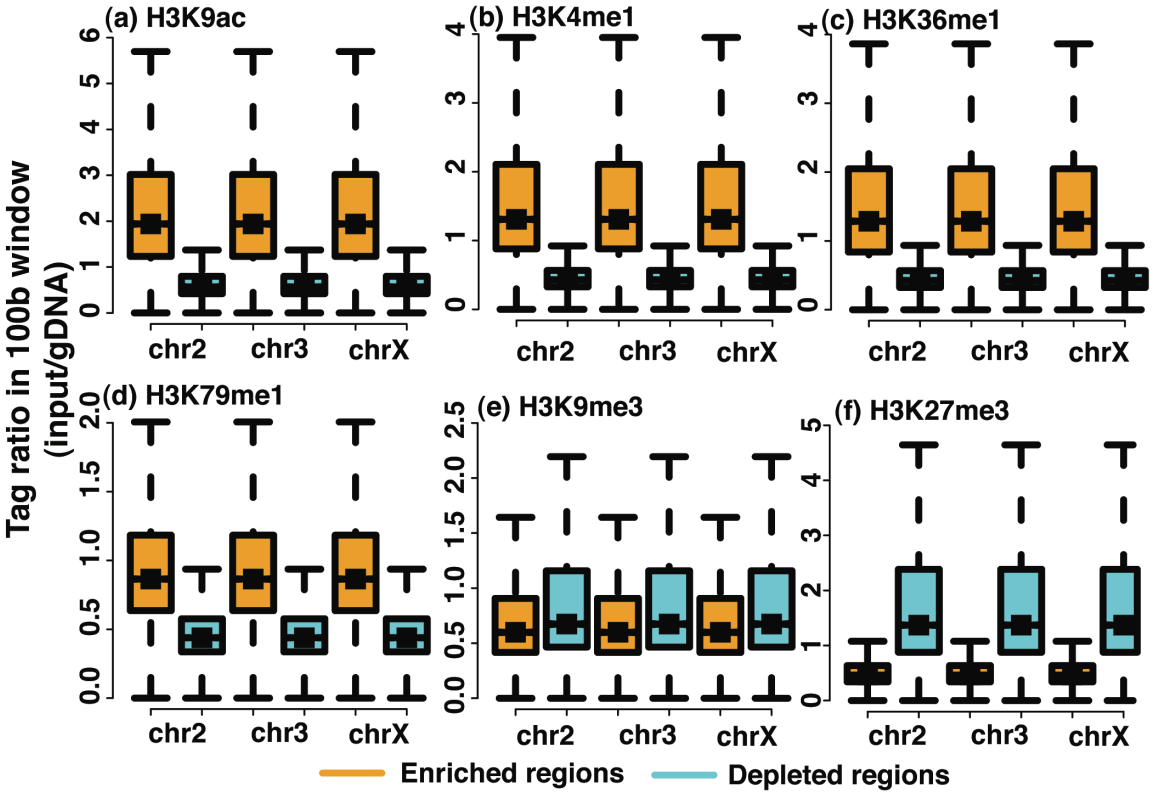
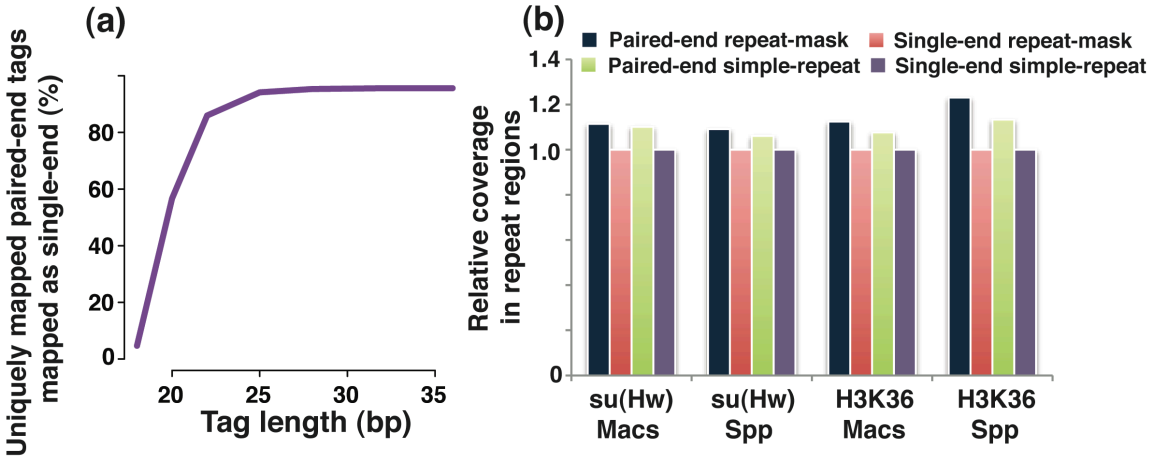
Supplementary material

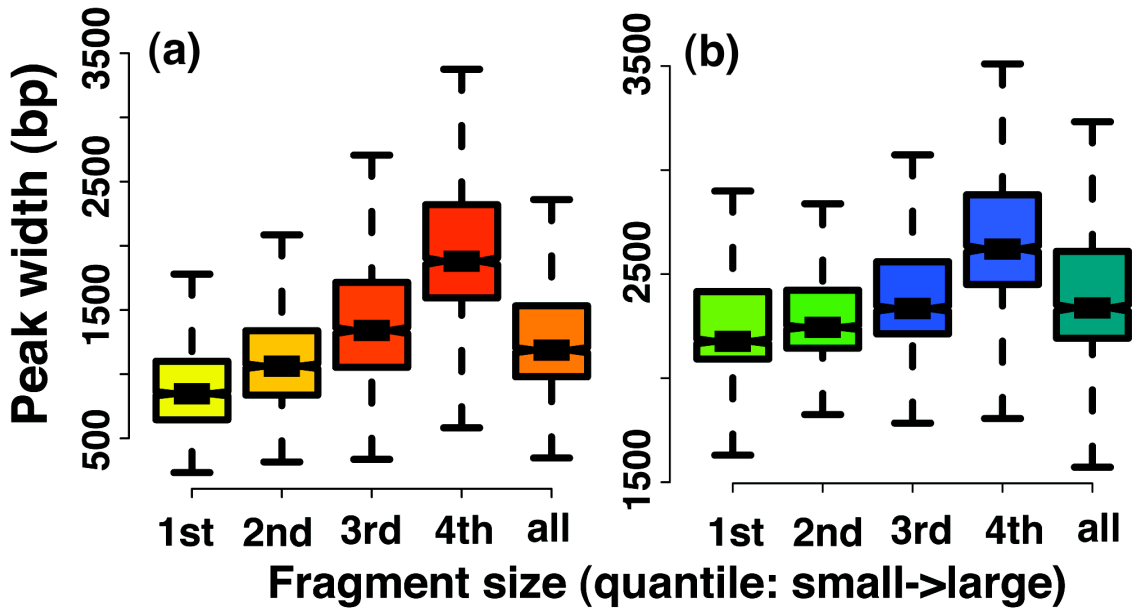
Supplementary Figure 1. The boxplots of read count ratio of chromatin input to gDNA sample are shown for (a) H3K9ac, (b) H3K4me1, (c) H3K36me1, (d)H3K79me1, (e) H3K9me3, and (f) H3K27me3.



