerase, *Escherichia coli* DNA polymerase I large fragment (Klenow polymerase) and T4 polynucleotide kinase. We treated the blunt, phosphorylated ends with Klenow fragment (3′ to 5′ exo$^-$) and dATP to yield a protruding 3′-end adenine ('A') base for ligation of Illumina's adaptors, which have a single thymine ('T') base overhang at the 3′ end. After adaptor ligation, we PCR-amplified DNA with Illumina primers for 15 cycles, and isolated bands of library fragments of ~250 bp from a 2% agarose gel. We captured the purified DNA on an Illumina flow cell for cluster generation. We sequenced libraries on the Genome Analyzer IIx following the manufacturer's protocols. At UNC, we used a slightly different protocol of library preparation in which we isolated bands of the library fragments of ~150–500 bp immediately after the ligation of Illumina's adaptors followed by 18 cycles of PCR amplification.

The definition of library complexity for paired-end and single-end data. The library complexity of SE data was defined as the number of nonredundant SE reads divided by the total number of reads, where redundant SE reads are those that are mapped to the same location with the same orientation in the genome. The library complexity of PE data was defined as a nonredundant pair of reads divided by a total pair of reads, where redundant PE reads are those that have identical genomic locations on both ends.

The mapping of sequencing reads. We used ELAND to align the SE sequencing reads to the Flybase BDGPv5 reference genome. We pooled together the uniquely mapped SE reads with no more than two mismatches from different runs up to 120 M for the ChIP-sample of Su(Hw) and H3K36me3. For chromatin input and gDNA samples, we added uniquely mapped PE reads to constitute the total 120 M reads, thereby compensating for the failed runs of SE sequencing of these two samples (**Supplementary Table 1**). To compare differences in the read mappability and coverage of the repeats region, we estimated the library complexity for PE and SE reads. We first used Bowtie 0.12.5 to map the PE reads with almost all of the default settings (-chunkmbs 120), and we constrained the fragment size to 80–600 bp. Next, we re-aligned the uniquely mapped PE reads in the SE mode using the same parameter settings.

The mappability, the heterochromatin and the repeat regions of the *Drosophila* genome. We obtained mappability data of *Drosophila* genome from a previous study[23] and the details of how the mappability was calculated have been described previously[10]. 'Repeat mask' refers to the interspersed repeats and the DNA sequences of low complexity that were identified by RepeatMasker. 'Simple repeat' refers to the simple tandem repeats (possibly imperfect repeats) that were identified by the Tandem Repeats Finding program[31]. Both the repeat region and the heterochromatin region annotation were based on UCSC dm3 and were downloaded from the UCSC genome browser.

ChIP-seq data analysis algorithms. The algorithms we evaluated were CisGenome (v1.2), MACS (v1.40beta), spp (v1.8), QuEST (v2.4), Useq (v6.9), SISSRS (v1.4) and PeakSeq (v1.01). We did not evaluate E-RANGE[32] and F-Seq[33], two highly cited algorithms, because the parameters of E-RANGE were optimized for the mammalian genome and F-Seq does not provide good support for peak finding in invertebrates. For evaluation of algorithm performance on ChIP-seq data of Su(Hw), we did not include PeakSeq[10] because it does not provide the information of the peak summit of each peak, whereas the evaluation of peak quality requires the information of the peak summit. When we used the middle point of each peak identified by PeakSeq as a surrogate of the peak summit, PeakSeq performed poorly, thereby making the comparison unfair. For the evaluation of the algorithm performance on the ChIP-seq data of H3k36me3, we did not include SISSRs because the width of the regions identified by SISSRs is too small to be consistent with that of the broad regions.

ChIP-seq data analysis at different sequencing depths. We randomly sampled reads at sequencing depths of 0.45 M, 0.9 M, 2.7 M, 5.4 M and 16.2 M reads from a pool of 120 M reads for both ChIP and chromatin input samples. These sequencing depths approximately correspond to 9 M, 18 M, 55 M, 109 M and 327 M reads in a human ChIP-seq experiment[10]. At each sampling depth,

we generated five independent 'replicates' of the sequencing data. We averaged the analysis results of each algorithm over the five replicates before comparison.

ChIP-chip peak calling for Su(Hw) and the fold change calculation for the sequencing and tiling array platform. We performed peak calling for Su(Hw) using the MAT[29] algorithm, which is among the best peak-calling algorithms for ChIP-chip data from Affymetrix data[5] with a band width of 250 bp, a *P*-value cutoff of 10^{-5} and a false discovery rate cutoff of 5%. We only considered the peaks with fold changes of no less than threefold (the detection limit of the Affymetrix tiling array[5]) for further comparison with ChIP-seq peaks. All 500-bp windows that centered on the summit of these ChIP-chip peaks of Su(Hw) were used as reference ChIP-chip peaks and as a proxy for the true positives to evaluate the sensitivity of different algorithms for identifying ChIP-seq peaks. The fold change of the signal (ChIP versus chromatin input) in each 500 bp-scanning window centered on the probes was calculated using Tiling Analysis software (Affymetrix). The fold change in 200-bp and 500-bp windows that were centered on the summit of ChIP-seq peaks was calculated as follows: $(\text{number of covered fragments} + 1)_{\text{ChIP}}$ / $(\text{number of covered fragments} + 1)_{\text{input}}$.

