BIOINFORMATICS ORIGINAL PAPER
Vol. 30 no. 20 2014, pages 2860–2867
doi:10.1093/bioinformatics/btu402

Genome analysis
Advance Access publication June 24, 2014

CNV-guided multi-read allocation for ChIP-seq

Qi Zhang[1,*] and Sündüz Keleş[1,2,*]

[1]Department of Biostatistics and Medical Informatics, 425 Henry Mall and [2]Department of Statistics, 1300 University Avenue, Madison, WI 53706, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: In chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) and other short-read sequencing experiments, a considerable fraction of the short reads align to multiple locations on the reference genome (multi-reads). Inferring the origin of multi-reads is critical for accurately mapping reads to repetitive regions. Current state-of-the-art multi-read allocation algorithms rely on the read counts in the local neighborhood of the alignment locations and ignore the variation in the copy numbers of these regions. Copy-number variation (CNV) can directly affect the read densities and, therefore, bias allocation of multi-reads.

Results: We propose cnvCSEM (CNV-guided ChIP-Seq by expectation-maximization algorithm), a flexible framework that incorporates CNV in multi-read allocation. cnvCSEM eliminates the CNV bias in multi-read allocation by initializing the read allocation algorithm with CNV-aware initial values. Our data-driven simulations illustrate that cnvCSEM leads to higher read coverage with satisfactory accuracy and lower loss in read-depth recovery (estimation). We evaluate the biological relevance of the cnvCSEM-allocated reads and the resultant peaks with the analysis of several ENCODE ChIP-seq datasets.

Availability and implementation: Available at http://www.stat.wisc.edu/~qizhang/

Contact: qizhang@stat.wisc.edu or keles@stat.wisc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 2, 2014; revised on June 8, 2014; accepted on June 18, 2014

1 INTRODUCTION

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is currently the dominant high-throughput assay in functional genomics and epigenomic research, especially for mapping transcription factor–DNA (TF-DNA) interactions and histone modifications. In a typical ChIP-seq experiment, tens of millions of DNA fragments (usually 100–1000 bp) are sequenced directly from one or both ends, and short reads (usually 25–100 bp) are recorded. The analysis of ChIP-seq data starts with the mapping of the short reads to a reference genome with a short-read aligner, e.g. Bowtie (Langmead *et al.*, 2009) and BWA (Li and urbin, 2009). Because of repetitiveness of the genomes, errors in reads and the short-read lengths, 10–30% of the reads can map to multiple locations (Chung *et al.*, 2011). Common practice in ChIP-seq analysis is to use only uniquely mapping reads (uni-reads) and discard the reads that map to multiple locations (multi-reads) (Landt *et al.*, 2012). Although using uni-reads alone is adequate in many cases, it makes any inference about the repetitive regions challenging, if not impossible. In addition, recently developed multi-read allocation methods have successfully argued that use of such reads leads to identification of many novel protein–DNA interactions in ChIP-seq experiments (Chung *et al.*, 2011; Newkirk *et al.*, 2011; Wang *et al.*, 2010, 2013). Furthermore, they illustrated that the effective increase in the coverage of the highly repetitive regions owing to multi-reads can not be simply achieved by adjusting for low mappability during peak calling (Chung *et al.*, 2011).

Although there are some specific modeling differences among different ChIP-seq aligners, they predominantly operate by allocating multi-reads to their mapping locations based on the read counts of the local neighborhoods of the mapping locations, which largely relies on the uni-read counts. Regions of high sequence similarity tend to have low uni-read counts, making it challenging for the algorithms to distinguish between different mapping locations. Furthermore, none of them take into account copy-number variations (CNVs), which are structural variations of a genome that reflect the differences between the sample genome and the reference genome in the number of copies of DNA segments. CNV has a direct effect on read depths of the sequencing samples. In the literature, this effect has been appreciated in peak calling (Ashoor *et al.*, 2013; Pickrell *et al.*, 2011; Rashid *et al.*, 2011) and in differential epigenome (ChIP-seq and RNA-seq) analyses (Robinson *et al.*, 2012). As of today, none of the multi-read mapping methods has considered the potential effect of CNV on multi-read allocation and the power it might provide for discriminating the mapping locations of multi-reads. The former methods (Ashoor *et al.*, 2013; Rashid *et al.*, 2011) estimate CNV from the mapped ChIP or control samples and incorporate this information during peak identification. However, if the reads are allocated without taking into account the CNV, CNV estimated from such mapped reads will be biased, especially in the repetitive regions (e.g. Table 1). Furthermore, even if the reads are mapped correctly, their genome-wide coverage is not uniform (Landt *et al.*, 2012), and the high copy-number regions identified from such data may be confounded with the open chromatin regions. This motivates us to correct the CNV bias at the stage of read allocation by using external CNV data instead of estimating it from the ChIP or control samples. We refer to Rozowsky *et al.* (2011) for a similar discussion in a different setting.

*To whom correspondence should be addressed.

To investigate the impact CNVs might have on multi-read allocation, we extracted all the reads that align to exactly two locations of the reference genome in a Ctcf ChIP-seq sample from GM12878 cells (such two locations are referred to as an alignment pair from hereon). We found that for 15.2% of reads that map two locations, the copy numbers of the alignment pairs differed by at least one (Supplementary Fig. S1 and Supplementary Table S1); and therefore, we expect the CNVs to inform the multi-read allocation for such reads. Even for the alignment pairs with the same copy number, CNVs may still have an indirect impact by affecting the allocation of other reads in the neighborhoods of these alignment locations. For multi-reads with more than two alignments, the copy numbers are more likely to vary among the alignment locations, and a larger impact of CNV is expected.