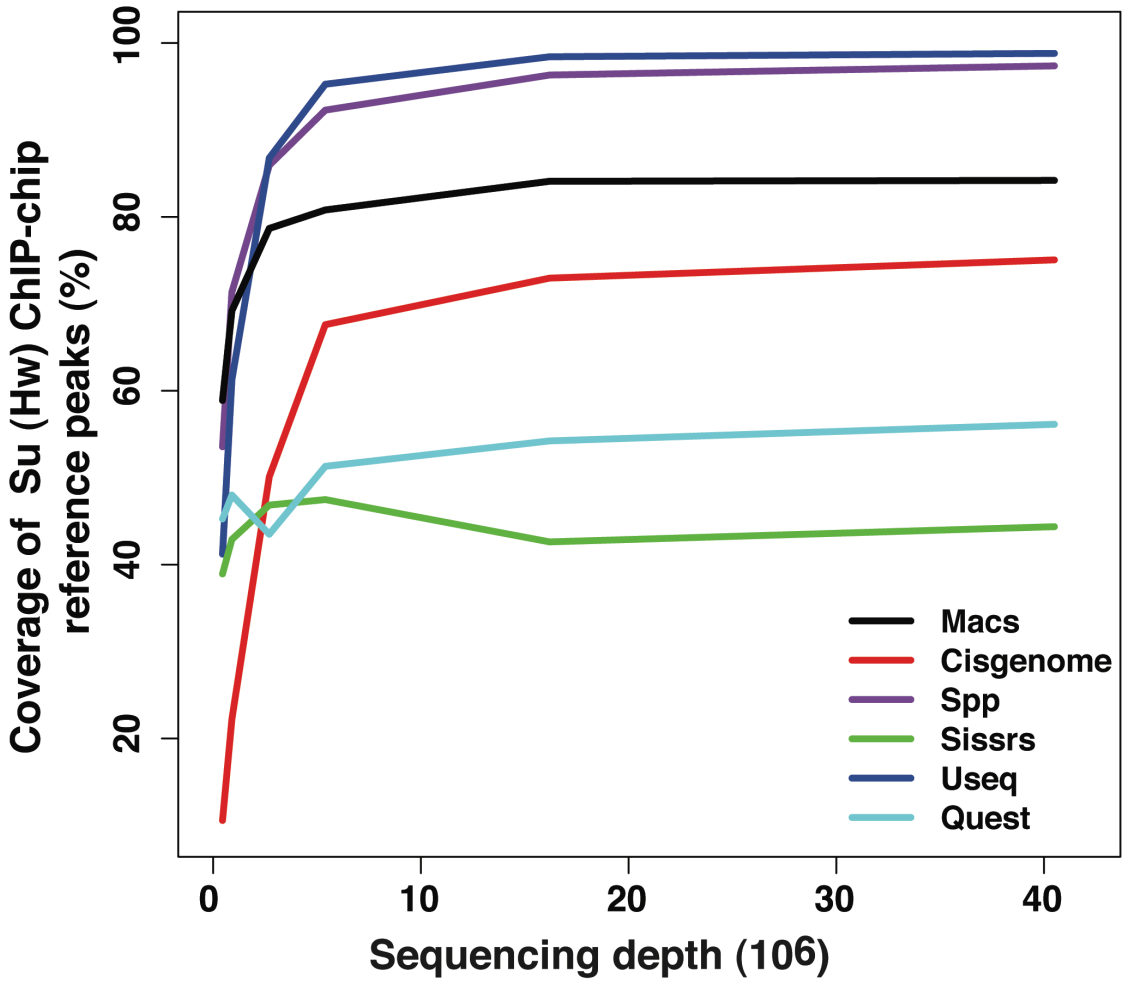
Supplementary Figure 2. A comparison of the difference in (a) the number of uniquely mapped reads between PE and SE reads, and (b) the difference in coverage of repeat regions by the Su(Hw) and H3K36me3 peaks that were identified by Macs and Spp is shown. The relative coverage of repeat regions by SE read is set to one for both algorithms.



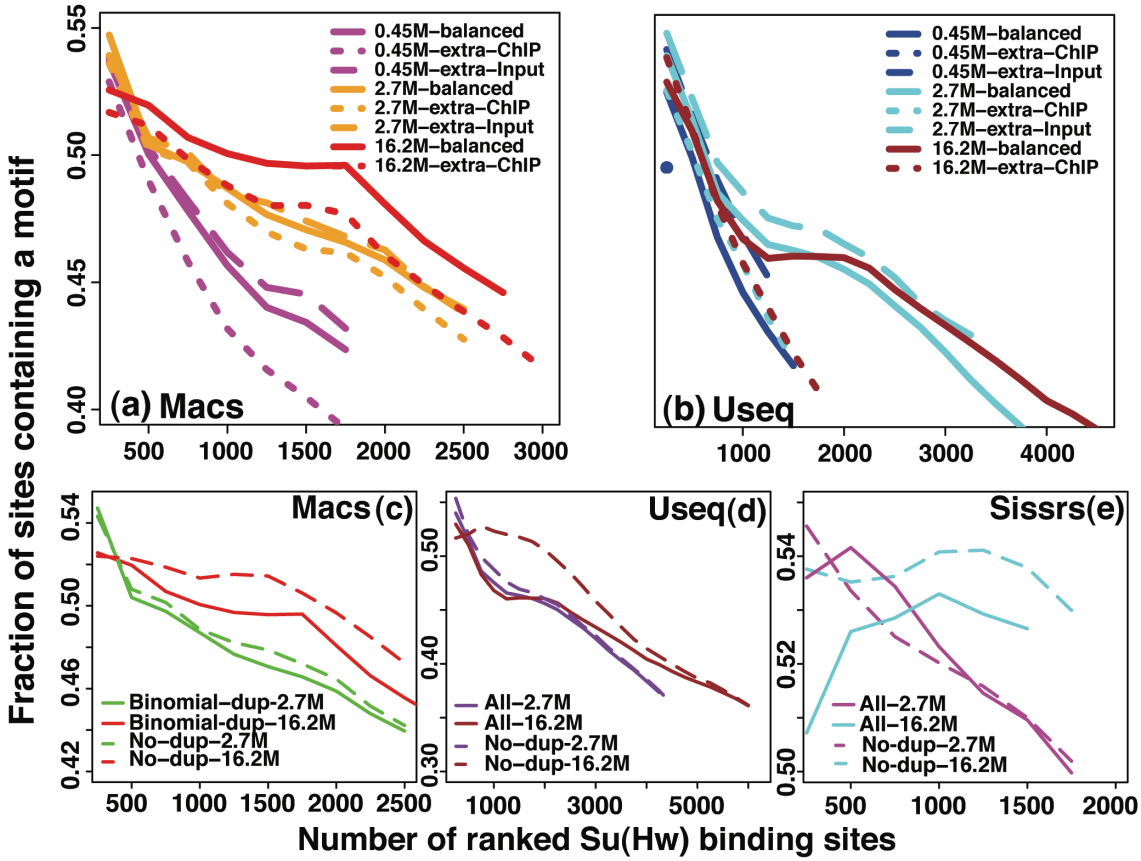
Supplementary Figure 3. The box-plot of the length of the peaks identified by Macs (a) and Spp (b) is shown across cases where paired-end reads from DNA fragments with different sizes were used.