The use of IDR to quantify reproducibility of peak calling between a pair of replicates. The reproducibility across replicates is essential to ChIP experiments not only at the level of read count data but also at the level of peak calling because the identified peaks usually are the primary substrates for downstream analysis. IDR is a statistical measure that assesses the consistency of the rank orders between a pair of rank lists[30]. Unlike the usual scalar measures of reproducibility (for example, the rank correlation), this measure describes reproducibility in terms of the extent to which the ranks of the entries on the lists are no longer consistent across replicates that are ordered in descending significance. Based on a copula mixture model, this measure provides a 'score' that estimates the probability that each pair of peaks is reproducible, and it reports the expected rate of irreproducible discoveries in the selected peaks in a fashion analogous to that of false discovery rate. The number of reproducible peaks across replicates at given IDR levels can be used to compare the relative reproducibility of different peak-calling algorithms.

IDR is independent of the threshold choice that is used for peak calling, and it emphasizes implicitly the consistency between the top-ranked peaks, rather than treating all of the ranks equally. Therefore, this method overcomes many limitations in traditional ways of measuring reproducibility and is suitable for our purposes. Detailed descriptions of the methodology and the implementation of IDR for narrow peak can be found in ref. 30. For broad peaks, often one peak overlaps with multiple (small) peaks on the other replicate. When this occurs, all these small peaks are lumped as one peak and the most substantial significance of these small peaks is used as the significance of the lumped peak (A. Kundaje, personal communication).

Because the number of identified peaks of some algorithms is much larger than others, we evaluated the top 3,000 significant peaks from all of the identified peaks. We used the R package in ref. 30 for all of the IDR analyses.

The RNA-seq and H3K4me3 ChIP-chip data. We calculated the gene expression summarized as the reads per kilobase per million reads (RPKM) value based on the RNA-seq data from a previous study[34]. The H3K4me3 ChIP-chip data were generated, and the ChIP-enriched regions were identified, by members of the White laboratory. We considered the 2-kb transcription start site–centered promoter to have H3K4me3 enrichment if it overlapped with the H3K4me3 peaks.

The data for enriched and depleted regions of various histone marks. The enriched and depleted regions of 15 histone marks that were identified from ChIP-chip data were obtained from modENCODE[17–19,35]. These histone marks include H3K18ac, H3K27ac, H3K27me3, H3K36me1, H3K36me3, H3K4Me3, H3K4me1, H3K4me2, H3K79Me1, H3K79me2, H3K9ac, H3K9me3, H4K16ac, H4K5ac and H4K8ac.

Motif-enrichment analysis. We mapped the position-specific weight matrix (PSWM) of Su(Hw) onto the genome of *Drosophila* (dm3) using CisGenome with a third-order Markov background model. We calculated the distance between the nearest mapped motif and the peak summit using a custom Perl script.

The definition of positive and negative regions for H3K36me3 and the measurement of sensitivity and the false positive rate. H3K36me3 is highly enriched in exonic regions but not in intronic regions of actively transcribed genes[14], which allowed us to approximate the positive and negative regions of H3K36me3 in the genome and to estimate the sensitivity and specificity of different algorithms using these predefined regions. We defined the 'positive' regions of H3K36me3 as the exons of the top 4,000 expressed genes (evaluation results were similar for the top 1,000 genes or top 2,000 genes) and the negative regions as the gene body of the non-expressed genes. We estimated the expression of each annotated gene based on the RNA-seq data from the S2 cells as the total number of reads of all of the unique exons per kilobase of total length of unique exons per million mapped reads (RPKM)[34]. We averaged the RPKM value of each gene over two biological replicates. We used the same criterion to define the non-expressed genes as in the previous study, where the non-expressed genes were those with the number of unique mapped reads per kilobase per million mapped reads (RPKM) smaller than or equal to 4 (ref. 34). This cutoff was chosen based on the distribution of RPKM values in intergenic regions, where the probability of observing an RPKM value greater than or equal to 4 is ~5%. To control for the difference in peak width among algorithms, we used the coverage of the positive regions normalized by the total width of all predicted enriched regions as a proxy for the sensitivity and the width-normalized coverage of the negative regions as a proxy for the false positive rate.

The calculation of the G+C composition, the read-count ratio and the coverage over different genomic features. We calculated the window-based G+C composition (window size of 36 bp, the same as read length) across the genome using the hgGcPercent program from J. Kent (UCSC). We calculated the G+C composition of sequencing reads of different samples using a custom Perl script. We calculated the window-based read-count ratio and the read coverage over different genomic features, including

doi:10.1038/nmeth.1985

exons, gene bodies and repeat regions using the combination of custom Perl scripts and BEDTools[36].

Statistical analysis. We performed all of the statistical analyses in R, and showed all of the *P* values smaller than 2.2×10^{-16} as $P < 2.2 \times 10^{-16}$, which is the default cutoff in R.

31. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
32. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
33. Boyle, A.P., Guinney, J., Crawford, G.E. & Furey, T.S. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537–2538 (2008).
34. Zhang, Y. *et al.* Expression in aneuploid *Drosophila* S2 cells. *PLoS Biol.* **8**, e1000320 (2010).
35. Celniker, S.E. *et al.* Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
36. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).