In this article, we develop cnvCSEM (CNV-guided ChIP-Seq by expectation-maximization algorithm), a flexible framework that guides multi-read allocation by CNVs. The cnvCSEM takes advantage of the state-of-the-art multi-read allocation algorithms and incorporates CNV information parsimoniously. We use the fact that various array- and sequence-based CNV detection algorithms have been developed (Abyzov *et al.*, 2011; Komura *et al.*, 2006), and large consortia such as 1000 Genomes (The 1000 Genomes Project Consortium, 2012) have produced high-throughput sequencing data for CNV detection. Combining these efforts, detecting CNV and estimating copy number with a reasonable accuracy is possible. Our data-driven simulation results show that (i) cnvCSEM increases multi-read allocation coverage and significantly reduces allocation ambiguity in the segmental duplication regions (SDR) with only a marginal loss in accuracy, and (ii) cnvCSEM also improves the accuracy of the read-depth recovery, especially in the highly repetitive regions with low copy numbers. Our study of several ENCODE TF ChIP-seq experiments demonstrate the biological relevance of the cnvCSEM-allocated reads and the ChIP-seq peaks identified in the downstream analysis.

2 METHODS

2.1 Preliminaries

We model the reads with the following generative model, which underlies the CSEM algorithm (Chung, 2012). Let M be the total number of genomic locations, and N be the total number of reads. For each read R_i, $i=1,\ldots,N$, we define $Z_i=(Z_{i1},\ldots,Z_{iM})$ as the indicator of the origin of R_i, i.e. $Z_{ij}=1$ if the read R_i originates from location j in the genome, and $Z_{ij}=0$ otherwise. We model $Z_1,\ldots,Z_N$ as independent and identically distributed samples from a multinomial distribution with parameter vector $\pi=(\pi_1,\ldots,\pi_M)$. Multi-read allocation can be achieved in two steps based on this model. The first step is to estimate the posterior distributions of $Z_1,\ldots,Z_N$, which allocate the multi-reads fractionally. Then, the most likely originating location of each read can be determined based on its posterior. A natural way of estimating the posteriors is by an expectation-maximization (EM) algorithm. The expectation and the maximization steps of the CSEM have explicit form solutions, so numerical optimization is not needed, and the computation is fast. Specifically, for $i=1,\ldots,N$ and $j=1,\ldots,M$, we define $H_{ij}=1\{R_i \text{ can be aligned to location j}\}$, and let $\pi^{(t)}=(\pi_1^{(t)},\cdots,\pi_M^{(t)})^T$ be the estimate of π at iteration t. Then, E- and M- steps for iteration $(t+1)$ are as follows.

E-step: For a given read R_i, $i=1,\ldots,M$, the expectation of Z_{ij} is given by the following equation:

$$z_{ij}^{(t+1)}=\frac{\pi_j^{(t)}H_{ij}}{\sum_{k=1}^{M}\pi_k^{(t)}H_{ik}}. \tag{1}$$

M-step:

$$\mu_j^{(t+1)}=\frac{1}{N}\sum_{i=1}^{N}z_{ij}^{(t+1)}. \tag{2}$$

CSEM uses a smoothed EM (EMS) algorithm to enable information sharing among the neighboring locations on the genome. Specifically, in the EMS algorithm, $\mu_j^{(t+1)}$ in (2) denotes an initial estimate of π_j and, for a preselected window size w, $\pi_j^{(t+1)}$ is determined by smoothing $\mu_j^{(t+1)}$ over the $(2w+1)$ bp window around j in a smoothing step (S-step) that follows the M-step.

S-step:

$$\pi_j^{(t+1)}=\frac{1}{2w+1}\sum_{k=j-w}^{j+w}\mu_k^{(t+1)} \tag{3}$$

Although the EMS algorithm enjoys convergence to a local maximum of the likelihood, its outcome significantly depends on the initial estimates, and there is no guarantee that it converges to the maximum likelihood estimator. This is especially true in our setting, where the model is multimodal, and the parameter space is of a high dimension. Although largely ignored in many applications of the EM algorithm in computational biology, appropriate initialization of the EM algorithm has been recognized as an important issue and non-trivial problem in the practice of ‘Big Data’ (Fayyad *et al.*, 1998; Toutanova and Galley, 2011). CSEM uses a uniform distribution as its starting configuration, which ignores the natural variation of ChIP-seq data. Our main contribution in this work is a CNV-guided initialization scheme for the EMS algorithm. Given the CNV estimates from the same cell/tissue types as the ChIP sample, this initialization for the EMS algorithm encodes the natural genomic variation in ChIP-seq read density, and hence, it is more biologically relevant than the initial uniform configuration. Consequently, we expect the final estimate of the read density π to be closer to the underlying true read density.

2.2 cnvCSEM: a general pipeline for CNV-guided multi-read allocation

We propose the following CNV-aware multi-read allocation framework:

(1) CNV detection and copy-number estimation in the cells that ChIP-seq experiment is performed.

(2) Initial weight assignment: estimating the initial value, $\pi^{(0)}$, so that it adapts to the CNV of the sample.

(3) EMS algorithm as in (1)–(3) with the estimated $\pi^{(0)}$ as the initial value.