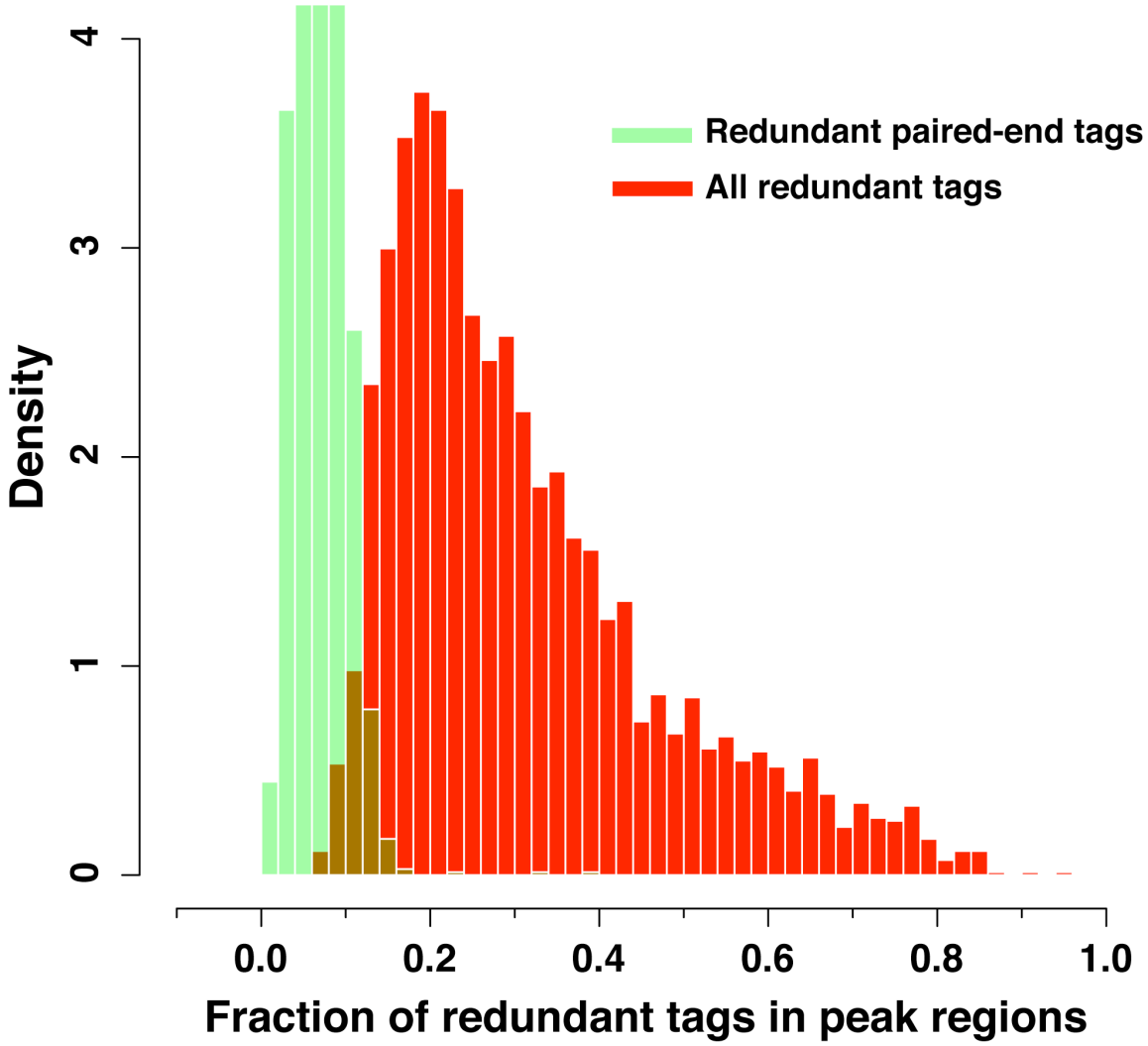
Supplementary Figure 4. The coverage of Su(Hw) ChIP-chip reference peaks by ChIP-seq peaks that were identified by Macs, Cisgenome, Spp, Sissrs, Useq and Quest at different sequencing depths is shown.



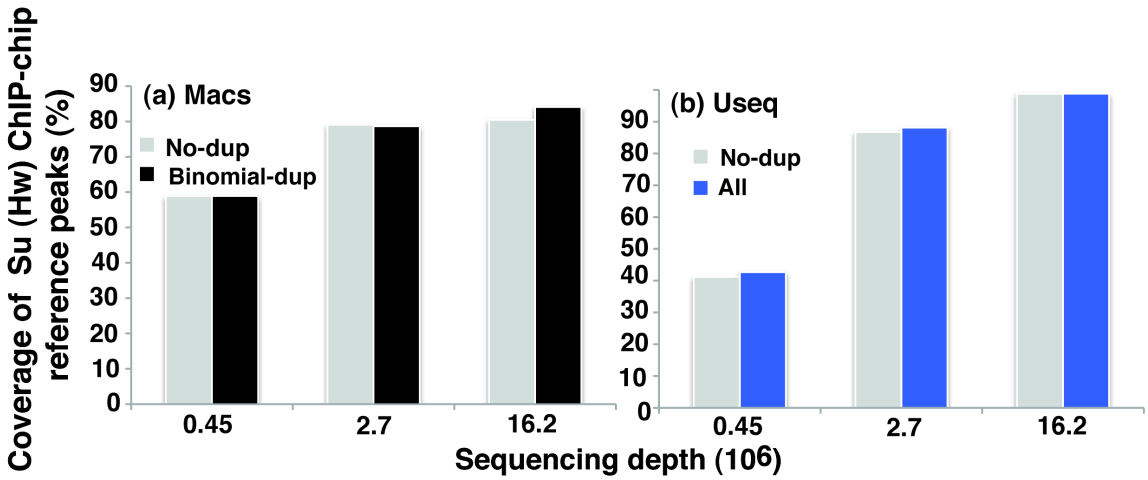
Supplementary Figure 5. The effect of sequencing depth imbalance between ChIP and chromatin input samples on the quality of the Su(Hw) peaks identified by Macs (a) and Useq (b) is shown. The term "extra-ChIP" refers to a condition in which the sequencing depth of the ChIP sample is larger than that of the chromatin input sample, and "extra-Input" indicates the opposite condition. For illustration purposes, only the cases in which the depth of the chromatin input sample is 6 times as large as that of the ChIP sample, or vice versa, are shown for the ChIP-sample sequencing depths of 0.45 and 2.7 M reads. When the difference in the sequencing depth is smaller, the difference in the peak quality is more subtle. At a sequencing depth of 16.2 M reads for the ChIP-sample, the "extra-ChIP" corresponds to a condition in which the sequencing depth of the ChIP sample is 36 times as large as that of the chromatin input sample. The effect of removing redundant reads on the quality of the Su(Hw) peaks identified by Macs (c), Useq (d) and Sissrs (e) is also shown.