We estimate the initial value $\pi^{(0)}$ non-parametrically as follows. We first run CSEM on the ChIP and the control data to be analyzed, and also annotate the reference genome with copy numbers estimated either from sequence- or array-based data. For each value x of the copy-number estimates, we then calculate the average read-depth of all the genomic locations with the same copy number. This step essentially aggregates ChIP or input read counts across locations with the same estimated copy number. Then, this average read depth is used as the initial weight at all the locations with copy number x. As copy number is also an estimated quantity, we truncate high copy numbers to achieve robustness. Specifically, we set the copy numbers ≥ 4 to 4 for the applications presented in this article. Further details on the initialization procedure are provided in the Supplementary Materials, where we also illustrate with a theoretical analysis and simulation studies that incorporating CNV information into initialization is sufficient and the gain by incorporating it into the actual likelihood and the update steps of the EMS algorithm is expected to be negligible.

In many CNV datasets, only the break point positions of the CNV regions, instead of the absolute copy numbers, are available. For example, for the HepG2 cell line commonly used in the ENCODE project, the highest resolution CNV information is available from an array-based experiment, where the genome is segmented into amplification regions, normal regions and heterozygous and homozygous deletion regions. To accommodate the use of CNV information for these cases, we assign pseudo copy numbers to these regions. Specifically, we encode 0 = homozygous deletion, 1 = heterozygous deletion, 2 = normal and 3 = amplification. The genomic regions that are not annotated with one of these four categories are treated as being normal. We remark that the numerical valued pseudo copy numbers of 0 and 1 do not necessarily reflect the numerical order of the copy numbers. Although it is natural to assign 2 to the normal copy-number regions, one may question the choice of the values assigned to the other three classes. The estimated $\pi^{(0)}$ is invariant to the actual values of the (pseudo) copy numbers. This is because the elements of $\pi^{(0)}$ are estimated from the regions with the same copy number, and the estimation only depends on the break points of the CNV regions, but not the actual numerical values of the (pseudo) copy numbers.

Our strategy of estimating $\pi^{(0)}$ is flexible and does not rely on CSEM or any other specific models or software. For example, one can use the read depth of the uni-reads instead of the CSEM output for calculating the aggregated read depth for initialization.

3 RESULTS

In this section, we compare the following three read allocation strategies using data-driven simulations and actual ChIP-seq experiments.

- Uni: only uni-reads are kept, and all the multi-reads are discarded.
- CSEM: in addition to the uni-reads, all the alignments of multi-reads with a CSEM posterior probability ≥ 0.5 are kept.
- cnvCSEM: in addition to the uni-reads, all the alignments of multi-reads with a cnvCSEM posterior probability ≥ 0.5 are kept.

Although keeping multi-read allocations with a posterior probability $\geq$0.5 reduces the number of multi-reads used, it facilitates downstream analysis with commonly adapted ENCODE ChIP-seq uniform analysis pipeline, which requires each read to map to a single location on the reference genome (Landt *et al.*, 2012).

Throughout this article, we used Bowtie aligner (Langmead *et al.*, 2009). Specifically, multi-read allocation with CSEM and cnvCSEM used reads with at most two mismatches and a maximum of 99 alignment locations. Reads with only one reported alignment location are referred to as uni-reads.

3.1 Simulation-based evaluation

We conducted a simulation study to examine the impact of CNVs on multi-read allocation, especially in the repetitive regions. We first simulated short reads from a repetitive sequence-enriched segment of the human genome with synthetic binding events and copy numbers, and then compared the three read allocation methods (Uni, CSEM and cnvCSEM) in terms of coverage, read allocation accuracy and the read-depth recovery.

We chose the segment chr22:21460001-21920000 of the human genome for the simulation experiment. There are two pairs of long repetitive regions that are similar to each other in this interval based on the hg19 (build 37) segmental duplication database of the human genome (Bailey *et al.*, 2002). These pairs are [21465673, 21548303], [21613746, 21695794] and [21727109, 21797378], [21846050, 21917116] and have percent sequence identities of 90%. We artificially divided this segment into 11 regions and assigned them different copy numbers. We simulated 50 binding events with various binding strengths and the read density that captured the synthetic binding signals and CNV information. We used a read length of 36 bp, as this is the typical read length for the large collection of publicly available ChIP-seq data. To address both the randomness of read sampling and the variation in the ChIP read density, we generated 10 read densities and simulated 10 samples from these densities resulting in 100 simulation replications. Each sample included 2000 simulated reads. Supplementary Materials provide further details on the specifics of this simulation study.

Figure 1 compares the three allocation methods in terms of their coverage and allocation accuracy. We observe that on average 21.1% of the simulated reads are uni-reads. The allocation accuracy of these uni-reads is on average 99.7%; however, the coverage is low, with an average of only 422 reads of 2000 reads across the simulation replications. In contrast, CSEM allocates 36.6% of the reads in the sense that each such read has a unique alignment location with a posterior probability ≥ 0.5 and has an average accuracy rate of 90.4%. Therefore, in addition to the average of 422 uni-reads, it recovers on average 310 multi-reads, a 73.4% increase at the cost of 9.3% loss in the overall accuracy. Detailed analysis of the CSEM allocations reveals that, because of the repetitiveness of the segment used in the simulation, CSEM allocates on average 30.1% (613 of them) of the multi-reads ambiguously in the sense that they are distributed to two alignment locations with a posterior probability of 0.5 for each. This is a direct consequence of two alignment locations with high sequence similarity providing a small number of uni-reads to CSEM to discriminate between them. In comparison,

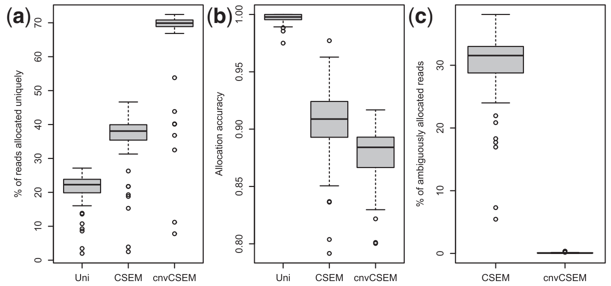

Fig. 1. Simulation-based evaluations. (**a**) Percentage of reads that are uniquely allocated; (**b**) accuracy rates of the allocations; (**c**) Percentage of ambiguously allocated reads. Boxplots display the results over 100 simulation replications with 2000 reads each