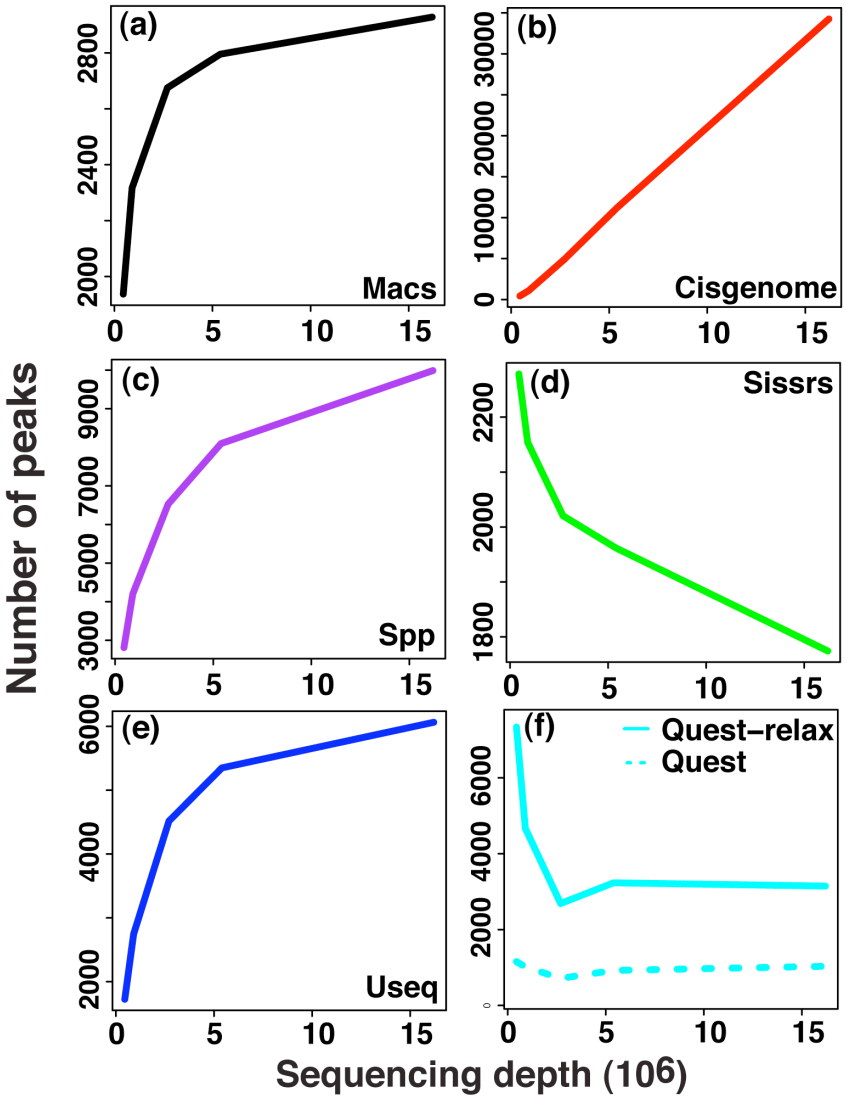
Supplementary Figure 6. The histograms of the fraction of redundant tags belonging to duplicate fragments in the paired-end library (green) and the fraction of all redundant tags (red) in peak regions are shown.



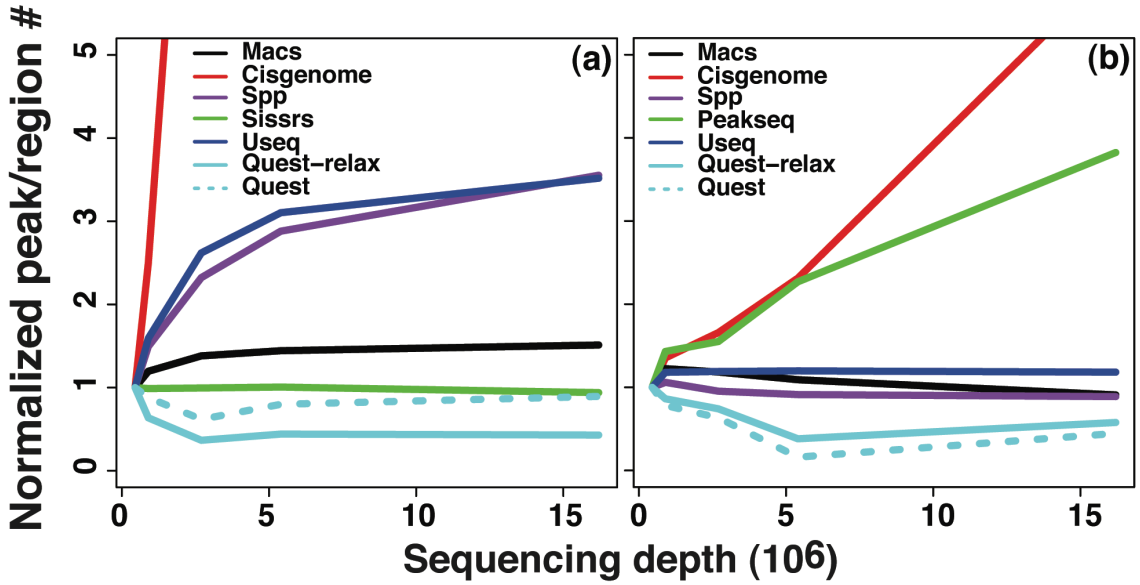
Supplementary Figure 7. The difference in the coverage of Su(Hw) ChIP-chip peaks by ChIP-seq peaks before and after removing redundant reads is shown for two algorithms (a) Macs and (b) Useq at different sequencing depths.