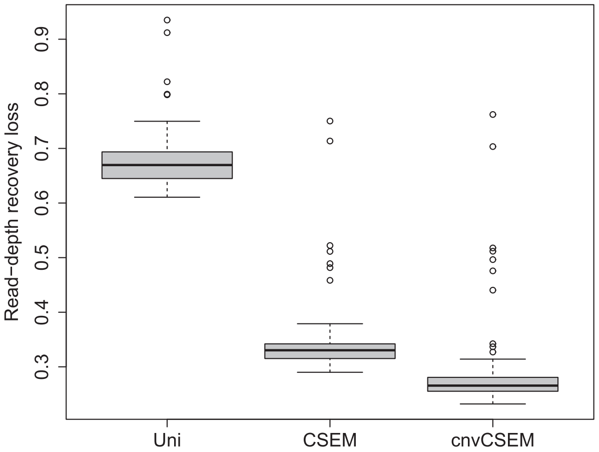

Fig. 2. Simulation-based evaluation. Read-depth recovery loss of Uni, CSEM and cnvCSEM read allocation methods

cnvCSEM allocates 67% of the reads uniquely (on average 918 multi-reads and 422 uni-reads) with an average accuracy rate of 88%. Therefore, by incorporating CNV information in multi-read allocation, we increase the coverage by another 83.1% with a small loss of 2.4% in accuracy.

We next compared the read-depth recovery by the three allocation methods. We treated the true and the recovered (estimated) read-depth curves as probability densities and measured the distance between them by the total variation distance of the probability measures (see Section 5.7 of the Supplementary Materials for details). Figure 2 illustrates that both CSEM and cnvCSEM perform better than Uni, and the distance between the true and the estimated read densities for cnvCSEM is also smaller than that of CSEM.

Figure 3 displays a typical example of the true and recovered read-depth curves. We observe that Uni completely misses the signals in the repetitive regions. Although CSEM successfully recovers some of the signal patterns in the repetitive regions, especially in the regions with high copy numbers, it leads to false-positive results in the repetitive regions with low copy numbers. Finally, cnvCSEM shifts the reads from the regions with low copy numbers to those with high copy numbers, removes most of the false-positive results in the repetitive regions with low copy numbers, and also refines the read density estimates in the regions with high copy numbers.

3.2 Evaluation on multiple ENCODE datasets

In this section, we analyzed several ChIP-seq samples along with their control samples in GM12878 (Ctcf, Atf3, Gabpa, Pol2), GM12891 (Pax5, Pou2, Pu1, Pol2) and HepG2 (Ctcf) cells from the ENCODE project. For GM12878 and GM12891 samples, high-coverage aligned sequence data measuring CNV were available from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2012). We used CNVNator (Abyzov *et al.*, 2011) to estimate copy numbers in these samples. For HepG2, we used the array-based categorical CNV result provided by the ENCODE project (GEO accession ID: GSM999286). We compared the three read allocation methods in terms of their consequences and biological implications in peak calling, and we focused most of our attention to the comparison of CSEM and cnvCSEM. For peak calling, we used ENCODE's uniform ChIP-seq processing pipeline (Landt *et al.*, 2012), where SPP (Kharchenko *et al.*, 2008) is used for calling the peaks, and the irreducible discovery rate (IDR) (Li *et al.*, 2011) is used for determining the optimal numbers of peaks for further downstream analysis.

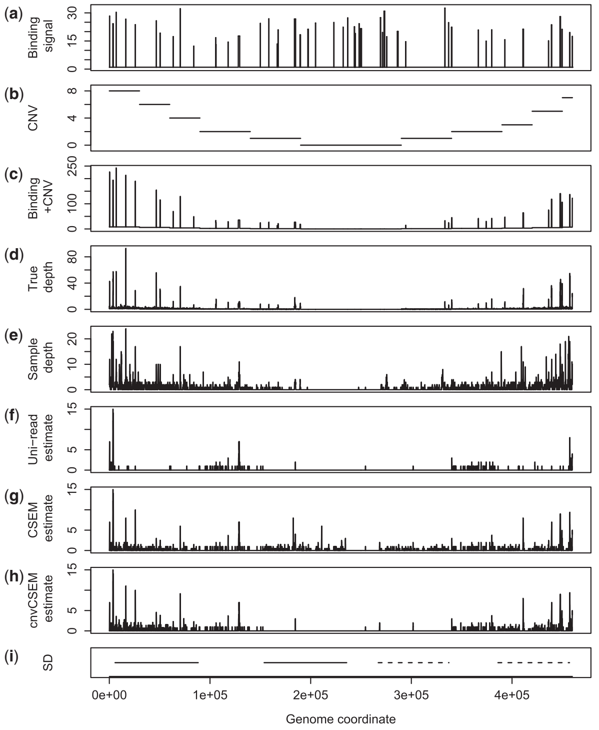

Fig. 3. Simulation-based evaluation: comparison of read-depth estimates of Uni, CSEM and cnvCSEM. (**a**) Simulated binding signal; (**b**) simulated CNVs; (**c**) convolution of the binding signal and CNV information; (**d**) true read depth; (**e**) simulated sample read depth; (**f**) read-depth estimates by Uni; (**g**) read-depth estimates by CSEM; (**h**) read-depth estimates by cnvCSEM; (**i**) segmental duplication annotation, where the two solid lines and the two dashed lines represent the two pairs of regions with high sequence similarity