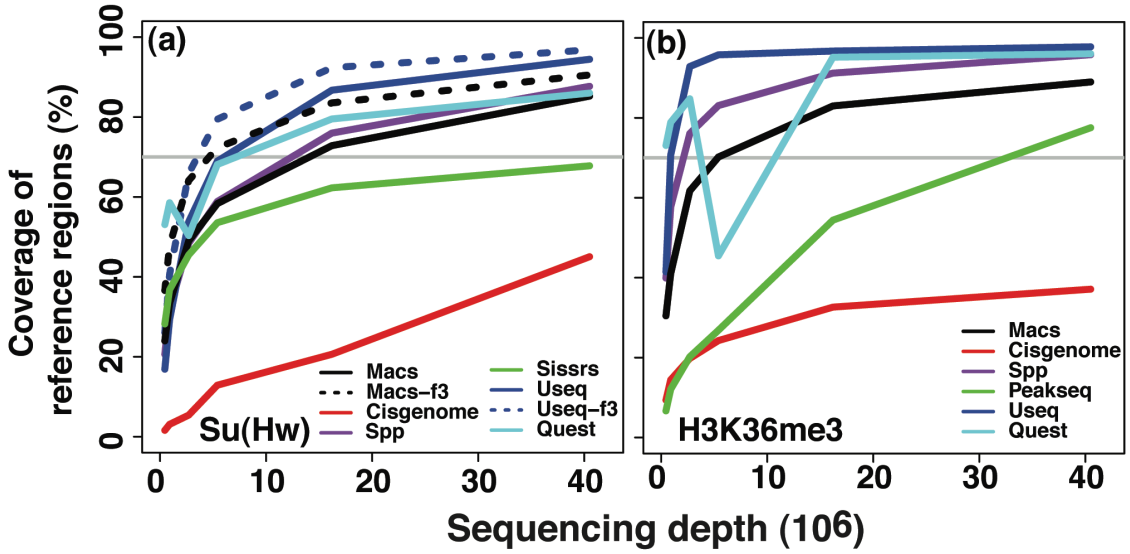
Supplementary Figure 8. The dependence of the number of Su(Hw) peaks that were identified by (a) Macs, (b) Cisgenome, (c) Spp, (d) Sissrs, (e) Useq, and (f) Quest on the sequencing depth is shown. Quest-relax stands for the condition under which less stringent parameters compared with the default ones were used for peak-calling.



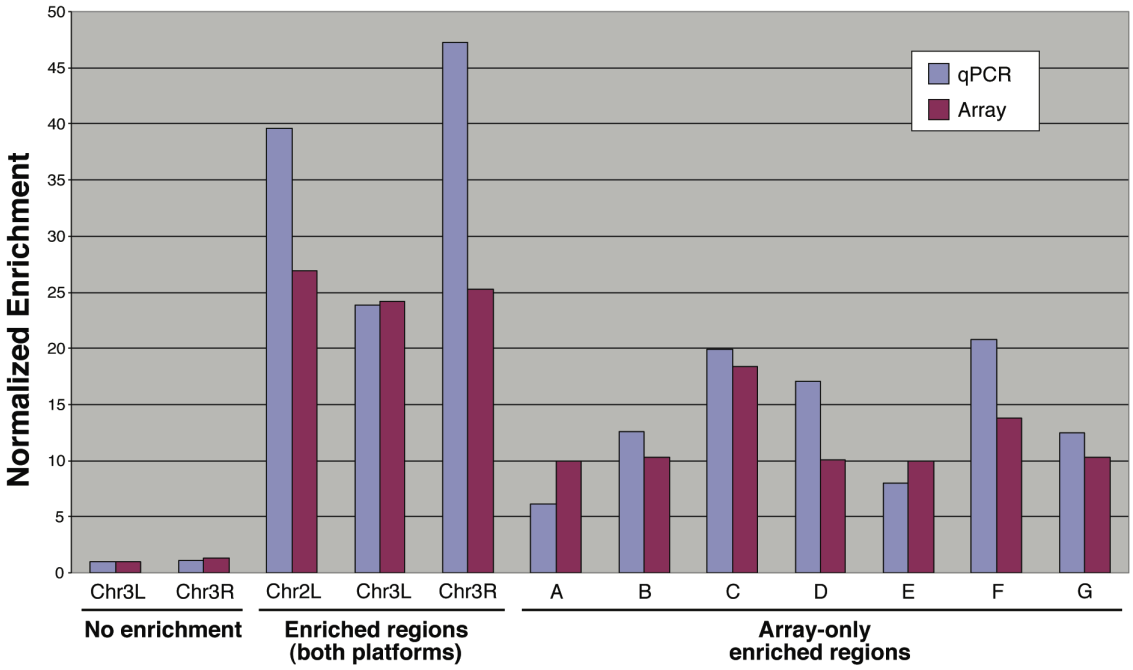
Supplementary Figure 9. The dependence of the number of Su(Hw) (a) and H3K36me3 (b) peaks that were identified by different algorithms on the sequencing depth is shown. For cross-algorithm comparison, the peak number at the sequencing depth of 0.45M tags of individual algorithm was used to normalize the number of peaks at other sequencing depths.



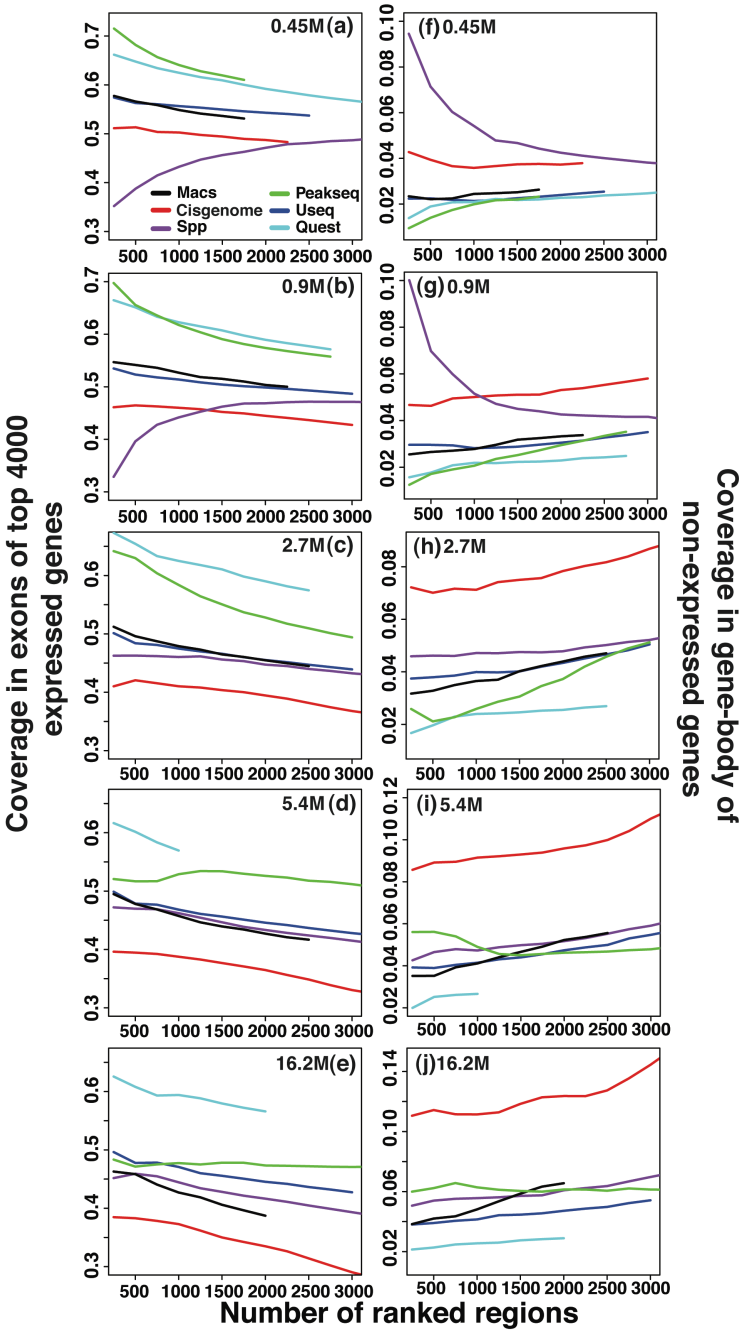
Supplementary Figure 10. The change in identified ChIP-enriched regions of (a) Su(Hw) and (b) H3K36me3 with respect to the regions that were identified using the complete data is shown with the increase of sequencing depth for different algorithms. Macs-f3 and Useq-f3 denote the Su(Hw) regions that have more than 3 fold enrichment and were identified by Macs and Useq, respectively.