3.2.1 cnvCSEM tends to allocate more reads to the regions with high copy numbers: an essential change that the CNV information brings to multi-read allocation As discussed in Section 1, CNV has a direct effect on read densities of sequencing data. Assuming there is no binding signal and all the reads can be aligned uniquely to their true origins, the read depth should be roughly proportional to the copy number in the region, and consequently, it should be 0 in the regions with a copy number of 0. However, in practice, observed read depths are far from this ideal scenario because of sequencing errors and sequence repetitiveness in ChIP-seq and other high-throughput experiments. This phenomenon is more prevalent in the SDR, where the sequence is repetitive, and a much higher portion of reads are multi-reads compared with other regions (non-SDR). We also remark that we do not expect CNV and the ChIP-seq read-depths to match perfectly because of potential associations between the binding events (i.e. peaks) and CNV (see Section 3 in Supplementary

Materials for more discussion). Nonetheless, a level of consistency between the ChIP-seq read depth and the copy number is expected.

We first analyzed Ctcf ChIP-seq data from GM12878 cells with the three allocation methods. Table 1 displays the average read depth of the regions with the same copy number (0, 1, 2, 3 and ≥ 4) and the segmental duplication status (SDR/non-SDR). For the non-SDR, the average read depths of Uni, CSEM and cnvCSEM are all roughly proportional to the copy numbers, at least at the lower end. Both CSEM and cnvCSEM have higher read depths, especially in the regions with high copy numbers, compared with the uni-reads allocation. In contrast, for SDR, Uni read depth is much lower and is inconsistent with CNV. Hence, the uni-read counts in SDR are not sufficient in characterizing the local read density and guiding multi-read allocation. As a result, CSEM allocates similar amounts of reads to the regions with copy numbers 0 and 1, which is not consistent with CNV. In comparison, cnvCSEM allocates much fewer reads to the regions with copy number 0, thus is more consistent with the copy numbers in these regions. In general, cnvCSEM allocates fewer reads to the regions with low copy number and more reads to regions with high copy numbers. We note that inferences in SDR are generally much more difficult than those of non-SDR. Although CSEM and cnvCSEM do not perform as well in SDR compared with non-SDR, their SDR coverages are better than uni-read coverage (see Section 3 in Supplementary Materials for more discussion).

Table 1 displays two results for multi-read allocation with CSEM: the result with the default number of 200 iterations is in the column titled CSEM, and the column CSEM1000 has the output with 1000 iterations. We observe that the final estimates from these two runs are close to each other, indicating that the utilization of CNV information in the initialization is essential and can not be achieved by larger number of iterations in the EMS algorithm.

Table 1. Average read-depth estimates for segmental duplication and non-SDR

Copy number	Uni	CSEM	cnvCSEM	CSEM1000
	SDR			
0	0.061	0.725	0.390	0.723
1	0.131	0.868	0.815	0.855
2	0.428	1.252	1.232	1.252
3	0.155	1.500	1.573	1.502
≥ 4	0.387	2.544	2.985	2.558
	non-SDR			
0	0.002	0.006	0.006	0.006
1	0.319	0.409	0.412	0.410
2	0.985	1.050	1.050	1.050
3	1.528	2.169	2.172	2.169
≥ 4	3.541	16.342	16.376	16.331

Notes: The average read depth of the uni-reads (Uni), multi-reads allocated by CSEM with 200 iterations (CSEM) and with 1000 iterations (CSEM1000) and multi-reads allocated by cnvCSEM (cnvCSEM) for GM12878 Ctcf ChIP-seq dataset.

3.2.2 cnvCSEM improves the consistency between ChIP-seq peaks and CNV We next evaluated the three allocation strategies in terms of their downstream effects on the identified peaks. Table 2 summarizes the definitions of the peak sets that we compared, and Supplementary Table S2 tabulates the sizes of all the peak sets compared in the rest of the article. For the GM12878 Ctcf dataset, we compared the three read allocation methods by comparing the set differences of the three peak lists (Table 3). ENCODE's uniform ChIP-seq processing pipeline calls two lists of peaks: optimal as determined by a specific IDR (ENCODE default of 0.02 for the presented analysis) threshold and relaxed, which are super sets of the optimal peaks and include both high signal peaks and regions that do not show any ChIP enrichment. We note that 99.5% of the Uni-only peaks are captured in the extended lists of both CSEM and cnvCSEM; however, only 38.4% of the Multi-common are identified in the extended peak list of uni-reads. We included comparisons with the relaxed peak sets to ensure that our results hold irrespective of the specific IDR threshold used for peak calling.

Table 3 reveals that Common peaks are rarely in the regions with CNV. Comparison of the CSEM-only peaks with the cnvCSEM-only peaks indicates that a much lower percentage of the cnvCSEM-only peaks are in the regions with lower copy

Table 2. Definitions of the peak lists for comparison

Uni-only	In the optimal list of Uni, but not in those of CSEM and cnvCSEM
CSEM-only	In the optimal list of CSEM, but not in those of Uni and cnvCSEM
cnvCSEM-only	In the optimal list of cnvCSEM, but not in those of Uni and CSEM
Multi-common	In the optimal lists of CSEM and cnvCSEM, but not in that of Uni
Common	In all the three optimal lists of Uni, CSEM and cnvCSEM