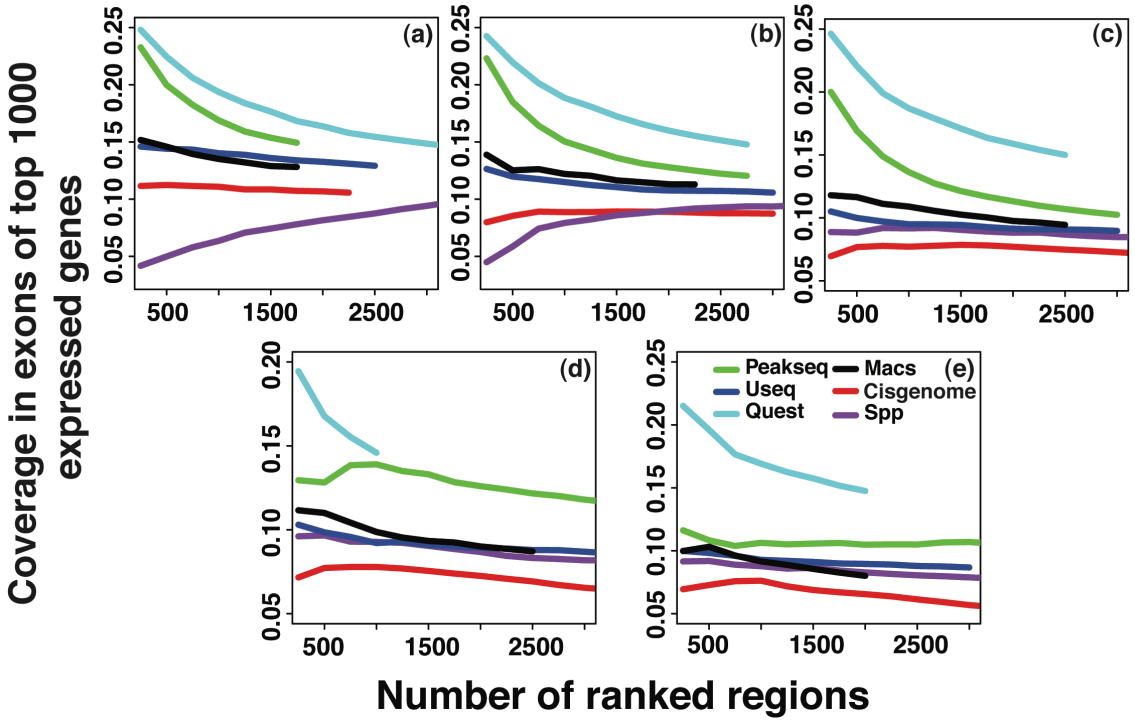
Supplementary Figure 11. The ChIP enrichment of 7 array-unique Su(Hw) sites, 3 positive control and 2 negative control sites that were quantified by both qPCR and tiling array is shown. All enrichment values were normalized relative to enrichment at the Chromosome 3L negative control region.



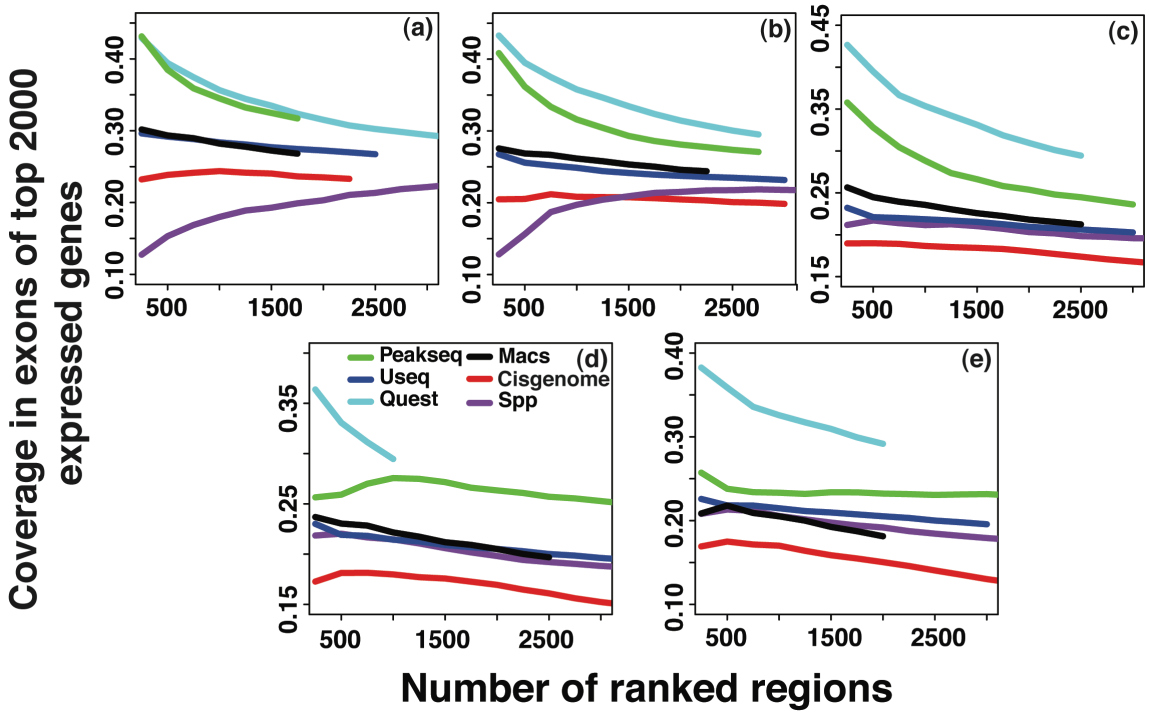
Supplementary Figure 12. An evaluation of the sensitivity and false-positive rate of six alogrithms (Macs, Cisgenome, Spp, Peakseq, Useq, and Quest) in identifying H3K36me3-enriched regions is shown. The sensitivity is approximated by the per-base region coverage of exons from top 4000 expressed genes and is plotted as a function of the number of top-ranked regions at the sequencing depths of 0.45M(a), 0.9M(b) and 2.7M(c), 5.4M(d), and 16.2 M (e) reads. The false-positive rate is approximated by the per-base region coverage of the bodies of genes with negligible expression level and is plotted as a function of the number of top-ranked regions at the sequencing depths of 0.45M(f), 0.9M(g) and 2.7M(h), 5.4M(i), and 16.2 M (j) reads.