Table 3. Comparison of GM12878 Ctcf peak lists from Uni, CSEM and cnvCSEM allocation strategies

	cnvCSEM-only	CSEM-only	Multi-common	Common
Number of peaks	249	78	1845	40752
% of copy number > 2	41.4	47.4	15.6	0.2
% of copy number < 2	3.6	18.0	7.0	0.3
% of copy number = 2	55.0	34.6	77.4	99.5
Number of copy number = 0	1	3	8	5
% in SDR	68.7	100	75.4	2.4

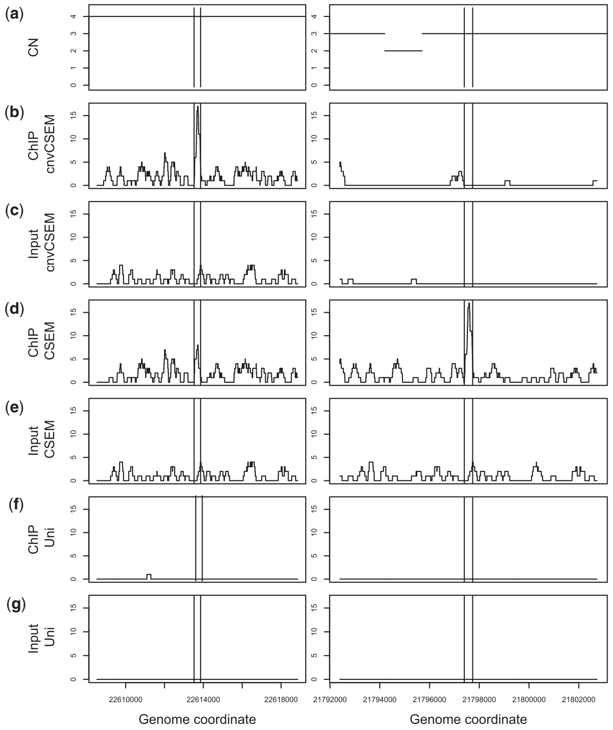

Fig. 4. cnvCSEM versus CSEM peaks across two regions in chr16 with high sequence similarity. (**a**) copy numbers (CN); (**b**) ChIP read counts by cnvCSEM; (**c**) input read counts by cnvCSEM; (**d**) ChIP read counts by CSEM; (**e**) input read counts by CSEM; (**f**) ChIP read counts by Uni; (**g**) input read counts by Uni

numbers. We note that some identified peaks may be located in regions with copy number 0 even when only uni-reads are used (five such peaks if only Uni-reads are used, and one and three such peaks in the cnvCSEM-only and CSEM-only categories, respectively). This can be due to errors in CNV estimation and/or read allocation. However, we observe that the number of such peaks is small.

The difference between the CSEM and the cnvCSEM optimal peak lists is due to CNV across the whole genome. As we have discussed in Section 3.2.1, cnvCSEM is more likely to align reads to the regions with high copy numbers compared with CSEM. In some extreme cases, it may even shift a peak from a region with low copy numbers to a region with high copy numbers. In Figure 4, the two peaks share >50% of the ChIP reads, and incorporating CNV information shifts the peak from the region in the right panel to the region in the left panel. Hence, CNV helps differentiate the regions with similar sequences, especially when the uni-read counts in these regions are low.

3.2.3 cnvCSEM-only peaks are supported by DNase-seq signals DNase-seq is a high-resolution assay that has been widely used in identifying the location of active regulatory regions (Hesselberth *et al.*, 2009; Song and Crawford, 2010). Genomic regions with high DNase-seq signals are more likely to harbor protein–DNA binding sites. Therefore, we use the enrichment of DNase-seq signal as an evidence of the biological relevance of the identified ChIP-seq peaks. Figure 5 compares the peak lists from Pol2 and Ctcf ChIP-seq experiments in GM12878 cells in terms of their DNase-seq signals. To highlight the influence of CNVs in the SDR, we perform the comparisons for the peaks in the SDR and non-SDR separately. cnvCSEM-only peaks have comparable or higher DNase-seq signal than the CSEM-only peaks both in SDR and in non-SDR. We also tested whether cnvCSEM-only peaks have higher DNase-seq signal than CSEM-only peaks in SDR (in terms of third quartile) by permutation tests (see Supplementary Materials for details). The *P*-values for Ctcf and Pol2 are 0.0086 and 0.3354, respectively.

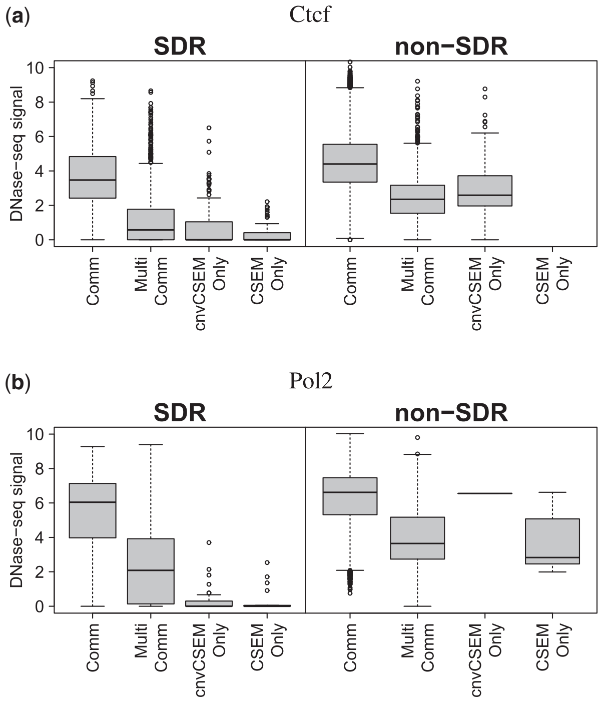

Fig. 5. Comparison of the peaks in terms of DNase-seq signal. An empty box implies that there are no peaks in this category. Both datasets are from GM12878 cells

3.2.4 cnvCSEM identifies additional Pol2 peaks in the promoters of expressed genes Pol2 is the largest subunit of RNA polymerase II and is essential for polymerase activity in the promoter regions of the active genes. To evaluate the biological relevance of the peaks identified by CSEM and cnvCSEM, we overlapped the identified Pol2 peaks with the promoter regions (–5 kb and + 0.5 kb of the transcription start site) of expressed genes, and compared the two lists of the promoter regions that harbor Pol2 peaks. For the GM12891 cells, we identified six promoter regions (for genes CLEC2D, TCEA2, SMN1, POLR2J3, CFP and ASMTL) that only contain cnvCSEM peaks, and two promoter regions (for genes BOLA2B and SLC25A6) with only CSEM peaks. The RNA-seq levels of these genes, for which a Pol2 promoter peak was only identifiable with either cnvCSEM or CSEM, were comparable with those genes with promoter Pol2 peaks from both CSEM and cnvCSEM. Specifically, the transcripts per million values computed by RSEM (Li and Dewey, 2011) varied between 20.13 and 57.74 for genes with cnvCSEM-only Pol2 peaks and 28.39 and 275.50 for genes with CSEM-only

Table 4. Percentage of peaks with the most significant *de novo* identified motif

Cell/factor	cnvCSEM-only	CSEM-only	Multi-common	Uni-only	Common-test
Ctcf (1)	43.4	39.7	61.1	65.9	83.9
Atf3 (1)	12.3	7.1	17.2	37.4	51.2
Gabpa (1)	9.6	13.5	26.8	47.0	71.4
Pax5 (2)	23.6	22.6	27.2	19.9	27.6
Pou2 (2)	7.9	14.0	15.5	13.8	24.4
Pu1 (2)	63.8	64.6	74.6	67.4	74.6

Notes: (1) and (2) in the first column denote GM12878 and GM12891 cells, respectively

peaks. We observed a similar phenomenon for the Pol2 dataset in the GM12878 cells. By incorporating the CNV information, we identified two additional expressed genes with cnvCSEM-only Pol2 peaks (POLR2J3 and AC006995.3 with transcripts per million values of 94.53 and 70.39).

3.2.5 cnvCSEM and CSEM lead to similar percentages of peaks with the binding motif We conducted *de novo* sequence analysis to further evaluate the cnvCSEM-only and CSEM-only peaks using the MEME Suite (Bailey *et al.*, 2009). We identified the most significant binding motif with a *de novo* sequence analysis of the top 500 Common peaks (using the sequences within ± 50 bp of the peak summits) using MEME. Then, we scanned both the rest of the Common peaks (Common-test) and the other peak sets from Table 2 (using the sequences within ± 150 bp of the peak summits). Overall, the motif occurrence percentages of the cnvCSEM-only peaks are comparable with their CSEM-only counterparts (Table 4).

3.2.6 CNV information with lower resolution is also useful Although current state-of-the-art for profiling CNVs is based on high-throughput sequencing assays, there are many datasets for which CNV information is available only from lower resolution array platforms. Using HepG2 cells, a cancer cell line for which many ChIP-seq datasets and array-based CNV information is available from the ENCODE project, we investigated whether lower resolution CNV information can be used in multi-read allocation. For HepG2 cells, the available CNV information is not the actual copy-number estimates along the genome, but a classification of the genome into the following four categories: homozygous deletion (hom.del), heterozygous deletion (het.del), normal and amplification (amp). We assigned pseudo copy numbers 0, 1, 2 and 3 to these regions, as described in Section 2, and performed the analysis presented in Sections 3.2.1 and 3.2.2 for the Ctcf ChIP-seq datasets from HepG2 cells. Although the differences in the performances of CSEM and cnvCSEM are smaller, the overall patterns are similar to those we obtained with higher resolution CNV data (Tables 5 and 6).

4 DISCUSSION

In ChIP-seq experiments and other short-read sequencing experiments, a significant fraction of the reads map to multiple locations. Processing of multi-reads can impact the downstream analysis of peak calling significantly. None of the currently available multi-read allocation methods take into account CNVs in the sample genomes. We showed in this article that using CNV provides additional information for discriminating mapping locations of a multi-read with similar uni-read counts. Our data-driven simulations revealed significant improvements in multi-read allocation accuracy and better read-depth recovery when using CNV information. Analysis of multiple ENCODE datasets indicated that peaks identifiable only when using copy-number information in multi-read allocation have biologically meaningful characteristics.

In our framework, we incorporated CNV in the initialization of the EMS algorithm instead of explicitly including it in the model because of the nature of the CNV information. Intuitively, copy numbers should influence the elements of the read density π as multipliers. If the copy numbers are already encoded in $\pi^{(0)}$ during the initialization step, their influence will be preserved in the updates of the iterations, except at the locations around the break points of the CNV regions. However, the typical numbers of break points are small. For GM12878 and GM12891 experiments, CNVnator detects <5000 CNV regions, and consequently <10000 break points. The array-based CNV annotation of HepG2 includes <600 break points. Compared with the length of the whole genome, this additional effect of

Table 5. Average read-depth estimates for segmental duplication and non-SDR: Ctcf ChIP-seq dataset from HepG2

	SDR			non-SDR		
CNV	Uni	CSEM	cnvCSEM	Uni	CSEM	cnvCSEM
hom.del	0.000	0.333	0.038	0.047	0.103	0.099
het.del	0.532	2.546	2.466	1.449	1.689	1.688
normal	1.254	6.677	6.674	3.563	3.899	3.899
amp	1.591	10.063	10.199	5.959	6.388	6.388

Notes: The average read depths of the uni-reads (Uni), multi-reads allocated by CSEM and multi-reads allocated by cnvCSEM.