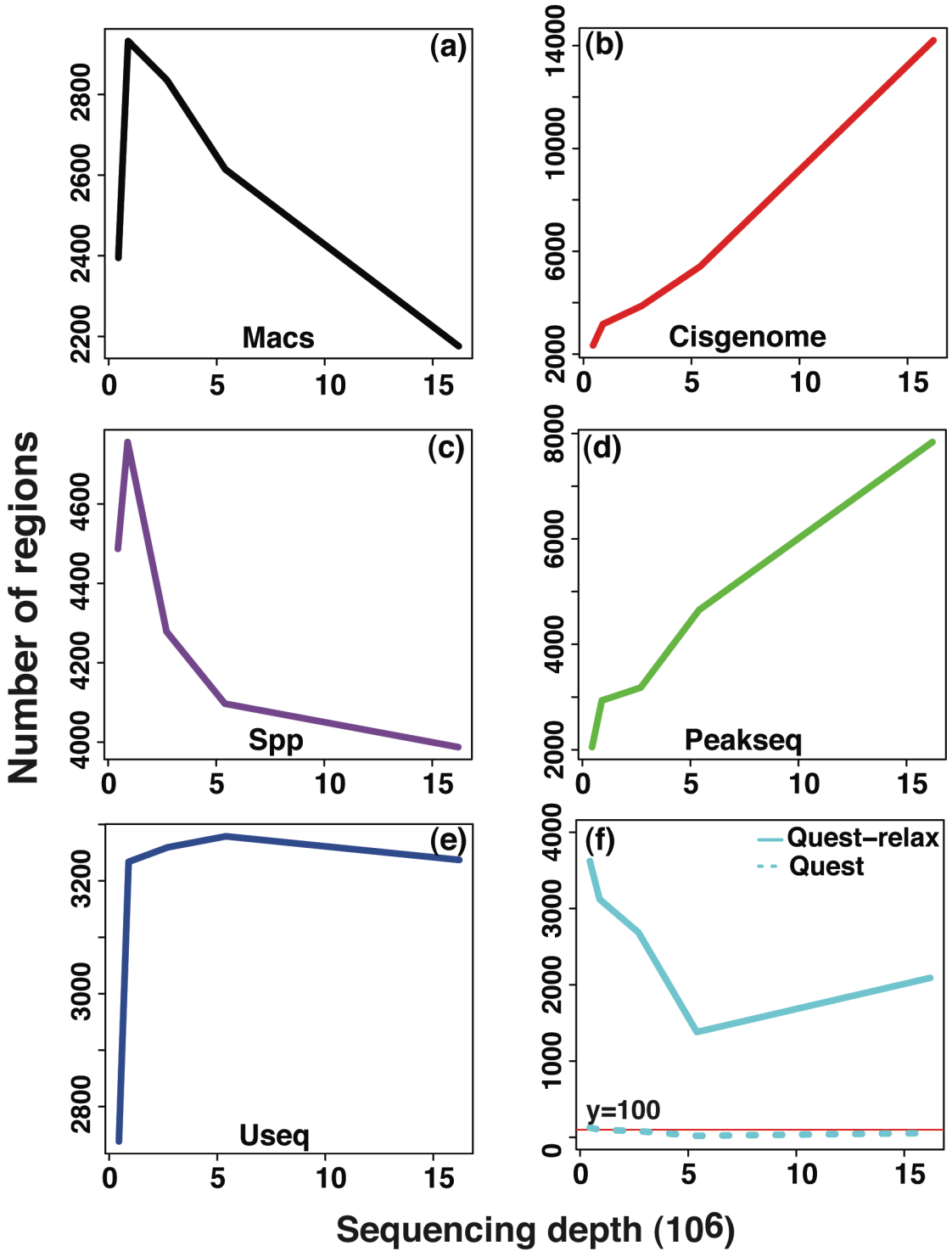
Supplementary Figure 13. An evaluation of the sensitivity of six algorithms (Macs, Cisgenome, Spp, Peakseq, Useq, and Quest) in identifying H3K36me3-enriched regions is shown. The sensitivity is approximated by the per-base region coverage of exons from top 1000 expressed genes and is plotted as a function of the number of top-ranked regions at the sequencing depths of 0.45M(a), 0.9M(b) and 2.7M(c), 5.4M(d), and 16.2 M (e) reads.



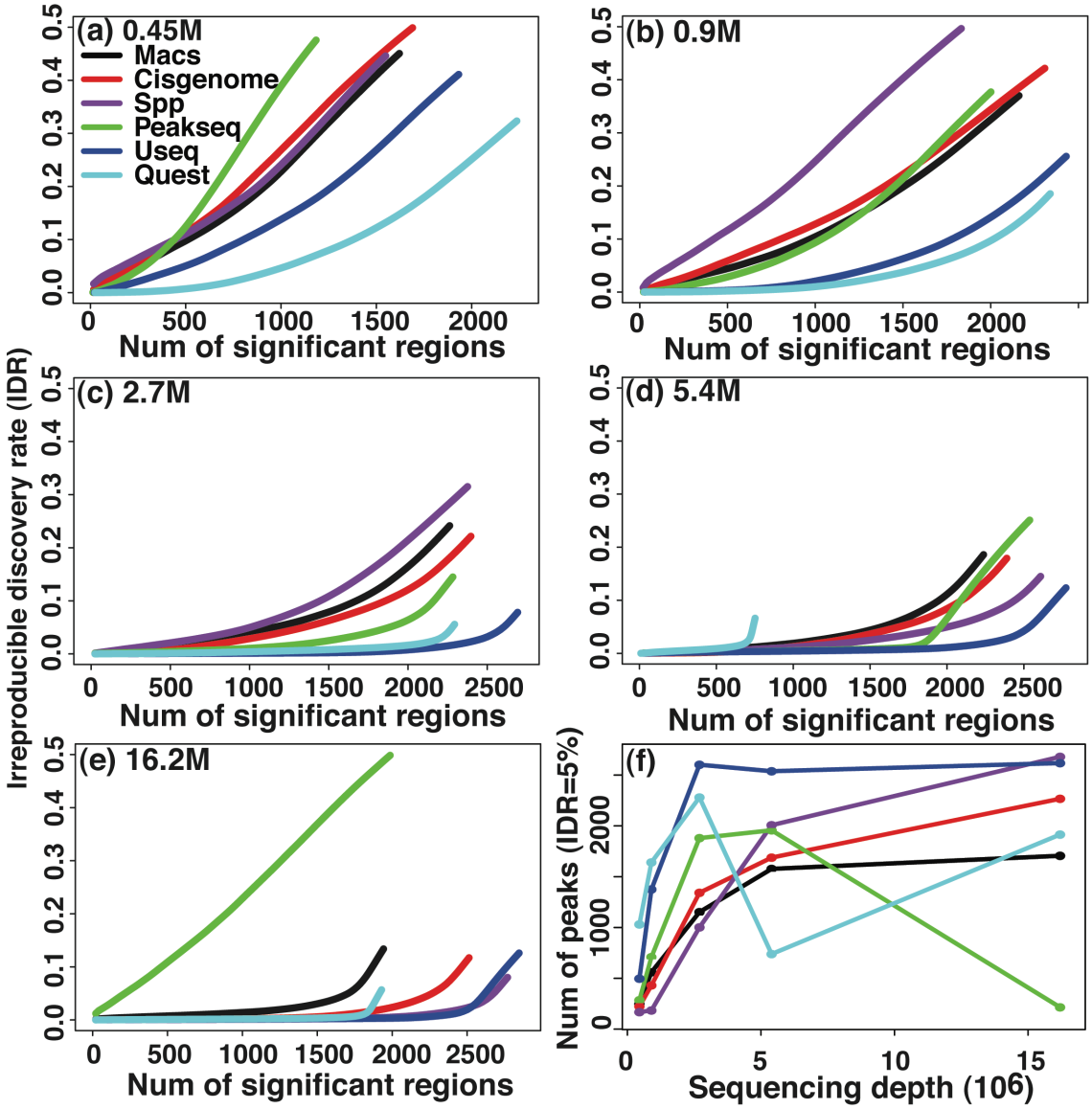
Supplementary Figure 14. An evaluation of the sensitivity of six algorithms (Macs, Cisgenome, Spp, Peakseq, Useq, and Quest) in identifying H3K36me3-enriched regions is shown. The sensitivity is approximated by the per-base peak coverage of exons from top 2000 expressed genes and is plotted as a function of the number of top-ranked peaks at the sequencing depths of 0.45M(a), 0.9M(b) and 2.7M(c), 5.4M(d), and 16.2 M (e) reads.



Supplementary Figure 15. The dependence of the number of H3K36me3-enriched regions that were identified by (a) Macs, (b) Cisgenome, (c) Spp, (d) Peakseq, (e) Useq, and (f) Quest on the sequencing depth is shown. Quest-relax stands for the condition under which less stringent parameters compared with the default ones were used.



Supplementary Figure 16. An evaluation of the consistency of identified H3K36me3-enriched regions between replicates by six algorithms (Macs, Cisgenome, Spp, Peakseq, Useq, and Quest) is shown. The number of reproducible regions at various IDR levels is plotted at the sequencing depths of 0.45M(a), 0.9M(b) and 2.7M(c), 5.4M(d), and 16.2 M (e) reads. In (f), the number of significant regions identified at the IDR of 5% is plotted as a function of sequencing depth.



Supplementary Table 1. The total number of uniquely mapped paired-end and single-end tags of gDNA, input sample, and the ChIP sample of Su(Hw) and H3K36me3 from all sequencing runs.

	Single-end	Paired-end
gDNA	85,689,583	61,673,578
Input	104,304,896	24,327,040
Su(Hw)	117,390,170	27,703,152
H3K36me3	148,703,534	37,378,264

Supplementary Notes.

1. Chromatin input

Chromatin input samples are generated by chromatin fragmentation via sonication or enzymatic reaction and are often used to model the background signal in a ChIP experiment. Chromatin extraction was done by first fixing the cell and then lysing the cell with lysis buffer.

2. Antibody for ChIP-seq experiments

A high-quality antibody with high specificity and sensitivity is the key determinant to the quality of raw ChIP-seq data. The antibody quality and the suitability for ChIP should always be evaluated before starting a ChIP-seq experiment[1].

3. The dependence of percentage of the uniquely mapped PE and SE reads on read length

We compared the percentage of the uniquely mapped PE reads that were also uniquely mapped when the PE reads were treated as if they were independent SE reads at different read lengths. We generated the reads of different lengths by trimming the 3' end of each read. We observed a general increase in uniquely mapped SE reads with an increased read length. The percentage of uniquely mapped SE reads was below 10% at a read length of 18 bp and was over 80% when the read length exceeds 22 bp. The increase in uniquely mapped SE reads gradually saturated once the read length exceeds 25 bp(**Fig. 2a**). The sequencing error rate increases toward the 3’ end of the reads, and the trimmed reads that have a shorter length have a lower error rate. Therefore, the observed trend in unique mappability is a composite effect of the change in both the length and the error rate of the read.

4. The low mappability of the Su(Hw) enriched regions that were detected on a tiling array but were missed by ChIP-seq

We obtained the mappability data of *Drosophila* genome from an early study[2] (Methods). We found that the vast majority of the peaks occurred in the genomic regions that have low mappability[2, 3].

Supplementary methods

ChIP-qPCR validation of array-specific Su(Hw)-enriched regions

Seven sites were randomly selected from the Su(Hw) binding sites that were unique to array platform and have at least a fold change of 4 for experimental validation. The ChIP enrichment of 3 positive control sites, 2 negative control sited and the 7 selected sites were quantified by real time PCR analysis using amplicons spanning the peak summits at these regions. Primer sequences for these regions, as well as for positive and negative control regions, are available upon request. Real time PCR was performed per the manufacturer's instructions (Applied Biosystems), and ChIP enrichment was quantified with respect to input using the delta Ct method. All enrichment values were normalized relative to enrichment at the Chromosome 3L negative control region.

References

1. **Egelhofer, T.A. et al. An assessment of histone-modification antibody quality. *Nat Struct Mol Biol* 18, 91-93 (2011).**
2. **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).**
3. **Rozowsky, J. et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 27, 66-75 (2009).**