Table 6. Comparison of HepG2 Ctcf peak lists from uni-read, CSEM and cnvCSEM allocation strategies

	cnvCSEM-only	CSEM-only	Multi-common	Common
Number of peaks	51	16	2903	53763
% of amp	19.6	0.0	12.0	13.2
% of deletion	2.0	25.0	4.1	2.7
% of normal	78.4	75.0	83.8	84.1
% in SDR	45.1	100	65.2	1.8

The deletion regions include both heterozygous and homozygous deletions. In all, 96.5% of the Uni-only peaks are captured in both the extended lists of CSEM and cnvCSEM; however, only 44.1% of the Multi-common are identified in the extended peak list of uni-reads.

CNV around the break points is rather small, especially if the EMS algorithm is already appropriately initialized under the guidance of CNV. In Supplementary Materials, we analytically investigated our read model and showed how an EMS algorithm can incorporate CNV in the updating steps. However, our calculations and simulations suggest that the results from this mathematically more complicated model will not be notably different from the cnvCSEM.

One curious question is how any given two regions with high sequence similarity can have different CNV estimates, while the uni-read counts in both regions are similarly low. We remark that this is mainly due to the local uniqueness of the repetitive regions and the high sequencing depths used for CNV estimation. Repetitive regions such as segmental duplications are usually composed of alternating short stretches of unmappable intervals (usually ≤ 200 bp), sequences of which are almost identical to some other genomic regions and the mappable intervals with more sequence uniqueness. The median maximum run length of unmappable bases across all the segmental duplications in the hg19 segmental duplication database is 66 bp. Overall, only 14.7% of the segmental duplications have a maximum unmappable run length >300 bp (data not shown). When the read coverage is high or the reads are long, there will be sufficiently large numbers of uni-reads overlapping the mappable intervals and differentiating regions with high sequence similarity. The sequencing-based CNV datasets we used from the 1000 Genomes Project are deeply sequenced (total of 811 and 741 million aligned reads for GM12878 and GM12891, respectively) and hence able to generate sufficiently large numbers of uni-reads to differentiate regions with high sequence similarity. For the array-based HepG2 CNV data, the median probe length is about 1500 bp, much larger than the typical read lengths in sequencing experiments. In contrast, typical ChIP-seq experiments usually have only 10–50 million aligned reads, and the uni-read counts in the repetitive regions are usually too low to differentiate the regions with similar sequences reliably.

In summary, our work can be viewed as part of the collective effort in removing background variation in sequencing-based genomic data analysis. Existing methods for ChIP-seq data analysis remove potential bias due to CNVs in peak calling (Rashid *et al.*, 2011) and differential epigenomic analysis (Robinson *et al.*, 2012). cnvCSEM incorporates CNVs at the level of alignment and improves the accuracy of multi-read mapping for individual datasets.

ACKNOWLEDGEMENTS

The authors thank Dr. Peng Liu in Keles' group for help in RNA-seq data.

Funding: This work was supported by National Institutes of Health Grants (HG007019 and HG003747 to S.K.).

Conflict of Interest: none declared.

REFERENCES

Abyzov,A. *et al.* (2011) CNVnator: an approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.

Ashoor,H. *et al.* (2013) HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics*, **29**, 2979–2986.

Bailey,J.A. *et al.* (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.

Bailey,T.L. *et al.* (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37** (**Suppl. 2**), W202–W208.

Chung,D. (2012) Statistical methods and software for ChIP-seq data analysis. Ph.D Thesis, University of Wisconsin, Madison.

Chung,D. *et al.* (2011) Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLoS Comput. Biol.*, **7**, e1002111.

Fayyad,U.M. et al. (1998) Initialization of iterative refinement clustering algorithms. In Agrawal,R. *et al.* (eds.), *KDD-98 Proceedings*, pp. 194–198. AAAI Press.

Hesselberth,J.R. *et al.* (2009) Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.

Kharchenko,P.V. *et al.* (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.

Komura,D. *et al.* (2006) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.*, **16**, 1575–1584.

Landt,S.G. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,Q. *et al.* (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.

Newkirk,D. *et al.* (2011) AREM: aligning short reads from ChIP-sequencing by expectation maximization. *J. Comput. Biol.*, **18**, 1495–1505.

Pickrell,J.K. *et al.* (2011) False positive peaks in chip-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics*, **27**, 2144–2146.

Rashid,N. *et al.* (2011) 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.

Robinson,M.D. *et al.* (2012) Copy-number-aware differential analysis of quantitative DNA sequencing data. *Genome Res.*, **22**, 2489–2496.

Rozowsky,J. *et al.* (2011) Alleleseq: analysis of allele-specific expression and binding in a network framework. *Mol., Syst. Biol*, **7**, 522.

Song,L. and Crawford,G.E. (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, 2010, pdb.prot5384.

The 1000 Genomes Project Consortium *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Toutanova,K. and Galley,M. (2011) Why initialization matters for IBM model 1: multiple optima and non-strict convexity. In: *HLT'11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers*. Vol. 2, pp. 461–466. Association for Computational Linguistics: Human Language Technologies.

Wang,J. *et al.* (2010) A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics*, **26**, 2501–2508.

Wang,R. *et al.* (2013) LOcating non-unique matched tags (LONUT) to improve the detection of the enriched regions for ChIP-seq data. *PLoS One*, **8**, e67788.